TRABALLO FIN DE GRAO
GRAO EN ENXEÑARÍA INFORMÁTICA
MENCIÓN EN COMPUTACIÓN

Ranking de usuarias de Reddit aplicando Modelos de Relevancia para trastornos depresivos

Estudante: Eliseo Bao Souto

Dirección: Álvaro Barreiro García

Miguel Anxo Pérez Vila

A Coruña, xuño de 2022.

A meus pais

Agradecementos

Quixera comezar dándolle as grazas aos meus directores de proxecto. A Álvaro, pola súa experiencia e os seus valiosos consellos. A Anxo, pola súa implicación e axuda infinita, incluso desde a distancia nos últimos meses. Traballar con vós supuxo unha experiencia moi enriquecedora tanto a nivel académico como persoal, que de seguro me permitiu evolucionar.

Tamén me gustaría agradecer profundamente o traballo e a atención de David Otero, parte indispensable deste proxecto, así como tamén a todas aquelas persoas do IRLab que dunha ou outra maneira participaron neste traballo. A vós, Javier e Alfonso, grazas.

Grazas, por suposto, á miña familia, fonte inesgotable de comprensión e apoio. Traballo constantemente para ser quen de vos devolver todo o que me levades dado e me continuades a dar. A María, por ser e estar sempre. E, como non, grazas a Lucía, por ser a mellor compañeira de vida e un piar para todos nós.

A todos e a todas, **grazas**.

Resumo

Os trastornos depresivos conforman un dos grupos de enfermidades máis comúns no mundo. Se ben é certo que existen tratamentos eficaces, ben por falta de recursos, ben polo estigma aínda a día de hoxe asociado, en moitas ocasións as consecuencias para quen padece este tipo de trastornos son devastadoras. Sabendo que a linguaxe manifestada polas persoas que sofren este tipo de enfermidades pode amosar evidencias da súa saúde mental, o obxectivo deste proxecto é explotar as posibilidades dos Modelos de Linguaxe baseados na Relevancia para seren aproveitados de cara á detección temperá. En concreto, tomando como punto de partida coleccións CLEF eRisk, búscase construír vocabularios de depresión. Estes vocabularios identifican termos de peso e relevancia en persoas con tendencias depresivas, e deben ser sometidos ás pertinentes fases de avaliación e comparación respecto a outros lexicóns validados. Complementariamente, preténdese ser quen de realizar *ranking* de suxeitos, isto é, a partires de textos escritos por unha serie de persoas, establecer unha ordenación para as mesmas en función do posible grao de depresión. Para a xestión do proxecto, utilizouse unha metodoloxía áxil, de modo que fose posible adaptar o proxecto en función dos resultados obtidos na experimentación. Conseguíronse resultados satisfactorios, especialmente en canto ao *ranking*, así como tamén se pautaron novas vías para a experimentación e ampliación.

Abstract

Depressive disorders are one of the most common groups of illnesses in the world. Although it is true that effective treatments exist, either due to the lack of resources or the stigma that is still associated, in many cases the consequences for those suffering from this type of disorders are devastating. Knowing that the language manifested by people suffering from this type of diseases can denote evidence of their mental health, the aim of this project is to exploit the possibilities of Relevance-Based Language Models to be used for early detection. Specifically, taking CLEF eRisk collections as a starting point, the goal is to build depression vocabularies. These vocabularies identify terms of weight and relevance in people with depressive tendencies, and must undergo phases of evaluation and comparison with other validated lexicons. In addition, we focus in being able to perform ranking, i.e., from texts written by a number of people, to establish a ranking for them according to the possible degree of depression. For the management of the project, an agile methodology has been used, so that it has been possible to adapt the project according to the results obtained in the experimentation. Satisfactory results have been achieved, especially in terms of ranking, as well as new avenues for experimentation and expansion have been set.

Palabras chave:

- Recuperación da Información
- Modelos de Relevancia
- Trastornos depresivos
- Depresión
- CLEF eRisk
- Lexicón
- Vocabulario
- Reddit
- Ordenación
- Avaliación

Keywords:

- Information Retrieval
- Relevance Model
- Depressive disorders
- Depression
- CLEF eRisk
- Lexicon
- Vocabulary
- Reddit
- Ranking
- Evaluation

Índice Xeral

Índice de Figuras

Índice de Táboas

Capítulo 1

Introdución

A depresión é un trastorno mental que vai máis alá das variacións habituais do estado de ánimo e as respostas emocionais breves aos problemas da vida cotiá. A súa intensidade é variable, pero en casos recorrentes supón un problema de saúde serio para a persoa que a sofre, pois pode padecer gran sufrimento e ver como se alteran as súas actividades laborais, escolares e familiares. En casos extremos, pode estar ligada a episodios suicidas, que representan a cuarta causa de morte no grupo de idade de 15 a 29 anos segundo a Organización Mundial da Saúde (OMS)[1]. Así, é unha das enfermidades máis comúns no mundo, afectando ao 3.8% da poboación. Por grupos de idade (ver figura 1.1), estímase que o 5.0% das persoas adultas a padecen, elevándose esta cifra ata o 5.8% no caso das persoas de entre 50 e 70 anos. Calcúlase que, en todo o mundo, uns 280 millóns de persoas teñen depresión [2].

Aínda que existen tratamentos eficaces contra os trastornos mentais, máis dun 75% das persoas afectadas en países de baixos e medianos ingresos non recibe tratamento algún [3]. Ademais da falta de recursos, resulta determinante a estigmatización asociada aínda a día de hoxe aos trastornos mentais. Resulta, por tanto, imprescindible un adecuado e temperán diagnóstico, ao que o traballo que aquí se presenta espera poder axudar.

[1] https://www.who.int/news-room/fact-sheets/detail/depression

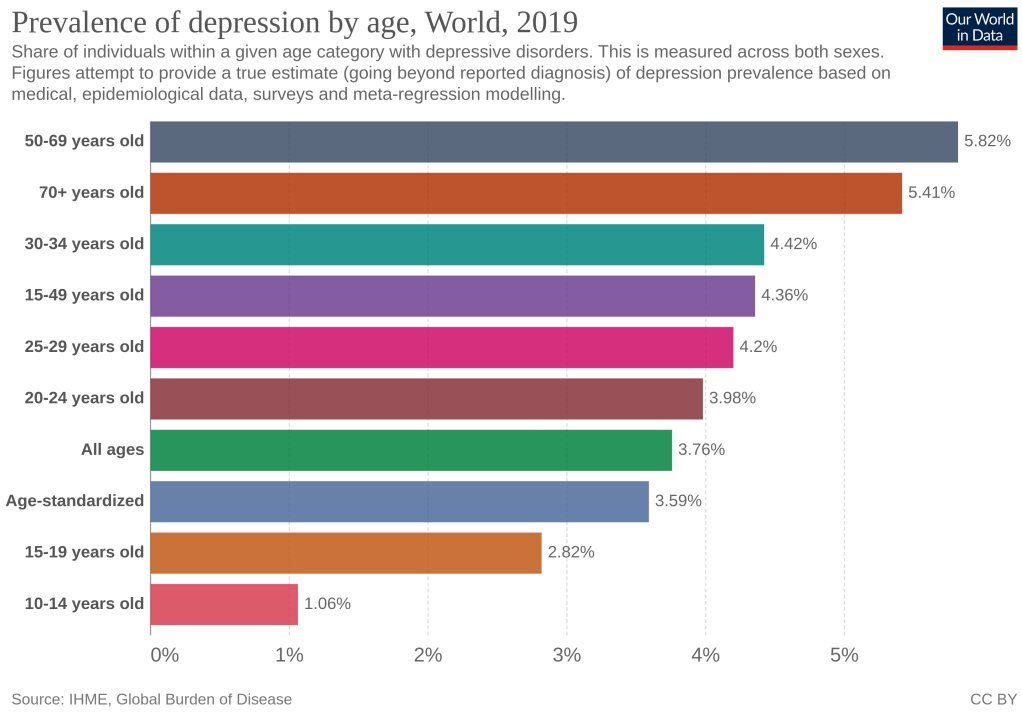

Figura 1.1: Prevalencia da depresión por idade [1].

1.1 Motivación

Existen estudos que apuntan a que as situacións de emerxencia aumentan a prevalencia dos trastornos depresivos na poboación. Estímase que o 22% das persoas que viviron un conflito bélico ou suceso violento nalgún momento dos 10 anos previos padece depresión, ansiedade, trastorno por estrés postraumático, trastorno bipolar ou esquizofrenia [4]. Así mesmo, emerxencias sanitarias como a recente pandemia da COVID-19 tamén supoñen un aumento nos casos de depresión [5, 6]. A maiores, a prevalencia da depresión no grupo de adolescentes e adultos novos segue unha tendencia alcista nos últimos anos [7], dato que resulta máis preocupante se cabe ao tratarse de grupos de idade especialmente vulnerables.

Por fortuna, as investigacións apuntan a que unha adecuada intervención pode diminuir significativamente os efectos dos trastornos depresivos [8]. Máis concretamente, unha detección temperá resulta crucial para minimizar as consecuencias deste tipo de trastornos [9, 10]. A maiores, puido comprobarse a relación existente entre a evolución dos trastornos e o uso que fan da linguaxe as persoas que os padecen [11, 12]. A este efecto, a xa transversal pero aínda a día de hoxe crecente popularidade de Internet en xeral e as redes sociais en concreto, supón unha gran oportunidade. As investigacións amosan que o uso da linguaxe neste contexto ten gran potencial para ser explotado [13], xa que as persoas usan en moitas ocasións

as redes sociais para expresarse e comunicar os seus sentimentos [14].

Deste xeito, a principal motivación do proxecto radica na investigación das posibilidades dos Modelos de Relevancia (ver sección 2.4) para aproveitar os puntos anteriormente expostos.

1.2 Obxectivos

Os obxectivos que se pretenden acadar durante a realización deste proxecto son varios:

Adecuado estudo do problema e posibilidades
Antes de abordar outros obxectivos, é preciso comprender tanto os fundamentos e conceptos básicos que se presentan no capítulo 2 como procurar e investigar as tecnoloxías e ferramentas que se expoñen no capítulo 3. Así pois, o primeiro que se busca neste proxecto é entender os *porqués* e os *comos* das fases que se pretenden completar.

Planificación axustada
Unha vez estudado o problema, e previo a calquera desenvolvemento, cómpre dotar o proxecto dunha boa planificación. Canto máis axustada sexa en orixe, menores serán as desviacións e erros a corrixir.

Cómputo dos Modelos de Relevancia e obtención dos vocabularios de depresión
Logo de planificar, é tempo de desenvolver. O obxectivo neste punto é computar os Modelos de Relevancia, de tal xeito que sexa posible obter vocabularios propios para o ámbito dos trastornos depresivos. Isto, que se explicará con máis detalle en capítulos posteriores, farase a partires de coleccións de texto obtidas para a Early Risk Prediction on the Internet (eRisk)[2].

Comparativa dos vocabularios de depresión con *lexicons* de referencia
Obtidos os vocabularios propios, o obxectivo é realizar unha comparativa entre estes e outros léxicos reputados e validados que foron obtidos no mesmo dominio.

Optimización do *ranking* e comparativa cos baselines
O *lexicon* obtido para o Modelo de Relevancia permite, despois dun proceso de optimización, construir unha ferramenta que ordene os suxeitos por grao de severidade estimada do trastorno depresivo coa información dos seus textos. É un obxectivo, por tanto, optimizar este útil e comparar os seus resultados con valores de referencia.

Análise global e busca de posibles melloras e/ou ampliacións
Finalmente, debe realizarse unha reflexión sobre o traballo realizado. Neste punto, búscanse vías que permitan obter mellores resultados. Isto pode pasar ben por melloras sobre o realizado, ben por ampliacións do traballo.

[2] https://erisk.irlab.org/

1.3 Estrutura da memoria

Esta sección trata de dar unha visión global da propia memoria, describindo brevemente cada unha das partes que a compoñen:

Introdución
Busca contextualizar o presente proxecto, presentando tamén a temática a tratar e definindo a motivación e os obxectivos. Así mesmo, contén a estrutura da memoria e o plan de traballo que se seguiu.

Fundamentos e conceptos básicos
Trata de introducir as bases teóricas sobre as que se sustenta este traballo.

Método
Pretende introducir máis en detalle o tema do proxecto e, ademais, presenta as tarefas realizadas ao longo do proxecto.

Tecnoloxías e ferramentas
Describe e analiza as linguaxes de programación, librarías e ferramentas empregadas no contexto do proxecto, xustificando para cada unha delas a súa utilización.

Desenvolvemento
Expón os detalles de deseño e implementación do software realizado para o proxecto.

Experimentación
Realiza unha descrición das métricas de avaliación utilizadas nos experimentos. Así mesmo, ofrece unha visión en detalle dos experimentos concretos levados a cabo.

Resultados
Presenta os resultados obtidos na experimentación, comentándoos e ofrecendo unha análise crítica. Tamén aporta unha comparación con resultados de referencia.

Metodoloxía e xestión do proxecto
Detalla a metodoloxía seguida durante o transcurso do proxecto, así como a xestión do mesmo. Para isto, xustifícase e preséntase a propia metodoloxía, así como tamén a adaptación realizada sobre a mesma.

Conclusións e traballo futuro
Analiza o cumprimento ou non dos obxectivos inicialmente establecidos, detallando tamén aspectos positivos e negativos. A maiores, introdúcense posibles liñas futuras e ampliacións para o traballo.

Apéndices
Anexa unha serie de seccións adicionais:

- **Relación de acrónimos:** acrónimos presentes na memoria.
- **Glosario:** termos técnicos utilizados.
- **Bibliografía:** referencias bibliográficas consultadas durante o proxecto.

1.4 Plan de traballo

Para poder levar a cabo o proxecto seguíronse as seguintes fases relacionadas coa metodoloxía empregada:

- Estudo da propia metodoloxía, así como conceptos de base necesarios para levar a cabo o proxecto, centrándose principalmente nos modelos de relevancia no ámbito da recuperación de información.
- Análise dos requisitos funcionais e non funcionais para fixar os obxectivos correspondentes.
- Estudo das tecnoloxías e ferramentas dispoñibles, seleccionando aquelas que se consideren óptimas en cada momento.
- Dado que a metodoloxía queda encadrada dentro das denominadas "áxiles" e é incremental, en cada unha das iteracións realizáronse os seguintes pasos:
  - Xestión das tarefas identificadas.
  - Desenvolvemento e implementación das mesmas.
  - Avaliación dos resultados obtidos.
- Documentación do traballo e elaboración da memoria.

Capítulo 2

Fundamentos e conceptos básicos

O obxectivo deste capítulo é definir as bases do marco teórico sobre o que se constrúe este proxecto. Así, na sección 2.1 farase unha breve introdución ao concepto de Recuperación de Información e a súa definición. A continuación, na sección 2.2, explicarase a idea xeral da que é a tarefa principal no eido da Recuperación de Información. De xeito complementario, na sección 2.3 presentarase o concepto de Modelo de Linguaxe, necesarios para a tarefa previamente descrita. Finalmente, a sección 2.4 serve tamén como carta de presentación dos Modelos de Relevancia, extensións estes dos Modelos de Linguaxe.

2.1 Recuperación de Información

A Recuperación de Información, en inglés *Information Retrieval (IR)*, é un campo das Ciencias da Computación que busca satisfacer as necesidades de información das persoas usuarias [15, 16]. De xeito máis formal, pode utilizarse a seguinte definición:

> A Recuperación de Información trata coa representación, almacenamento, organización e acceso a elementos de información tales como documentos, páxinas web, catálogos, rexistros estruturados ou semi-estruturados e obxectos multimedia. A representación e organización dos elementos de información debe ser tal que lle permita ás persoas usuarias un acceso sinxelo á información que sexa do seu interese [15].

A IR xa era un campo plenamente asentado con anterioridade, pero foi sen dúbida a aparición da World Wide Web (WWW) o factor que fixo que a súa importancia aumentara de xeito expoñencial. Así pois, o gran aumento do volume de información derivou no desenvolvemento de novos modelos de recuperación, capaces de adecuarse ás novas necesidades de información. Tanto é así que, entre os múltiples sistemas de IR existentes, os motores de busca para *web* erixíronse coma o exemplo máis prominente.

2.2 Recuperación *ad hoc*

A recuperación *ad hoc* é a tarefa principal e, por tanto, máis estudada no ámbito da IR [16]. Esta tarefa realízase sobre unha colección de documentos (unidades de recuperación) previamente indexada para crear a estrutura de información denominada *índice invertido*. O obxectivo será entón atopar aqueles documentos da colección que sexan relevantes[1] para a necesidade de información da persoa usuaria, que tipicamente se presenta como unha breve descrición textual denominada *query* ou consulta. O motor de recuperación procesará entón a *query* contra a colección utilizando un modelo de recuperación e obterá un ***ranking*** ou ordenación dos documentos.

Os modelos de recuperación *ad hoc* deben producir un *ranking* de documentos presentándoos en orde decrecente de relevancia estimada. Para levar a cabo esta tarefa, é necesario comparar a representación do documento coa representación da necesidade de información. Existen múltiples modelos matemáticos propostos para conseguir o devandito *ranking* de documentos en función dunha determinada *query*. Son exemplos o Modelo *Booleano* [17], o Modelo de Espazo Vectorial [18] ou os modelos probabilísticos coma o Modelo de Independencia Binaria [19] e os Modelos de Linguaxe [20, 21]. Na seguinte sección (2.3), abordaranse estes últimos.

2.3 Modelos de Linguaxe

Os Modelos de Linguaxe, en inglés Language Model (LM), son modelos probabilísticos de recuperación cunha base estatística sólida. Todos eles seguen o Probability Ranking Principle, que se define formalmente do seguinte xeito:

> Se a resposta dun sistema de recuperación a cada petición é un *ranking* dos documentos na colección en orde decrecente de probabilidade de seren útiles para a persoa usuaria, onde as probabilidades son estimadas de xeito tan preciso como sexa posible en base á información dispoñible que o sistema teña para este propósito, entón a eficacia do sistema para as persoas usuarias será a máxima obtible para a información da que se dispón [22].

Neste tipo de modelos, as ocorrencias das palabras nos documentos e as *queries* vense coma un proceso de xeneración aleatorio, tipicamente usando unha distribución multinomial. Deste xeito, é posible inferir un LM para cada documento da colección. Para ordenar os documentos de acordo a unha determinada *query*, haberá que estimar a probabilidade D de cada documento dada a *query* Q, de acordo ao amosado na ecuación 2.1:

[1] Un documento é considerado relevante cando resulta valioso en relación á necesidade de información.

$$p\left(D|Q\right) = \frac{p\left(Q|D\right) \cdot p\left(D\right)}{p\left(Q\right)} \overset{rank}{=} p\left(Q|D\right) \cdot p\left(D\right) \tag{2.1}$$

onde $p\left(Q|D\right)$ representa o *query likelihood* e $p\left(D\right)$ o *prior* do documento. Ignorando o *prior* $p\left(Q\right)$ da *query* e supoñendo uniforme o *prior* do documento, tan só hai que computar $p\left(Q|D\right)$ para obter o *ranking* de documentos para a *query*. Deste xeito, estase a asumir que a *query* é unha mostra do modelo de linguaxe do documento, polo que para obter a puntuación dun documento tan só se computará o *query likelihood* dado o modelo de linguaxe do documento. Usando un modelo de unigramas, a formulación resulta tal e como se pode observar na ecuación 2.2:

$$P(Q|D) = \prod_{i=1}^{n} P(q_i|D) \tag{2.2}$$

onde q_i é un termo da *query* e n o número de termos na *query*. A probabilidade condicionada $P(q_i|D)$ pode computarse mediante a Maximum Likelihood Estimate (MLE) dunha distribución multinomial, do xeito que se ilustra na ecuación 2.3:

$$P(q_i|D) = \frac{f_{q_i,D}}{|D|} \tag{2.3}$$

onde $f_{q_i,D}$ é o número de veces que o termo q_i aparece no documento D, e $|D|$ é o número de termos en D. O problema desta estimación radica no feito de que cando algún dos termos da *query* non está presente no documento, a probabilidade $P(Q|D)$ será cero. Adicionalmente, dado que se está a contruír un modelo de linguaxe, termos relacionados co tema deben ter unha probabilidade asociada, aínda que non se mencionen explicitamente no documento.

Ante esta problemática aparece o concepto de *smoothing*[23], que non é máis que unha técnica para estimar a probabilidade dos termos non presentes nos documentos. Para conseguir isto, debe descontarse unha masa de probabilidade dos termos *si* vistos nos documentos, sendo despois esta masa repartida entre os termos *non* vistos. Así, as súas probabilidades xa non serán cero. Ao aplicar *smoothing*, a estimación de probabilidades dos termos realízase do seguinte xeito:

- Para termos *non* vistos, o estimado será o resultante da ecuación 2.4, onde $P(q_i|C)$ é a probabilidade para o termo i da *query* no modelo de linguaxe para a colección C e α_D se corresponde cun parámetro.

$$\alpha_D P(q_i|C) \tag{2.4}$$

- Para termos *si* vistos, o estimado calcúlase segundo a ecuación 2.5, onde de novo α_D se corresponde cun parámetro, $P(q_i|D)$ coa probabilidade para o termo i da *query* no

documento e $P(q_i|C)$ coa probabilidade para o termo no modelo de linguaxe para a colección C.

$$(1 - \alpha_D)P(q_i|D) + \alpha_D P(q_i|C) \tag{2.5}$$

Este parámetro α_D pode calcularse de diversas maneiras, dando así lugar aos diferentes métodos de *smoothing*.

Jelinek-Mercer

No caso de Jelinek-Mercer, o parámetro de suavización fíxase segundo a ecuación 2.6. Como se pode ver, utilízase un valor constante λ, que tipicamente oscila entre 0 e 1. A probabilidade estimada do modelo de linguaxe da colección para o termo q_i é $c_{q_i}/|C|$, onde c_{q_i} se corresponde co número de veces que un termo da *query* aparece na colección de documentos. Á súa vez, $|C|$ representa o número total de ocorrencias do termo na colección. Complementariamente, defínese $f_{q_i,D}$ como o número de veces que o termo q_i ocorre no documento D, sendo $|D|$ o número de termos en D. Isto, todo xunto, resulta no estimado para $P(q_i|D)$ que se amosa na ecuación 2.7.

$$\alpha_D = \lambda \tag{2.6}$$

$$P(q_i|D) = (1 - \lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|} \tag{2.7}$$

Dirichlet

Para Dirichlet, o parámetro de suavización establécese segundo a ecuación 2.8. Habitualmente, o valor de μ vai desde 0 ata 5000. Tendo en conta as consideracións xa explicadas no anterior caso, agora o estimado para $P(q_i|D)$ será o da ecuación 2.9.

$$\alpha_D = \frac{\mu}{|D| + \mu} \tag{2.8}$$

$$P(q_i|D) = \frac{f_{q_i,D} + \mu\frac{c_{q_i}}{|C|}}{|D| + \mu} \tag{2.9}$$

Independentemente do tipo de *smoothing* utilizado, os *search engines* realizan transformacións logarítmicas para atacar os problemas de precisión que xorden ao multiplicar números pequenos. A este respecto, as ecuacións 2.10 e 2.11 amosan as transformacións utlizadas para Jelinek-Mercer e Dirichlet, respectivamente.

$$\log P(Q|D) = \sum_{i=1}^{n} \log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|}) \tag{2.10}$$

$$\log P(Q|D) = \sum_{i=1}^{n} \log \frac{f_{q_i,D} + \mu\frac{c_{q_i}}{|C|}}{|D| + \mu} \tag{2.11}$$

2.4 Modelos de Relevancia

Os Modelos de Linguaxe baseados na relevancia, en inglés *Relevance-Based Language Models (RM)* e comunmente abreviados como *Relevance Models*, foron concebidos co obxectivo de introducir explicitamente o concepto de *relevancia*, intrínseca aos modelos probabilísticos, nos Modelos de Linguaxe [24]. Nun RM, en tanto que é un LM, a *query* orixinal é considerada unha mostra de palabras obtidas do propio RM (R). De querer obter máis palabras de R, resulta razonable escoller aquelas con maior probabilidade estimada ao considerar as palabras para a distribución xa vista. Deste xeito, os termos do *lexicon* da colección son ordenados en función da súa probabilidade estimada, que baixo RM1 se calcula segundo a formulación que se presenta na ecuación 2.12:

$$p\left(w|R\right) = \sum_{D \epsilon C} \cdot p\left(w|D\right) \cdot p\left(D\right) \cdot \prod_{i=1}^{n} p\left(q_i|D\right) = \sum_{d \epsilon C} \cdot p\left(w|D\right) \cdot \prod_{i=1}^{n} p\left(q_i|D\right) \tag{2.12}$$

onde o *prior* $p\left(d\right)$ do documento se supón uniforme, $\prod_{i=1}^{n} p\left(q_i|D\right)$ é a denominada *query likelihood* dado o modelo do documento e $p\left(w|D\right)$ a probabilidade do termos dado un documento D. Estas dúas últimas probabilidades é habitual computalas facendo uso de métodos de suavización [23]. Tendo en conta isto, para o proceso completo de *ranking*, deben seguirse catro pasos:

1. Ordenar os documentos da colección C segundo a súa *query likelihood*, sobre a que tipicamente se aplica un método de suavización de probabilidades.

2. Seleccionar dos documentos recuperados, en lugar da colección C enteira, un certo *top* r que denominaremos *pseudo-relevance set* RS.

3. Calcular as probabilidades $p\left(w|R\right)$ do modelo utilizando a ecuación 2.12 con RS en lugar de C.

4. Seleccionar os e termos con maior probabilidade $p\left(w|R\right)$ estimada para construír con eles a *query* expandida.

Posteriormente xurdiron novas extensións para os RM. Un exemplo paradigmático é Relevance Model 3 (RM3) [25], que interpola os termos seleccionados por RM1 coa *query* orixinal tal e como se amosa na ecuación 2.13.

$$p\left(w|q'\right) = (1-\lambda) \cdot p\left(w|q\right) + \lambda \cdot p\left(w|R\right) \tag{2.13}$$

Os RM foron introducidos no eido da IR probabilística como un método para facer *pseudo relevance feedback*. Neste TFG desenvólvese unha aplicación novedosa dos mesmos, estimando cos RM modelos de depresión e construindo un *ranking* de usuarias con depresión.

Capítulo 3

Tecnoloxías e ferramentas

Cando obter un resultado de calidade e completar con éxito o traballo plantexado é unha prioridade, resulta de gran importancia adicar tempo á escolla e estudo das tecnoloxías e ferramentas dispoñibles. Para iso, na sección 3.1 describiranse as linguaxes de programación empregadas. A continuación, na sección 3.2, farase o propio coas librarías e compoñentes. Tamén se detallarán as ferramentas de desenvolvemento seleccionadas na sección 3.3. Por último, adicarase a sección 3.4 á explicación das ferramentas de soporte.

3.1 Linguaxes de programación

O código implementado neste proxecto pode dividirse en dúas categorías ben diferenciadas. Por unha banda, estarían os *scripts* para tarefas concretas, que se desenvolveron en Python. Para as aplicacións responsables da indexación das coleccións, o cómputo do modelo de relevancia e o *ranking* utilizouse Java. A continuación, preséntanse estas linguaxes de programación.

3.1.1 Python

Python[1] é unha linguaxe interpretada de alto nivel que incorpora módulos, excepcións, tipado dinámico e clases. Soporta múltiples paradigmas de programación, podendo ser usada para programación orientada a obxectos, funcional, imperativa, etc. Un dos seus puntos fortes radica na súa gran potencia combinada cunha sintaxe limpa e sinxela. Conta tamén con interfaces para efectuar chamadas ao sistema e librarías. Así mesmo, cabe destacar a portabilidade desta linguaxe, que pode ser executada tanto en Windows coma en múltiples variantes Unix, incluíndo Linux e macOS.

Como punto negativo, o feito de tratarse dunha linguaxe interpretada provoca que o código se execute máis lento en comparación a se fora escrito nunha linguaxe compilada, como por

[1] https://docs.python.org/3/faq/general.html

exemplo C. Non obstante, as tarefas para as que su utliza Python neste proxecto son puntuais e non críticas, polo que a característica anteriormente descrita non afecta á hora de escoller a linguaxe. Ademais, considérase que esta desvantaxe queda amplamente compensada por outros dous factores: a rapidez coa que se desenvolve o código e o feito de ser a linguaxe máis popular na actualidade[26], tal e como se pode ver na figura 3.1. Resulta importante a popularidade porque implica que a maioría de librarías e extensións serán compatibles con esta linguaxe.

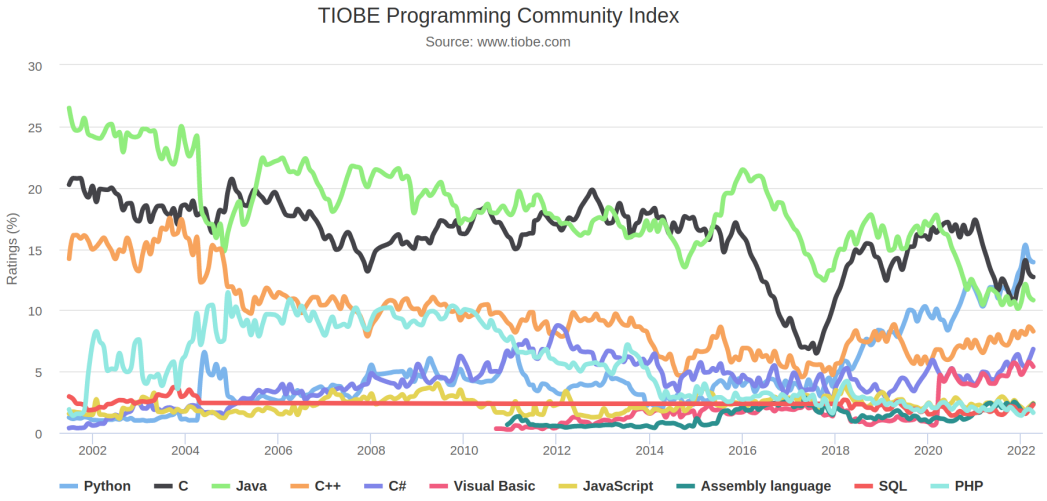

Figura 3.1: Índice TIOBE de popularidade.

3.1.2 Java

Java[2] é unha linguaxe de alto nivel de propósito xeral orientada a obxectos e baseada en clases. O seu principio fundamental é que o código se compila a un bytecode que pode ser executado en calquera JVM, independentemente da arquitectura da máquina. Durante moitos anos foi a linguaxe máis popular, polo que a súa valía está máis que avalada. En concreto, utilízase neste proxecto por ser a linguaxe para a que está escrita a libraría Apache Lucene, que se detallará no punto 3.2.3.

[2] https://www.java.com/en/download/help/whatis_java.html

3.2 Librarías e compoñentes

3.2.1 lxml

lxml[3] é un toolkit para o procesado de ficheiros XML e HTML empaquetado para Python. Dado que encapsula as librarías libxml2[4] e libxslt[5] orixinalmente escritas para C, o seu punto forte radica en combinar a velocidade e completitude destas librarías coa simplicidade de Python. Ademais, está distribuída baixo licenza BSD de software libre.

O seu funcionamento consiste en construir unha árbore co contido dos documentos, de modo que se poida extraer a información. No contexto deste proxecto, utilízase para procesar as coleccións eRisk[6] de cara á fase de indexación.

3.2.2 PRAW

Python Reddit API Wrapper[7] é un paquete para Python que facilita o acceso á API de Reddit[8]. O seu obxectivo é ser o máis sinxelo de utilizar posible, ademais de estar deseñado para seguir todas as regras da API que permite acceder. A maiores, cabe indicar que está distribuído baixo licenza GPL de software libre.

Neste proxecto, emprégase para facer scrapping de publicacións e comentarios en Reddit[9], podendo construir así unha colección complementaria ás eRisk.

3.2.3 Apache Lucene

Apache Lucene[10] é unha libraría escrita en Java e distribuída baixo licenza Apache de software libre co propósito de ofrecer servizos de indexación e, de xeito complementario, busca de documentos. Actualmente poden atoparse encapsulacións para outras linguaxes, como por exemplo PyLucene[11]. É importante poñer énfase no termo "encapsulación" en contraposición ao que se denominaría *port*, xa que deste modo o que ocorre é que se crea unha capa escrita noutra linguaxe por enriba do núcleo de Lucene, que continúa a executarse en Java sobre unha JVM. Como se comentou con anterioridade, este foi un factor decisivo á hora de decidir desenvolver as aplicacións en Java.

[3] https://lxml.de/index.html
[4] https://gitlab.gnome.org/GNOME/libxml2/-/wikis/home
[5] https://gitlab.gnome.org/GNOME/libxslt/-/wikis/home
[6] https://erisk.irlab.org/
[7] https://praw.readthedocs.io/
[8] https://www.reddit.com/dev/api/
[9] https://www.reddit.com/
[10] https://lucene.apache.org/
[11] https://lucene.apache.org/pylucene/

3.3 Ferramentas de desenvolvemento

3.3.1 Eclipse IDE

Eclipse[12] é un IDE que, se ben a día de hoxe está a perder cota de mercado fronte a outros competidores, continúa a ser amplamente usado para o desenvolvemento de aplicacións en Java, aínda que tamén soporta o desenvolvemento de aplicacións escritas noutras linguaxes. Conta con todas as características habituais que se esperan dun IDE (plugins, refactorizado, depuración, integración Git (ver 3.4.3), macros, etc.).

Por ser unha ferramenta coñecida polo autor e por atoparse dispoñible baixo licenza EPL de software libre foi seleccionado como entorno para o desenvolvemento Java de todo o proxecto.

3.3.2 Visual Studio Code

Visual Studio Code[13] é un editor de ficheiros fonte lixeiro desenvolto por Microsoft e dispoñible para Windows, Linux e macOS. Entre as súas características, inclúese a posibilidade de depurar o código, o autocompletado, as macros, facilidades para refactorización, etc. Ademais, permite a instalación de extensión e é moi personalizable tanto en apariencia coma en comportamento. Por todo isto, entre outros motivos, actualmente sitúase como a ferramenta de desenvolvemento máis popular, cunha cota de uso do 70%[27]. O código fonte atópase dispoñible baixo licenza MIT, mentres que as versións liberadas por Microsoft son de tipo propietario.

Polos motivos anteriormente expostos, ademais de ser unha ferramenta xa coñecida polo autor, é o editor empregado para todo o desenvolvemento en Python deste proxecto.

3.4 Ferramentas de soporte

3.4.1 Xestión de paquetes

Pip

Pip[14] é un sistema de xestión de paquetes empregado para instalar e administrar paquetes de Python, moitos deles dispoñibles no Python Package Index. Unha vantaxe de Pip é que permite xestionar ditos paquetes de maneira moi sinxela mediante a súa interfaz de liña de comandos. Outra característica de interese é que permite controlar listas de paquetes e as súas versións mediante un ficheiro de requisitos, permitindo unha recreación eficaz dun conxunto

[12] https://www.eclipse.org/ide/
[13] https://code.visualstudio.com/
[14] https://pip.pypa.io/en/stable/

de paquetes nun novo entorno. Por último, destacar que está distribuido baixo licenza MIT de software libre.

venv

venv[15] é un modulo que proporciona acceso á creación de entornos virtuais lixeiros, cos seus propios directorios opcionalmente illados dos directorios do sistema. Cada entorno virtual ten o seu propio Python e pode ter o seu propio conxunto independente de paquetes. Este módulo forma parte da libraría estándar de Python.

Apache Maven

Apache Maven[16] pretende estandarizar a construción de proxectos, ao tempo que busca automatizar a compilación, execución de probas e despregue da aplicación, entre outros obxectivos. Un dos seus aspectos máis destacados, e o motivo polo que se utiliza neste proxecto, é a simplificación do manexo de paquetes en aplicacións Java. Para isto faise uso dun ficheiro XML denominado POM, onde se detalla a carga de módulos e dependencias, así como a orde de construción. No contexto deste proxecto, utilízase para controlar as dependencias das aplicacións Java coa libraría Apache Lucene (ver 3.2.3). Por último, sinalar que esta ferramenta se atopa dispoñible baixo licenza Apache de software libre.

3.4.2 Xestión do proxecto

Un desenvolvemento software de calidade implica unha boa planificación e un seguemento adecuado. Estas necesidades poden ser cubertas coa ferramenta que a continuación se detalla.

Taiga

Taiga[17] é unha ferramenta para a xestión de proxectos encadrados dentro das metodoloxías denominadas áxiles: Scrum e Kanban. A súa principal funcionalidade é a xestión de tarefas, xa que permite crear as historias de usuario, a estimación de puntos de historia e a asignación das mesmas aos diferentes *sprints*. Todo isto facilita o seguemento do proxecto e dá unha visión global tanto dos puntos completados coma dos restantes ata a finalización en cada un dos *sprints*. A maiores, ofrece integración con GitLab (ver 3.4.3), o que unido a que está deseñada para ser empregada coa metodoloxía seguida neste proxecto (ver 6.1) fixo que fora a ferramenta seleccionada para a xestión do proxecto. Destacar, finalmente, que está distribuída baixo licenza Mozilla de software libre e código aberto.

[15] https://docs.python.org/3/library/venv.html
[16] https://maven.apache.org/
[17] https://www.taiga.io/

3.4.3 Control de versións

Git e GitLab

Git[18] é un sistema de control de versións distribuído deseñado por Linus Torvalds para o manter o desenvolvemento do kernel de Linux. Está deseñado para traballar tanto con pequenas aplicacións como con grandes proxectos, sendo a facilidade para o uso e o rendemento dous dos seus piares fundamentais.

O ser estritos no control de versións permite ter unha correcta trazabilidade nos cambios e evolución do proxecto, de modo que é requisito indispensable para un desenvolvemento de calidade. Fronte a outros sistemas de control de versións, coma SVN[19], Git ofrece certas vantaxes que o levaron a ser o SCM de referencia:

- Posibilidade de traballar sen conexión a internet: Git funciona con repositorios locais e remotos, o que permite efectuar o control de versións desconectados da rede e conservar en local toda a funcionalidade.
- Desenvolvemento non lineal: Git potencia o uso do branching, facilitando o traballo dos equipos de desenvolvemento e maximizando a produtividade.

Para xestionar os repositorios Git, existen dúas plataformas amplamente asentadas: GitLab[20] e GitHub[21]. Para este proxecto, faise uso dunha conta en GitLab facilitada polo IRLab[22], laboratorio no cal se enmarca o desenvolvemento deste proxecto.

[18] https://git-scm.com/
[19] https://subversion.apache.org/
[20] https://about.gitlab.com/
[21] https://github.com/
[22] https://www.dc.fi.udc.es/irlab/

Capítulo 4

Método

4.1 Iniciativa eRisk e coleccións

CLEF eRisk [28, 29, 30, 31, 32] é unha iniciativa organizada co obxectivo de propoñer metodoloxías e métricas e avaliar técnicas e algoritmos para a detección temperá de riscos en internet. Céntrase en riscos relacionados coa saúde como poden ser a depresión, os trastornos alimentarios ou o dano autoinflixido. Con este propósito, cada edición realízase unha proposta de tarefas e libéranse coleccións de textos escritos por persoas nas redes sociais. De cara a coñecer máis en detalle a distribución das tarefas propostas para cada unha das edicións celebradas, cómpre revisar a Táboa 4.1. É importante saber que, a efectos deste proxecto, unicamente se utilizan os conxuntos de datos asociados ás tarefas de detección e estimación da severidade de casos de depresión.

Tarefas	**2017**[1]	**2018**[2]	**2019**[3]	**2020**[4]	**2021**[5]
Detección de depresión	X	X			
Anorexia		X	X		
Comportamentos autolesivos			X	X	X
Severidade dos signos de depresión			X	X	X
Ludopatía					X

Táboa 4.1: Distribución da proposta de tarefas ao longo das edicións de eRisk.

Estas coleccións de depresión teñen dúas naturezas ben diferenciadas. Para as dúas primeiras edicións da iniciativa o obxectivo era a detección dos casos de depresión. Deste xeito, as coleccións son "binarias", é dicir, para cada suxeito do conxunto de datos só hai información de se padecía depresión ou non. A partires de 2019, introduciuse a tarefa de estimación do índice de depresión. Tendo en conta isto, queda de manifesto a necesidade dunha variación

no tipo dos *datasets*. Por tanto, para as edicións de 2019, 2020 e 2021 as coleccións elaboráronse en colaboración con persoas usuarias do foro *Reddit*[6], concretamente usuarias do subforo de saúde mental. Voluntariamente, estas persoas ofrecéronse a responder o inventario de depresión BDI-II [33] (ver Apéndice B), que permite avaliar a gravidade da sintomatoloxía depresiva. Con esta información, xunto coa recompilación de todos os *posts* destas persoas no foro, é posible construir as devanditas coleccións para a tarefa de estimación do índice de depresión.

As coleccións, independentemente da súa natureza, gardan un formato moi semellante. Están formadas por un ficheiro por cada unha das persoas participantes, no que se recompilan todos os *posts* en *Reddit* desa persoa, entendendo por *post* tanto as publicacións propias coma os comentarios nas publicacións doutras persoas. A información aparece etiquetada segundo o caso concreto, é dicir, se ben todos os *posts* teñen unha data asociada, tan só as publicacións teñen título, por exemplo. A maiores, existe un ficheiro de *golden truth* no que se especifica, para as coleccións de 2017 e 2018, se a persoa padece depresión ou non. Para as coleccións de 2019, 2020 e 2021 ofrécense detalladas todas as respostas ao inventario BDI-II. Deste xeito, a Táboa 4.2 recolle unha serie de estatísticas para todas elas, tratando de aportar información que permita coñecer mellor o conxunto de datos.

	2017	**2018**	**2019**	**2020**	**2021**
Suxeitos con depresión	135	79			
Suxeitos non deprimidos	752	741			
Número total de suxeitos	887	820	20	70	80
Número total de posts	531294	544447	10941	35562	32237
Media de posts	599	663	547	508	402
Mínimo número de posts	10	9	29	16	19
Máximo número de posts	2000	1999	1510	1355	1218

Táboa 4.2: Resumo das estatísticas para as coleccións CLEF eRisk empregadas.

4.2 *Lexicons*

O termo *lexicon* xa se introduciu con anterioridade, mais aínda non se definira claramente cales son os *lexicons* usados na experimentación deste proxecto e cales son as súas características. En concreto, preséntanse nos Puntos 4.2.1 e 4.2.2 dous *lexicons* posteriormente com-

[6] https://www.reddit.com/

parados e aumentados por Losada e Gamallo [34] mediante diversas técnicas de expansión, nomeadamente as denominadas *WordNet* e *Distribucional*.

4.2.1 Pedesis

O *lexicon* aquí denominado *Pedesis* [35] definiuse nun contexto de cribado proactivo e automático dos trastornos depresivos. En concreto, fíxose unha recollida na *Web* na procura de relacións metafóricas nas cales a depresión está implícita, estraendo aqueles dominios conceptuais que a describen. Esta información é usada por un grupo de persoas expertas para a elaboración do léxico de depresión, que puido ser usado para detectar indicios depresivos en textos, así como para avaliar automaticamente o nivel de gravidade.

4.2.2 Choudhury

Choudhury [36], que recibe o seu nome do autor principal da investigación na que xorde, é o segundo dos *lexicons* utilizados neste proxecto. A investigación asociada explora o potencial uso das redes sociais para a detección e o diagnóstico de trastornos depresivos en suxeitos individuais. Primeiramente, selecciona un conxunto de persoas usuarias de Twitter[7] que reportan ter sido diagnosticadas clinicamente con depresión. Para elas, mídense atributos de comportamento relativos á súa interacción na propia rede, tales como as emocións, o uso e estilo da linguaxe, etc. Obtívose, por tanto, etiquetando as palabras e termos máis relevantes entre suxeitos depresivos a través das súas contas de Twitter.

Ambos os dous *lexicons* presentan os seus termos (tanto na forma orixinal como na expansión) categorizados por substantivos, adxectivos e verbos. Para este proxecto, utilízanse todos os termos orixinais e tamén os adxectivos por separado, tanto os orixinais coma os expandidos. A Táboa 4.3 recolle información co número de termos de cada unha destas categorizacións usadas, ademais de referenciar os apéndices onde se poden consultar ao completo.

	Pedesis	**Choudhury**
Adexctivos únicos non ambiguos orixinais	153 (ver A.4)	7 (ver A.3)
Axd. únicos non amb. expandidos Dist	312 (ver A.7)	13 (ver A.5)
Adx. únicos non amb. expandidos WN	549 (ver A.8)	16 (ver A.6)
Termos únicos totais orixinais	636 (ver A.1	106 (ver A.2)

Táboa 4.3: Detalle das categorizacións usadas dos *lexicons*.

[7] https://twitter.com

4.3 Tarefas

4.3.1 Tarefa 1 - Obtención dos vocabularios de depresión

De maneira xeral, o obxectivo que se persegue nesta tarefa é poder obter os vocabularios de depresión a partires das coleccións coas que se traballa, asociando cada un dos termos co seu peso. Unha vez obtidos, será posible comparalos con outros *lexicons* do ámbito dos trastornos depresivos, tales como os presentados nos Puntos 4.2.1 e 4.2.2. Para isto, reformúlase o introducido na Sección 2.4 ata obter a ecuación 4.1:

$$P(w|R) = \sum_{d \epsilon F} P(w|d) BDI(d) \tag{4.1}$$

Esta ecuación, que modifica o cálculo estándar do RM, ten en conta os seguintes aspectos:

- O *prior* ou preferencia $P(d)$ dun documento será o mesmo para todos, polo que se pode considerar $1/|C|$, onde $|C|$ é o número de documentos da colección. Dada esta característica, pode ser eliminado da ecuación.
- Na formulación que se presenta, o produtorio correspóndese coa denominada *Query-Likelyhood* ou, o que é o mesmo, coa probabilidade de xerar a *query* co modelo de linguaxe dos documentos. De cara a este proxecto, este valor é substituído nos cálculos pola puntuación *BDI-II* asociada a cada un dos documentos. Isto pretende potenciar aqueles documentos escritos por persoas que efectivamente padecen un trastorno depresivo.

Por tanto, será esta a formulación que se utilize no cómputo dos RMs para pesar os termos dos vocabularios de depresión. Así pois, a probabilidade de observar unha palabra na colección será o sumatorio do, para todos os documentos da propia colección, produto da probabilidade[8] da palabra dado o documento pola puntuación *BDI-II* asociada. Tendo en conta isto, deben computarse as seguintes series de RMs, cada un deles co seu vocabulario de depresión asociado:

- **Grupo A:** Suavizacións Dirichlet no rango [0, 5000] con salto 500 e Jelinek-Mercer no rango (0, 1] con salto 0.2. Todas sobre as coleccións CLEF eRisk 2019, 2020 e 2021.
- **Grupo B:** Suavizacións Dirichlet no rango [0, 5000] con salto 500 e Jelinek-Mercer no rango (0, 1] con salto 0.2. Todas sobre as coleccións CLEF eRisk 2019, 2020 e 2021, usando tamén as coleccións binarias de 2017 e 2018 para a probabilidade de fondo dos modelos.

[8] Sobre a que se aplica unha suavización [23], como pode ser a de Dirichlet ou Jelinek-Mercer.

- **Grupo C:** Suavizacións Dirichlet no rango [2500, 5000] con salto 500 sobre un conxunto de datos propio.

4.3.2 Tarefa 2 - Análise dos vocabularios de depresión

Superada a Tarefa 1, téñense en disposición unha serie de vocabularios de depresión, recordando que un vocabulario non é máis que unha lista de termos cos seus pesos asociados. Cómpre incidir en que cada que un dos RMs computados cos diferentes valores de suavización (ver Punto 4.3.1) ten o seu propio vocabulario de depresión. Así, para cada un deles, realizaranse análises tanto individuais coma comparativas. Son as correspondentes cos Puntos 5.2.1 e 5.2.2 da Sección 5.2 de Experimentos.

4.3.3 Tarefa 3 - *Ranking*

A derradeira tarefa consiste na elaboración e avaliación de *rankings*. Cómpre pormenorizar que se entende por *ranking* no contexto deste proxecto e cales son os pasos que se deben dar. De maneira xeral, existirá unha *query*, é dicir, unha consulta que debe ser lanzada contra as indexacións das coleccións, obtendo unha ordenación dos documentos en función da relevancia que se lles estima. No marco deste proxecto, isto permitiría ordenar unha serie de persoas polo seu grao estimado de depresión en función dos seus textos escritos. Os experimentos concretos que se realizaron están detallados nos Puntos 5.2.3, 5.2.4, 5.2.5, 5.2.6 e 5.2.7 da Sección 5.2 de Experimentos.

Capítulo 5

Experimentación e resultados

Unha vez obtidos os vocabularios de depresión dos RMs, xa se está en disposición de realizar experimentos con eles. Estes experimentos están detallados na Sección 5.2. De cara á avaliación dos resultados de *ranking*, deben definirse unha serie de métricas que permitan emitir valoracións sobre os resultados observados. Estas preséntanse na Sección 5.1.

5.1 Metodoloxía de avaliación do *ranking*

Para medir a eficacia de calquera solución proposta cómpre facer uso de métricas que permitan avaliar os resultados, favorecendo tamén a comparación dos mesmos. No caso concreto deste proxecto, o que se pretende medir é a eficacia dun *ranking* de persoas usuarias de *Reddit* no que estas son ordenadas tratando de atopar vinculacións con trastornos de tipo depresivo. O que se está a facer, por tanto, é construír unha consulta que simule a necesidade de información "*que persoas están deprimidas?*". Isto garda relación cun exemplo clásico de IR, onde se busca medir se o conxunto de documentos recuperados como resposta a unha necesidade de información son ou non relevantes para esta. No contexto deste proxecto, e seguindo coa analoxía, son *relevantes* aquelas persoas que padecen depresión. Isto resulta importante porque son estes xuízos de valor os que permiten efectuar as medidas que a continuación se describen.

5.1.1 *P@n*

A precisión defínese como a proporción de documentos recuperados que son relevantes para a persoa usuaria [37]. Se se define Rec como o conxunto de documentos (*persoas*) recuperados e R como o conxunto de documentos que son relevantes (*deprimidas*), a formulación é a da ecuación 5.1.

$$P = \frac{|R|}{|Rec|} \tag{5.1}$$

Adoita ocorrer que, de cara á avaliación, non interesa ver a totalidade do *ranking*, senón que é preferible establecer un corte n para efectuar a avaliación nese *top*. Así, pode medirse a precisión cando só se teñen en conta os primeiros 5, 10 ou, en xeral, n resultados. Para isto, debe aplicarse a ecuación 5.2.

$$P@n = \frac{|R_n|}{n} \tag{5.2}$$

5.1.2 *AP@n*

A Precisión Media, *Average Precision* en inglés, é outra métrica amplamente usada. Defínese como unha medida que combina o *recall* (exhaustividade) e a precisión [37], favorecendo os *rankings* nos que os documentos relevantes aparecen nas primeiras posicións. De acordo coa ecuación 5.3, calcúlase dividindo entre o total de relevantes $|R|$ o sumatorio das precisións (5.2) cada vez que aparece na posición r do *ranking* un documento relevante.

$$AP = \frac{\sum_r P@r}{|R|} \tag{5.3}$$

Ao igual que xa se viu previamente para a precisión, é interesante poder establecer un valor de corte n para o cómputo da métrica, de tal xeito que só se consideren os *top* n documentos recuperados. Neste caso, debe terse en conta que existen diferentes alternativas para a normalización. Para este proxecto, decidiuse normalizar polo total de relevantes $|R|$, de modo que a formulación resulta na ecuación 5.4.

$$AP@n = \frac{\sum_{r=1}^{n} P@r}{|R|} \tag{5.4}$$

5.2 Experimentos

Defínense como experimentos todos aqueles exercicios realizados sobre os vocabularios de depresión dos RMs. Estes experimentos poden ser de diversa índole: análise individual do vocabulario, comparativa cunha referencia ou elaboración e avaliación do *ranking* a partir dos vocabularios. Defínense todos os experimentos realizados nos sucesivos puntos.

5.2.1 Ocorrencias das familias de pronomes persoais nos vocabularios de depresión

Segundo as investigacións de Ortega-Mendoza et al. [38], as frases de carácter persoal representan unha poderosa fonte de información de cara á discriminación de trazos depresivos no contexto das redes sociais. Deste xeito, seleccionar e pesar os termos tendo en conta a súa ocorrencia en afirmacións persoais resulta moi útil para a detección de trastornos depresivos. Isto significa que, segundo esta hipótese, as persoas que padecen depresión tenden a revelar este tipo de trastornos nas oracións nas que fan uso da primeira persoa pronominal. Para ver se esta característica se mostra presente nos vocabularios de depresión obtidos neste proxecto, contáronse as ocorrencias de todas as familias de pronomes dentro do *top* 20 dos seus termos. As familias de pronomes son as seguintes cinco:

- ***I - family*:** I, me, my, mine, myself.
- ***You - family*:** You, your, yours, yourself, yourselves.
- ***He/She - family*:** He, she, him, her, his, hers, himself, herself.
- ***We - family*:** We, us, our, ours, ourselves.
- ***They - family*:** They, them, their, theirs, themselves.

A Táboa 5.1 recolle todos os resultados en detalle. As súas filas correspóndense con todos os RMs computados (en función da suavización utilizada). A maiores, preséntase a categorización por grupos *A - B - C* definida no Punto 4.3.1. As columnas, pola súa parte, representan as diferentes familias de pronomes. Se ben a presenza da familia de pronomes *I* é a maior de todas, non se pode comprobar que isto se deba ao carácter depresivo dos suxeitos, e non a unha característica propia da linguaxe inglesa, pois esta marca non é o suficientemente predominante nos resultados. Por tanto, non se puido demostrar nin tampouco descartar a hipótese de partida.

5.2.2 Termos dos *lexicons* dentro de *top n* termos dos vocabularios de depresión

Definidos os *lexicons* (ver 4.2), o obxectivo que se persegue con este experimento é ver cantos termos dos *lexicons* están presentes dentro dos diferentes *top n* termos dos vocabularios de depresión asociados aos RMs. A Táboa 5.2 detalla os resultados observados para o experimento. As súas filas correspóndense con todos os RMs computados (en función da suavización utilizada). A maiores, preséntase a categorización por grupos *A - B - C* definida no Punto 4.3.1. As columnas, pola súa parte, representan as ocorrencias dentro dos *top*

termos considerados para os vocabularios de depresión: o *top tamaño-do-lexicon* e o *top*. En vista dos resultados, o solapamento entre os vocabularios de depresión obtidos e os *lexicons* de referencia foi menos do agardado.

Táboa 5.1: Ocorrencias das familias de pronomes persoais dentro do *top* 20 dos vocabularios de depresión.

Vocabulario depresión		*I*	*You*	*He/She*	*We*	*They*
Dirichlet (0)	A	3	1	0	0	0
	B	3	1	0	0	0
	C					
Dirichlet (500)	A	3	2	0	0	0
	B	3	2	0	0	0
	C					
Dirichlet (1000)	A	3	2	0	0	0
	B	3	2	0	0	0
	C					
Dirichlet (1500)	A	3	2	0	0	0
	B	3	2	0	0	0
	C					
Dirichlet (2000)	A	3	2	0	0	0
	B	3	2	0	1	0
	C					
Dirichlet (2500)	A	3	2	0	0	0
	B	3	2	0	1	0
	C	3	2	1	1	0
Dirichlet (3000)	A	3	2	0	1	0
	B	3	2	1	1	0
	C	3	2	1	1	0
Dirichlet (3500)	A	3	2	0	1	0
	B	3	2	1	1	0
	C	3	2	1	1	0
Dirichlet (4000)	A	3	2	0	1	0
	B	3	2	1	1	0
	C	3	2	1	1	0
Dirichlet (4500)	A	3	2	0	1	0
	B	3	2	1	1	0
	C	3	2	1	1	0
Dirichlet (5000)	A	3	2	0	1	0
	B	3	2	1	1	0
	C	3	2	1	1	0
	A	3	1	0	0	0

. . . . . *(continúa na páxina seguinte)* . . . . . .

Táboa 5.1 – *(vén da páxina anterior)*

Vocabulario depresión		*I*	*You*	*He/She*	*We*	*They*
Jelinek-Mercer (0.0)	B	3	1	0	0	0
	C					
Jelinek-Mercer (0.2)	A	3	2	0	0	0
	B	3	1	0	0	0
	C					
Jelinek-Mercer (0.4)	A	3	2	0	0	0
	B	3	2	0	0	0
	C					
Jelinek-Mercer (0.6)	A	3	2	0	0	0
	B	3	2	0	0	0
	C					
Jelinek-Mercer (0.8)	A	3	2	0	0	0
	B	3	2	1	1	0
	C					
Jelinek-Mercer (1.0)	A	3	2	1	0	0
	B	3	2	1	1	0
	C					

Táboa 5.2: Termos dos *lexicons* en *top* termos dos vocabularios de depresión.

Vocabulario depresión		Pedesis		Choudhury	
		Top 636	*Top* 1000	*Top* 106	*Top* 1000
Dirichlet (0)	A	21	35	5	34
	B	21	35	5	34
	C				
Dirichlet (500)	A	18	37	5	36
	B	21	41	5	36
	C				
Dirichlet (1000)	A	18	39	5	36
	B	20	41	5	36
	C				
Dirichlet (1500)	A	18	39	5	36
	B	20	40	5	37
	C				
Dirichlet (2000)	A	18	38	5	37
	B	19	40	5	37
	C				
Dirichlet (2500)	A	18	37	5	37
	B	20	40	5	36

. . . . . *(continúa na páxina seguinte)* . . . . . .

Táboa 5.2 – *(vén da páxina anterior)*

Vocabulario depresión		Pedesis		Choudhury	
		Top 636	*Top* 1000	*Top* 106	*Top* 1000
	C	18	34	4	34
Dirichlet (3000)	A	19	38	5	37
	B	20	40	5	36
	C	18	34	4	34
Dirichlet (3500)	A	19	39	5	37
	B	20	40	6	36
	C	18	34	4	34
Dirichlet (4000)	A	19	40	5	37
	B	20	40	6	36
	C	18	35	4	34
Dirichlet (4500)	A	19	40	5	37
	B	21	40	6	36
	C	18	35	4	34
Dirichlet (5000)	A	20	40	5	37
	B	21	40	6	36
	C	18	35	4	34
Jelinek-Mercer (0.0)	A	21	35	5	34
	B	21	35	5	34
	C				
Jelinek-Mercer (0.2)	A	19	38	5	36
	B	21	41	5	36
	C				
Jelinek-Mercer (0.4)	A	18	37	5	36
	B	21	41	5	35
	C				
Jelinek-Mercer (0.6)	A	18	40	5	36
	B	20	41	5	36
	C				
Jelinek-Mercer (0.8)	A	18	39	5	37
	B	19	40	5	36
	C				
Jelinek-Mercer (1.0)	A	18	40	4	37
	B	21	38	5	38
	C				

5.2.3 *Ranking* de *baseline*

Cada unha das seguintes *queries* permiten obter candanseu *ranking*, de modo que sexa posible atopar un valor de referencia co cal comparar os resultados medidos para os *rankings* propios:

Q1: { *sad*, *lonely*, *hopeless*, *worthless* }[1]

Q2: { Todos os termos únicos (106) de Choudhury (ver A.2) }

Q3: { Todos os termos únicos (636) de Pedesis (ver A.1) }

Q4: { Adxectivos non ambiguos (7) de Choudhury (ver A.3) }

Q5: { Adxectivos non ambiguos (153) de Pedesis (ver A.4) }

Q6: { Adx. non amb. (13) de Choudhury expandido co método *Dist* [34] (ver A.5) }

Q7: { Adx. non amb. (16) de Choudhury expandido co método *WN* [34] (ver A.6) }

Q8: { Adx. non amb. (312) de Pedesis expandido co método *Dist* [34] (ver A.7) }

Q9: { Adx. non ambiguos (549) de Pedesis expandido co método *WN* [34] (ver A.8) }

Para cada unha das *queries*, o proceso debe ser o seguinte:

Fase de *training*

Usando a colección CLEF eRisk 2017, optimizar a configuración de suavización buscando maximizar a AP@100 (ver 5.1.2). Isto é, debe lanzarse a *query* usando todas as configuracións de *Smoothing* xa coñecidas (Dirichlet [0, 5000] e Jelinek-Mercer (0, 1] - 0.2) e ver cal de todas elas reporta mellor AP@100.

Fase de *test*

Usando a colección CLEF eRisk 2018, debe validarse a optimización da configuración de suavización feita na fase anterior. Esta validación consiste na repetición da *query* sobre a nova colección (disxunta), efectuando novamente as medicións para as métricas contempladas.

Poden observarse os resultados obtidos na Táboa 5.3. Cada fila da táboa correspóndese cun dos *baseline* establecidos. As columnas permiten consultar o número de termos de cada

[1] A *query* non se define *ad-hoc* para este proxecto, senón que se toma do traballo de Losada e Gamallo [34]. Son estas catro palabras as que os autores identificaron para o vector asociado á depresión.

unha, así como tamén a suavización optimizada en *training* e as medicións feitas en *test*. Pode verse que o mellor dos *rankings* é o obtido pola *Q2*.

Táboa 5.3: Resultados observados para as *queries* de *baseline*.

	nº Termos	*Smoothing* óptimo	Medicións en *Test*		
			AP@100	P@5	P@10
q1	4	Dirichlet (1000)	0.08511	0.6	0.4
q2	**106**	**Dirichlet (500)**	**0.21458**	**0.4**	**0.5**
q3	636	Dirichlet (0)	0.01082	0	0
q4	7	Dirichlet (500)	0.09998	0.2	0.3
q5	153	Dirichlet (0)	0.01086	0	0
q6	13	Jelinek-Mercer (0)	0.01935	0.2	0.1
q7	16	Jelinek-Mercer (0)	0.07924	0.2	0.3
q8	312	Dirichlet (0)	0.01082	0	0
q9	549	Dirichlet (0)	0.01082	0	0

5.2.4 *Ranking* con Pedesis ordenado cos vocabularios de depresión

No anterior experimento os termos das *queries*, é dicir, o seu contido, estaban fixados de inicio. Neste experimento, que busca obter o primeiro dos *rankings* propios, os termos das *queries* son seleccionados a partires da ordenación que os vocabularios de depresión fagan para o *lexicon* Pedesis (ver 4.2.1). É importante ter en conta que *ordenar* un *lexicon* cun vocabulario de depresión consiste en comparar ambos e reordenar o *lexicon* pola orde de aparición dos seus termos no vocabulario. Sabendo isto, as fases de *training* e *test* introducidas no Punto 5.2.3 modifícanse de xeito que se incorpora un novo parámetro de cara á optimización. Este parámetro será o número de termos seleccionados do *lexicon*. Así, o proceso completo sería o seguinte:

Fase de *training*

Usando a colección CLEF eRisk 2017, optimizar a configuración de suavización buscando maximizar a AP@100 (ver 5.1.2), xunto co *top* de termos do *lexicon* seleccionados para a *query*. Isto é, deben lanzarse sucesivas *queries* usando todas as configuracións de *Smoothing* xa coñecidas (Dirichlet [0, 5000] e Jelinek-Mercer (0, 1] - 0.2) en combinación cos diferentes *top* termos estudados. Os *top* termos estudados son os seguintes: 5, 10, 20, 50 e 100.

Fase de *test*

Usando a colección CLEF eRisk 2018, debe validarse a optimización feita na fase anterior. Esta validación consiste na repetición da *query* sobre a nova colección (disxunta) cos parámetros de suavización e *top* gañadores na fase de *training*, efectuando novamente as medicións

para as métricas contempladas.

Os resultados obtidos poden ser visualizados na Táboa 5.4, que amosa unha ordenación do *lexicon* por cada fila. Nas columnas represéntanse os parámetros optimizados en *training* e as medicións feitas en *test*. Pode observarse que o mellor desta serie de *rankings* foi o resultante froito da ordenación co vocabulario de depresión computado con Dirichlet(0).

Táboa 5.4: Resultados observados para as *queries* con Pedesis ordenado cos vocabularios de depresión.

	Parámetros optimizados		Medicións en *Test*		
RM ordenación	*Top*	*Smoothing*	AP@100	P@5	P@10
Dirichlet (0)	**10**	**Dirichlet (1000)**	**0.24783**	**0.6**	**0.5**
Dirichlet (500)	10	Dirichlet (3000)	0.21983	0.6	0.6
Dirichlet (1000)	10	Dirichlet (3000)	0.21388	0.6	0.5
Dirichlet (1500)	10	Dirichlet (3000)	0.21388	0.6	0.5
Dirichlet (2000)	10	Dirichlet (3000)	0.21388	0.6	0.5
Dirichlet (2500)	10	Dirichlet (3000)	0.21388	0.6	0.5
Dirichlet (3000)	10	Dirichlet (2000)	0.22183	0.8	0.8
Dirichlet (3500)	10	Dirichlet (2000)	0.22183	0.8	0.8
Dirichlet (4000)	10	Dirichlet (1000)	0.21777	0.6	0.8
Dirichlet (4500)	5	Dirichlet (1000)	0.22218	0.4	0.4
Dirichlet (5000)	5	Dirichlet (1000)	0.22218	0.4	0.4
Jelinek-Mercer (0.0)	10	Dirichlet (1000)	0.24783	0.6	0.5
Jelinek-Mercer (0.2)	10	Dirichlet (1000)	0.24783	0.6	0.5
Jelinek-Mercer (0.4)	10	Dirichlet (3000)	0.21983	0.6	0.6
Jelinek-Mercer (0.6)	10	Dirichlet (3000)	0.21388	0.6	0.5
Jelinek-Mercer (0.8)	10	Dirichlet (2000)	0.22183	0.8	0.8
Jelinek-Mercer (1.0)	5	Dirichlet (1000)	0.11305	0.4	0.4

5.2.5 *Ranking* con Choudhury ordenado cos vocabularios de depresión

Este experimento realízase baixo as mesmas premisas e condicións que o visto no Punto 5.2.4, sendo a única diferenza que se usa o *lexicon* Choudhury (ver 4.2.2) onde antes se empregaba Pedesis. Os resultados obtidos están recollidos na Táboa 5.5, que amosa unha ordenación do *lexicon* por cada fila. Nas columnas represéntanse os parámetros optimizados en *training* e as medicións feitas en *test*. O mellor *ranking* para esta serie foi o obtido a partires da ordenación co vocabulario de depresión computado con Dirichlet(500).

Táboa 5.5: Resultados observados para as *queries* con Choudhury ordenado cos vocabularios de depresión.

	Parámetros optimizados		Medicións en *Test*		
RM ordenación	*Top*	*Smoothing*	AP@100	P@5	P@10
Dirichlet (0)	10	Dirichlet (1500)	0.23770	0.8	0.5
Dirichlet (500)	**20**	**Dirichlet (2000)**	**0.27216**	**0.6**	**0.7**
Dirichlet (1000)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (1500)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (2000)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (2500)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (3000)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (3500)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (4000)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (4500)	20	Dirichlet (2000)	0.24666	0.6	0.6
Dirichlet (5000)	20	Dirichlet (2000)	0.24666	0.6	0.6
Jelinek-Mercer (0.0)	10	Dirichlet (1500)	0.23770	0.8	0.5
Jelinek-Mercer (0.2)	20	Dirichlet (2000)	0.27216	0.6	0.7
Jelinek-Mercer (0.4)	20	Dirichlet (2000)	0.27216	0.6	0.7
Jelinek-Mercer (0.6)	20	Dirichlet (2000)	0.24666	0.6	0.6
Jelinek-Mercer (0.8)	20	Dirichlet (2000)	0.24666	0.6	0.6
Jelinek-Mercer (1.0)	10	Dirichlet (1500)	0.22670	0.2	0.5

5.2.6 *Ranking* cos vocabularios de depresión

Este experimento, ao igual que o experimento descrito no Punto 5.2.5, realízase baixo as mesmas condicións que o disposto no Punto 5.2.4 cunha única variación. Na fase de *training*, os *top* termos selecciónanse directamente dos vocabularios de depresión, e non de ningún *lexicon* ordenado. A Táboa 5.6 ilustra os resultados obtidos, amosando un vocabulario de depresión por cada fila. Nas columnas represéntanse os parámetros optimizados en *training* e as medicións feitas en *test*. O mellor *ranking* para esta serie, e tamén para o global da tarefa, obtívose co vocabulario de depresión obtido con Dirichlet (4500).

Táboa 5.6: Resultados observados para as *queries* cos vocabularios de depresión.

	Parámetros optimizados		Medicións en *Test*		
Vocabulario depresión	*Top*	*Smoothing*	AP@100	P@5	P@10
Dirichlet (0)	100	Dirichlet (500)	0.28952	0.8	0.6
Dirichlet (500)	100	Dirichlet (500)	0.30463	0.8	0.7
Dirichlet (1000)	100	Dirichlet (500)	0.34010	0.8	0.8
Dirichlet (1500)	100	Dirichlet (500)	0.34013	0.8	0.8
Dirichlet (2000)	100	Dirichlet (500)	0.31568	0.6	0.7
Dirichlet (2500)	50	Dirichlet (500)	0.31771	0.4	0.6
Dirichlet (3000)	50	Dirichlet (500)	0.32038	0.4	0.6
Dirichlet (3500)	50	Dirichlet (500)	0.30176	0.6	0.5
Dirichlet (4000)	50	Dirichlet (500)	0.32714	0.8	0.8
Dirichlet (4500)	**50**	**Dirichlet (500)**	**0.35163**	**0.8**	**0.7**
Dirichlet (5000)	50	Dirichlet (500)	0.33041	0.8	0.6
Jelinek-Mercer (0.0)	100	Dirichlet (500)	0.28952	0.8	0.6
Jelinek-Mercer (0.2)	100	Dirichlet (500)	0.32354	0.8	0.8
Jelinek-Mercer (0.4)	100	Dirichlet (500)	0.31301	0.8	0.7
Jelinek-Mercer (0.6)	100	Dirichlet (500)	0.34010	0.8	0.8
Jelinek-Mercer (0.8)	50	Dirichlet (500)	0.31171	0.4	0.6
Jelinek-Mercer (1.0)	50	Dirichlet (500)	0.28634	0.6	0.6

5.2.7 *Ranking* aleatorio

Este experimento debe verse como un complemento ao *ranking* de *baseline*, pois tamén pretende establecer unha referencia coa que comparar os *rankings* propios. Realizando un *ranking* aleatorio, o que se busca é determinar que resultados se obteñen simplemente debidos ao azar. En concreto, selecciónanse 100 persoas aleatorias da colección CLEF eRisk 2018 e efectúanse as medicións habituais para o *ranking*. Este proceso repítese dez veces, promediando os resultados dispoñibles na Táboa 5.7.

	AP@100	P@5	P@10	Depress@100
Media	0.1177	0.12	0.06	9.3
Mediana	0.1013	0	0	8
Máximo	0.1899	0.4	0.2	15
Mínimo	0.0633	0	0	5

Táboa 5.7: Resultados observados para un conxunto de *rankings* aleatorios.

5.2.8 Resultados destacados *ranking*

Este Punto pretende ser un breve resumo dos mellores e máis destacables resultados obtidos nos experimentos de *ranking*. Deste modo, as filas da Táboa 5.8 recollen os *rankings*

plantexados: *baseline*, Pedesis e Choudhury ordenados cos vocabularios de depresión e os propios vocabularios. As columnas amosan o número de termos da mellor *query* e o *smoothing* optimizado, así como tamén as medicións de avaliación feitas en *test*. Pode apreciarse, e é algo a destacar, que o mellor *ranking* se obtivo utilizando directamente os vocabularios de depresión.

Táboa 5.8: Mellores resultados de *ranking* e *baseline* de referencia.

			Medicións en *Test*		
	nº Termos	*Smoothing*	AP@100	P@5	P@10
Baseline	106	Dirichlet (500)	0.21458	0.4	0.5
Pedesis	10	Dirichlet (1000)	0.24783	0.6	0.5
Choudhury	20	Dirichlet (2000)	0.27216	0.6	0.7
Vocabularios	**50**	**Dirichlet (500)**	**0.35163**	**0.8**	**0.7**

Complementariamente, amósanse a continuación os termos das *queries* que reportaron os mellores *rankings*:

Termos para mellor *ranking* con Pedesis ordenado cos vocabularios de depresión:

```
feel, hurt, small, afraid, sad, deep, difficult, weight, numb, night
```

Termos para mellor *ranking* con Choudhury ordenado cos vocabularios de depresión:

```
like, girl, want, love, help, friend, date, weight, life, talk,
social, discuss, game, doctor, relationship, care, answer, adhd,
home, man
```

Termos para mellor *ranking* cos vocabularios de depresión:

```
i (0.03859225937360453), you (0.025079550443825892),
my (0.010815754652712214), have (0.009696152294924475),
so (0.009594713490141455), like (0.009002567370524617),
me (0.007544006131354665), just (0.007287057604773934),
your (0.007277128677351451), do (0.006354292453386365),
what (0.005921096070870912), get (0.005915968794477641),
can (0.005472110028430423), about (0.005391191361434873),
on (0.005272547418267872), would (0.004948309279394993),
think (0.004832577520143655), all (0.004826185199113026),
know (0.004742989976273174), we (0.004632796666553628),
peopl (0.004565053905214582), becaus (0.004198344716704538),
he (0.004196533808056153), time (0.004190579707536709),
http (0.004144049718138366), more (0.00406343178631741),
when (0.003974071632080296), go (0.003964243693105009),
out (0.003854747673850159), other (0.003808897036572563),
thing (0.0037042871552916126), feel (0.003695232471791368),
us (0.0036618415731294453), thank (0.003639540352110229),
```

```
make (0.0036039566214138494), from (0.00358339055980381),'
dont (0.003546734232373987), some (0.003523746154913984),'
im (0.0034862191743254747), realli (0.0033825100962233835),
how (0.0032166502526287787), want (0.0031950469386712913),
much (0.003175959878622647), too (0.003159537720945163),
up (0.00310697844599845), her (0.0030981291101763943),
them (0.002957650724629138), she (0.002912884170403016),
veri (0.002865516721577166), work (0.002828530395883663)
```

Capítulo 6

Metodoloxía e xestión do proxecto

Este capítulo analiza aspectos relacionados coa metodoloxía seleccionada para a xestión do proxecto. En concreto, utilízase *Scrum* cunha serie de adaptacións debido tanto ao feito de que o tipo de proxecto non é de enxeñaría clásica como a certas limitacións no equipo de desenvolvemento ao tratarse dun traballo de fin de grao. Preséntanse cuestións como a propia metodoloxía, as estimacións de custes e duración, a planificación e a xestión de riscos.

6.1 Metodoloxía

A metodoloxía de desenvolvemento resulta de gran importancia, xa que supón un marco de traballo que guía todo o proceso do proxecto. É por isto que resulta importante realizar unha escolla fundamentada e adecuada. Ao longo do tempo xurdiron diferentes metodoloxías para a xestión de proxectos, sendo o estudo e desenvolvemento das mesmas un campo dentro da disciplina da enxeñaría do software.

6.1.1 Escolla da metodoloxía

Nun proxecto de desenvolvemento en investigación como é este, a escolla da metodoloxía resulta de igual importancia. Habitualmente, no referido á metodoloxía de investigación, as eleccións adoitan estar baseadas no método científico. Por outra banda, para a xestión do proxecto científico, é preciso escoller unha metodoloxía. Dada a natureza do proxecto, esta escolla non resulta tan clara coma nun proxecto de enxeñaría clásico. A continuación preséntanse unha serie de características desexables que se deben ter en conta á hora de seleccionar a metodoloxía.

Ciclos de vida curtos
Búscase iterar de xeito constante, para adaptarse aos resultados que se van obtendo nos sucesivos ciclos e definir novos experimentos a partir destes.

Adaptación aos cambios
Se en calquera proxecto é habitual que aparezan cambios, nun proxecto deste tipo máis aínda. As necesidades varían en función dos resultados obtidos.

Xestion de riscos
É necesario ser quen de detectar o antes posible as ameazas que poidan xurdir durante a investigación para poder corrixir o curso do desenvolvemento, de tal modo que se acaden os obxectivos establecidos.

Visión do avance da investigación
Resulta desexable poder estudar os avances conseguidos durante a investigación. Isto faise coa intención de incorporar a información recadada e obter conclusións intermedias que permitan dirixir obxectivos concretos.

Todos estes requisitos apuntan á necesidade dunha metodoloxía áxil[1], amplamente estendidas a día de hoxe na xestión de proxectos. Están baseadas no uso de ciclos de vida iterativos, enfatizando na idea de obter un produto intermedio á finalización de cada un destes ciclos. Isto permite adaptarse mellor aos cambios nos requisitos e detectar o antes posibles os riscos. A maiores, facilita o estudo do avance do proxecto e supón unha motivación para o equipo. En concreto, escolleuse *Scrum*. As súas particularidades explícanse a continuación.

6.1.2 Scrum

Scrum[39] é unha das denominadas metodoloxías áxiles. Preséntase coma un marco deliberadamente incompleto, definindo unicamente as partes necesarias para implementar a súa teoría. Deste xeito, en lugar de proporcionar instrucións detalladas, serán as regras de Scrum as que guíen as relacións e interaccións.

Scrum (ver figura 6.1) utiliza un enfoque iterativo e incremental. Isto permite, por unha parte, mellorar a comunicación co cliente, xa que ao remate de cada incremento debería ser posible ter un entregable software. Ademais, tamén favorece a inspección continua, o que axuda a mellorar o control de riscos.

Para comprender mellor Scrum, cómpre definir o seu equipo (ver 6.1.2), os seus eventos (ver 6.1.2) e os artefactos dos que se dispón (ver 6.1.2). Todo isto fundaméntase nos piares empíricos da propia metodoloxía:

Transparencia
Favorece a inspección, e unha inspección sen transparencia pode levar a engano.

[1] https://agilemanifesto.org/iso/gl/manifesto.html

Inspección

Permite a adaptación, partindo de que a inspección sen adaptación resulta inútil. Debe ser continua e frecuente, para detectar desviacións e problemas co maior tempo posible. A inspección está, por tanto, moi relacionada cos eventos Scrum (6.1.2), deseñados para provocar cambios.

Adaptación

Indispensable cando se detecta unha desviación. No menor tempo posible, debe realizarse un axuste que permita minimizar a desviación adicional.

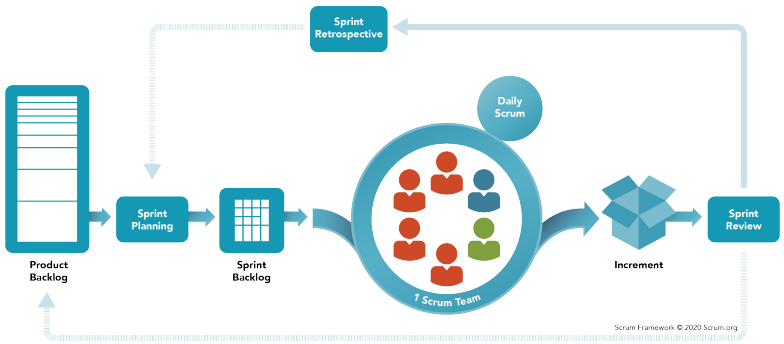

Figura 6.1: Detalle do proceso de *Scrum*.

Equipo

Un equipo de Scrum debe ser multifuncional, de modo que os seus membros teñan as habilidades necesarias para crear valor en cada *Sprint* (6.1.2). É, por tanto, unha unidade cohesionada que elimina os sub-equipos e as xerarquías. Os equipos Scrum, idealmente, deben manterse entorno ás 10 persoas para buscar o equilibrio entre ser áxil e ter a capacidade de completar un traballo significativo dentro dun *Sprint*. Ademais, estarán presentes os seguintes tres roles, coas súas responsabilidades específicas.

Desenvolvedores

Deben estar comprometidos a crear calquera aspecto dun incremento funcional dentro do *Sprint*. Serán sempre responsables de crear un plan para o *Sprint*, inculcar calidade e adaptar o seu traballo diario de cara á consecución do obxectivo do *Sprint*.

Product Owner

Actúa como persoa propietaria do produto, debendo maximizar o seu valor. Tamén é

responsable da xestión do *Product Backlog* polo que, entre outras tarefas, deberá desenvolver e comunicar explicitamente o *Product Goal*, crear e comunicar elementos de traballo pendentes do produto, etc.

Scrum Master
É o responsable de establecer Scrum tal e como se define na súa guía[39]. Buscará, por tanto, axudar a todas as persoas que dalgún modo participan no proxecto a comprender a teoría e a práctica da metodoloxía.

Eventos

O *Sprint* actúa como evento contedor de todos os demais eventos, que deberían ser vistos como oportunidades para a inspección e adaptación dos artefactos (ver 6.1.2). Os *Sprints*, idealmente, deben durar un mes ou menos e comezar un a continuación do anterior. Deste xeito, poderase garantir a inspección e a adaptación cunha periodicidade suficiente para que o *Sprint Goal* non se volva obsoleto, controlando tamén o risco e a complexidade. Todos os eventos que contén o *Sprint* son os seguintes:

Sprint Planning
Reunión que dá comezo ao *Sprint*, establecendo o traballo que se realizará dentro do mesmo e os obxectivos que se pretenden acadar. Debe responder certas cuestións, como son identificar porque o *Sprint* é valioso, que se pode facer nel e como se realizará o traballo escollido. Como máximo, este evento debería durar oito horas para *Sprints* dun mes.

Daily Scrum
Reunión diaria de quince minutos de duración para as persoas desenvolvedoras do equipo *Scrum* (ver 6.1.2). O seu obxectivo debe ser revisar o progreso de cara á consecución do *Sprint Goal*, inspeccionando o traballo realizado. Para isto, resulta útil poñer en común o traballo realizado no día anterior, definir o traballo que se realizará no propio día e comentar posibles problemas ou obstáculos que apareceran no proceso.

Sprint Review
Reunión coas partes interesadas ao remate do *Sprint*, de modo que sexa posible revisar e inspeccionar o traballo realizado durante o incremento e determinar futuras adaptacións. Permite ver que se logrou durante o *Sprint* e decidir que facer a continuación. Como máximo, este evento debería durar catro horas para *Sprints* dun mes.

Sprint Retrospective
Reunión posterior e complementaria á *Sprint Review* co propósito de planificar formas

de aumentar a calidade e a eficacia. O equipo Scrum debe analizar que foi ben durante o *Sprint*, que problemas se atoparon e como estes foron ou non foron resoltos. Como máximo, este evento debería durar tres horas para *Sprints* dun mes.

Artefactos

Ao longo do desenvolvemento do produto elabóranse unha serie de elementos denominados *artefactos*. Os artefactos de *Scrum* están deseñados para maximizar a transparencia da información clave. Son os seguintes:

Product Backlog
Traballo pendente do produto. Organízase como unha lista emerxente e ordenada do necesario para mellorar o produto, sendo os seus elementos implementados durante os *Sprints*. Exprésase en forma de Historias de Persoa Usuaria e é a única fonte de requisitos do produto. Dado que é dinámica e evoluciona conforme o fai o produto, nunca está completa.

En calquera traballo de fin de grao, e en especial se se trata dun proxecto de desenvolvemento en investigación, o *Product Backlog* debe estar practicamente completo ao comezo. Isto é debido a que se deben ter claros os obxectvos da investigación, as ferramentas das que se dispón e o plan de traballo, que foi especificado no anteproxecto. Non obstante, os requisitos serán refinados en función dos resultados que se vaian obtendo, quedando este aspecto cuberto polas facilidades para a adaptación que aporta *Scrum*.

Sprint Backlog
É o subconxunto de tarefas do *Product Backlog* que o equipo selecciona e incorpora ao *Sprint*. Elabórase durante o *Sprint Planning*. É posible engadir novos elementos durante a realización do *Sprint* en caso de consideralo necesario para cumplir co obxectivo establecido.

Incremento
Correspóndese co resultado do *Sprint*, sendo a suma de todas as tarefas, casos de uso, historias e elementos completados durante o *Sprint*. É posto en disposición da persoa usuaria en forma de software, e resulta un aspecto clave para o carácter iterativo e incremental de *Scrum*.

6.1.3 Adaptación de Scrum a este traballo

Como se comentou con anterioridade, existen unha serie de aspectos que imposibilitan unha fiel adopción da metodoloxía. Por exemplo, o equipo de desenvolvemento está formado

por unha única persoa, a dedicación non se mantén constante, etc. A continuación preséntanse as adaptacións realizadas sobre *Scrum* para poder utilizar esta metodoloxía no proxecto:

- O rol de *Product Owner* está desempeñado polos directores do proxecto Álvaro Barreiro García e Miguel Ánxo Pérez Vila.
- O rol de *Scrum Master* está desempeñado polo alumno Eliseo Bao Souto.
- O *Equipo de Desenvolvemento* está composto por unicamente unha persoa, o alumno Eliseo Bao Souto.
- Dado que o *Equipo de Desenvolvemento* é unipersoal, prescíndese das reunións diarias. Non obstante, o alumno debe avaliar as cuestións que se trataría nestas reunións.
- Dadas as obrigas académicas e formativas do alumno, o esforzo realizado nos *Sprints* non se puido manter uniforme. Complementariamente a isto, si que se intentou respetar a duración dos *Sprints*.
- A duración dos *Sprints* é de 4 semanas, cun esforzo de 60 puntos de historia.

6.2 Xestión do proxecto

A xestión do proxecto busca satisfacer unha serie de obexctivos que permitan maximizar o éxito do proxecto. Entre eles, destacan:

- Cumplir a planificación e os custes estimados.
- Acadar os termos de calidade desexados.
- Optimizar o uso de recursos.
- Ter constancia en calquera momento do proxecto do estado do mesmo.

6.2.1 Estimación

Para realizar unha estimación das historias de usuario entendible e non demasiado complexa, establécese que 1 hora de traballo equivale a 1 punto de historia. Isto permite estimar o custe de realizar unha historia de usuario en horas por persoa. A maiores, fíxanse 60 puntos de historia por *Sprint*, sendo a duración destes 4 semanas. Por tanto, en condicións ideais, o tempo de traballo é de 3 horas ao día. Como xa se comentou na sección 6.1.3, existen unha serie de condicionantes que fan que algúns *Sprints* se desvíen do ideal en canto ao esforzo. Non obstante, a duración dos mesmos si que se mantivo constante.

6.2.2 Recursos

Os recursos empregados para un proxecto destas características poden dividirse nas seguintes tres categorías.

Recursos humanos

Tal e como se introduce na sección 6.1.3, un total de tres persoas participan neste proxecto. Os directores Álvaro Barreiro García e Miguel Anxo Pérez Vila exercen a función de *Product Owners*, tomando parte activa na planificación e definición de requisitos. O resto de roles, tanto o de *Scrum Master* coma o de *Equipo de Desenvolvemento*, están representados polo alumno Eliseo Bao Souto. Isto implica que o alumno terá que, por un lado, asegurar o cumprimento do marco metodolóxico, así como tamén exercer de analista, deseñador e desenvolvedor.

Recursos materiais

Unicamente se fai uso dun recurso material, o ordenador persoal propiedade do alumno Eliseo Bao Souto.

Recursos software

Son todos os mencionados no capítulo 3. Todos eles son de uso gratuito, contan con versión de estudante ou foron proporcionados polos directores.

6.2.3 Custes

Para definir os custes dos recursos humanos faise uso do informe *Estudo salarial - Sector TIC Galiza 2015-2016* [40]. Na táboa 6.1 recóllese a estimación salarial para os diferentes perfís identificados. A maiores, deben terse en conta os seguintes aspectos:

- En total, realízanse 6 *Sprints*.
- Tal e como se define na subsección 6.1.3, os *Sprints* teñen unha duración de 4 semanas.
- Os *Product Owner* (directores) adican cada un 5 horas de traballo por *Sprint*.
- O equipo de desenvolvemento (alumno) adica 15 horas semanais de traballo: 1 como analista e 14 como desenvolvedor.

Deste xeito, o detalle completo dos custes dos recursos humanos queda especificado na táboa 6.2.

Respecto aos recursos materiais e de software, os custes completos son os recollidos na táboa 6.3.

Perfil	Custe
Director	45€/hora
Analista	25€/hora
Desenvolvedor	20€/hora

Táboa 6.1: Estimación salarial para os recursos humanos do proxecto.

Perfil	Tempo (h/*Sprint*)	Dedicación (nº *Sprints*)	Custe (€/h)	Total (€)
Director	10 (ambos)	6	45	2 700
Analista	4	6	25	600
Desenvolvedor	56	6	20	6 200
			Total	**9 500**

Táboa 6.2: Detalle dos custes totais dos recursos humanos.

Recurso	Unidades	Custe (€/ud)	Vida (meses)	Uso (meses)	Total (€)
Licenzas					0
Ordenador persoal	1	800	48	6	100
				Total	**100**

Táboa 6.3: Detalle dos custes totais dos recursos materiais e software

Sumando os custes dos recursos humanos, materiais e software, o total do proxecto ascende a 9 600€.

6.2.4 Xestión de riscos

En todo proxecto, de cara a evitar as ameazas e mitigar as súas consecuencias, resulta necesario realizar unha adecuada identificación e xestión dos riscos aos que se está exposto. Por este motivo, en primeiro lugar, debe levarse a cabo unha fase de identificación e clasificación dos mesmos, obtendo a exposición do proxecto a cada risco en función da probabilidade de aparición e o impacto en caso de ocorrer o risco. Na táboa 6.4 pode verse o resultado deste proceso de identificación e clasificación.

Código	Descrición	Probabilidade	Impacto	Exposición
R1	Planificación inadecuada	Alta	Alto	Alta
R2	Mal deseño	Media	Medio	Media
R3	Mala implementación	Baixa	Moi alto	Alta
R4	Baixa eficiencia	Media	Baixo	Baixa
R5	Falta de recursos	Baixa	Alto	Media
R6	Perda de información	Baixa	Moi alto	Alta

Táboa 6.4: Detalle da identificación e clasificación dos riscos presentes no proxecto.

Prevención

Aqueles riscos aos que se considera que se está altamente exposto deben ser tratados dun xeito especial, xa que de ocorrer poden supoñer grandes cambios na raíz do proxecto ou incluso o seu fracaso. Por iso, son analizados nunha fase temperá do proxecto, definindo plans de continxencia. A continuación detállanse os plans de continxencia para os riscos de alta exposición:

- **R1 - Planificación inadecuada.** Dada a escasa experiencia do alumno á hora de realizar planificación de proxectos, é probable que existan desviacións. En parte, este risco queda minimizado de xeito natural pola propia metodoloxía pero, aínda así, faise un exhaustivo control para evitar retrasos.
- **R3 - Mala implementación.** Unha mala implementación daría lugar á obtención de resultados erróneos na experimentación. Aínda que isto é pouco probable que ocorra, xa que a complexidade do software a implementar non é elevada, este suposto sería

catastrófico. Para evitalo, establécense controis rutinarios con perfís expertos, encarnados polos directores do proxecto, que se encargarán de validar a implementación. Tamén nesta liña, realízase un proceso de validacion dos resultados con respecto ás *queries* usadas como *baseline*.

- **R6 - Perda de información.** De novo, trátase dun risco pouco probable pero de gran impacto en caso de suceder. A perda de información pode darse por mútliples causas: erro humano, fallo hardware, roubo... En calquera caso, perder o propio código fonte, os resultados dos experimentos ou a documentación suporía un grave retraso no proxecto. Por iso, establécese en todo momento unha política de copias de seguridade e control de versións na nube.

Seguemento

Para os riscos de exposición media realízase un seguemento constante durante o proxecto para controlar que non se convirtan en riscos de alta exposición. En concreto:

- **R2 - Mal deseño.** Os ciclos de vida curtos e o carácter iterativo da metodoloxía permiten detectar erros no deseño e corrixilos antes de que se propaguen.
- **R5 - Falta de recursos.** Os recursos son xa coñecidos de antemán, polo que a planificación se fai acorde a eles e sempre con certas folguras para paliar posibles situacións de baixa dispoñibilidade dos recursos.

Asimilación

Os riscos con pouca exposición son simplemente aceptados, xa que é pouco probable que ocorran e ademais o seu impacto en caso de que se dea este suposto é pouco elevado. Así, realizar plans de continxencia ou manter un seguemento exhaustivo sería máis custoso que realizar a propia xestión de problemas en caso de que se chegaran a dar. Son riscos de baixa exposición os seguintes:

- **R4 - Baixa eficiencia.** Dada a escasa experiencia do alumno implementando solucións deste tipo, é posible que a proposta inicial non sexa tan eficiente como podería chegar a ser. Non obstante, hai que incidir en que o peso computacional sopórtao o cálculo dos RMs. Unha vez computados estes, realizar a fase de *ranking* ten pouco custe computacional e realízase con suficiente velocidade.

Capítulo 7

Desenvolvemento

O capítulo que a continuación se presenta recolle, primeiramente, a fase de análise do proxecto. Esta sección discute os actores e requisitos, así como tamén recolle todas as historias de usuario identificadas. A continuación, defínese a arquitectura a utilizar. Xa para finalizar, detállase o traballo realizado en cada un dos *Sprints*. Esta fase resulta de vital importancia para o traballo, debendo ser realizada nas etapas máis temperás do proxecto.

7.1 Análise de requisitos

A tarefa de análise de requisitos lévase a cabo, tipicamente, mediante reunións coas persoas cliente, que deberán especificar todas as súas necesidades. Destas necesidades, aquelas que sexan realizables (no sentido de ser implementables e satisfacibles) conformarán os requisitos funcionais. No caso deste proxecto, este proceso tivo lugar nas reunións iniciais entre o alumno e os directores.

7.1.1 Requisitos funcionais

En consonancia co explicado anteriormente, os requisitos funcionais poden definirse tamén como as características e/ou funcionalidades que debe ter o sistema. Nas reunións iniciais mantidas entre o alumno e os directores, os requisitos foron identificados a alto nivel. Entre eles:

Indexar coleccións de texto
É o primeiro paso a dar para poder levar a cabo a experimentación. O sistema debe permitir a indexación de coleccións de documentos, xa que os índices son a base sobre a que se computan os RM.

Obter pesado de termos dun RM para un determinado índice
O sistema debe ser quen de computar o RM e obter o seu vocabulario de depresión

asociado, que non é máis que realizar o pesado de termos dunha colección previamente indexada. É tamén requisito permitir esta operación tomando un amplo abano de métodos de suavización.

Creación dunha colección propia
De cara á experimentación, é necesario construir unha colección propia para o mesmo contexto que as detalladas na sección 4.1. O sistema, por tanto, debe ofrecer esta característica.

Analizar os vocabularios de depresión
O sistema debe permitir analizar individualmente os vocabularios de depresión(termos dos RM), así como tamén comparalos cos lexicons que se toman como referencia.

Optimizar *queries* de *baseline*
Para a tarefa de *ranking*, o sistema ten que permitir a optimización de parámetros sobre unha colección para as *queries* denominadas de *baseline* ou referencia. Así mesmo, o sistema ten que permitir validar os valores optimizados sobre unha colección que debe ser disxunta á utilizada en primeira instancia.

Optimizar *queries* propias
Tamén para a tarefa de *ranking*, o sistema debe permitir a optimización de paramétros sobre unha colección para unha serie de *queries* manuais que se constrúe coa información dos vocabularios de depresión. Tamén é necesario validar os valores optimizados sobre outra colección con intersección nula coa colección utilizada para adestrar.

7.1.2 Requisitos non funcionais

Os requisitos non funcionais especifican criterios que o sistema debe cumplir, incidindo isto de maneira directa nas decisións de deseño que se tomen. En concreto, para este proxecto identifícanse os seguintes requisitos non funcionais:

Eficacia
É necesario ter a certeza de que os resultados obtidos son correctos e que non inducen a erro.

Escalabilidade
O sistema debe ser quen de manexar coleccións de gran volume.

Extensibilidade e manteñibilidade
Resulta desexable poder engadir novas funcionalidades e modificacións ao longo da vida do produto.

Tempo de resposta asumible ante la elaboración do *ranking*
Os sistemas de detección de depresión non só deben ser fiables en canto á precisión das súas predicións, senón que tamén teñen que ser o suficientemente rápidos no tempo para que a detección se poida realizar antes de que a condición das persoas que padecen a enfermidade se vexa deteriorada.

7.1.3 Historias de Persoa Usuaria

Tomando como punto de partida os requisitos funcionais descritos a alto nivel (ver sección 7.1.1), é posible especificar a totalidade dos requisitos funcionais mediante a definición das Historias de Persoa Usuaria. Todas as Historias definidas para este proxecto atópanse recollidas na Táboa 7.1, cada unha delas coa súa estimación de Puntos de Historia. É importante saber que esta lista de Historias conforma o denomidado *Product Backlog*, desde onde se seleccionan as Historias para incorporar aos *Sprints*.

Táboa 7.1: Historias de Persoa Usuaria e puntos estimados.

ID	Historia de Persoa Usuaria	Puntos
1	Como *persoa autora* necesito coñecer en detalle as particularidades da construción de índices con Apache Lucene (ver 3.2.3).	20
2	Como *persoa autora* quero poder indexar coleccións CLEF eRisk.	40
3	Como *persoa autora* preciso coñecer en detalle os aspectos teóricos dos RMs.	20
4	Como *persoa autora* quero poder computar un RM sobre unha determinada colección, usando para tal fin os índices previamente construídos.	30
5	Como *persoa autora* quero poder obter os vocabularios de depresión dos RM obtidos anteriormente.	10
6	Como *persoa autora* necesito coñecer o estado da arte en canto ao tipo dos conxuntos de datos empregados para o proxecto.	10
7	Como *persoa autora* quero conseguir un novo conxunto de datos para a realización de experimentos complementarios.	50

. . . . . *(continúa na páxina seguinte)* . . . . . .

Táboa 7.1 – *(vén da páxina anterior)*

ID	Historia de Persoa Usuaria	Puntos
8	Como *persoa autora* quero poder emitir xuízos de valor sobre os vocabularios de depresión dos RMs, efectuando medicións nos mesmos.	20
9	Como *persoa autora* preciso documentarme sobre as características de *lexicons* profesionais.	10
10	Como *persoa autora* quero utilizar os *lexicons* profesionais para realizar comparativas entre eles e os vocabularios de depresión dos RMs.	20
11	Como *persoa autora* quero utilizar a información dos vocabularios de depresión dos RMs para operar sobre os *lexicons* profesionais.	10
12	Como *persoa autora* quero obter uns *rankings* de *baseline* por medio de *queries* predefinidas.	30
13	Como *persoa autora* quero poder optimizar, validar e avaliar os *rankings* de *baseline*.	30
14	Como *persoa autora* quero utilizar a información dos vocabularios de depresión dos RM e os *lexicons* profesionais para obter *rankings*.	30
15	Como *persoa autora* quero poder optimizar, validar e avaliar os *rankings* manuais.	30

7.2 Desenvolvemento

Esta sección describe o proceso de desenvolvemento levado a cabo durante o proxecto, detallando as Historias de Persoa Usuaria realizadas en cada *Sprint* e facendo un pequeno balance ao final de cada un deles. Logo da planificación inicial do proxecto, a suma total do *Product Backlog* ascende a 360 Puntos de Historia. Dado que a duración do proxecto será de 6 *Sprints*, para cada un deles deberanse completar 60 Puntos de Historia no caso ideal.

A Figura 7.1 amosa a estimación inicial do proxecto, e nela poden apreciarse unha serie de detalles:

Project points
Representan a totalidade dos puntos asignados para o proxecto.

Defined points
Representan a cantidade de puntos asignados ás Historias de Persoa Usuaria.

Closed points
Representa a suma dos Puntos de Historia xa completados.

Points/Sprint
Representa a media de Puntos completados en cada *Sprint*. Na terminoloxía de Scrum 6.1 coñécese como *velocity* e, de acordo co explicado previamente, debe manterse sempre arredor dos 60 Puntos de Historia por *Sprint*.

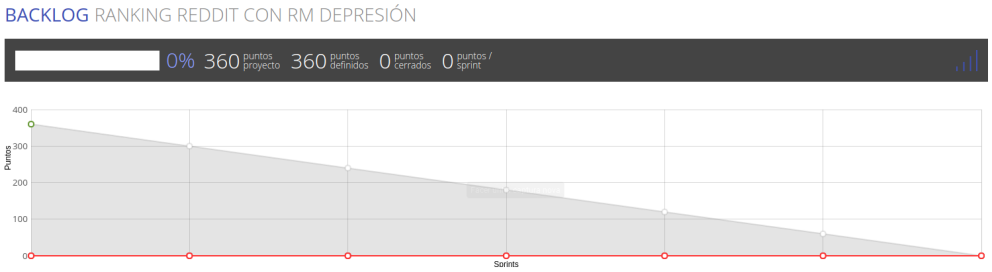

Figura 7.1: Planificación inicial do proxecto.

7.2.1 *Sprint* 1: Indexación de coleccións CLEF eRisk

O primeiro *Sprint* do proxecto supón a base sobre a que se levantará todo o desenvolvemento posterior. Inicialmente, para o *Sprint* 1 seleccionáronse as Historias 1 e 2, tal e como ilustra a Figura 7.2. Como parte do proceso, realizouse unha división das Historias en Tarefas, identificando as seguintes:

T1 - Estudo e preparación da indexación.

T1.1 - Estudo e asimilación de conceptos IR.

T1.2 - Estudo e documentación sobre Apache Lucene (3.2.3).

T1.3 - Instalación e configuración do entorno.

T2 - Preparación dos documentos (*parsing*).

T2.1 - Estudo do formato das coleccións.

T2.2 - Estudo da ferramenta lxml (3.2.1).

T2.3 - Implementación e uso do *parser*.

T3 - Indexación das coleccións CLEF eRisk (4.1).

T3.1 - Implementación indexador con Apache Lucene (3.2.3).

T3.2 - Construción dos índices.

T3.2.1 - 2019, 2020 e 2021.

T3.2.2 - 2017, 2018, 2019, 2020 e 2021.

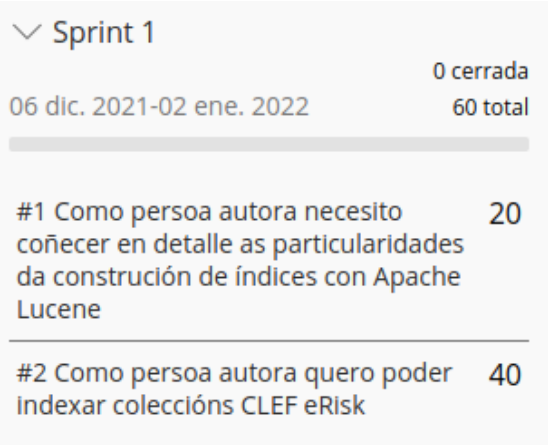

Figura 7.2: Planificación inicial do *Sprint* 1.

As Historias seleccionadas inicialmente completáronse antes da finalización do *Sprint*. Isto debeuse a dous motivos. Por unha parte, os coñecementos aquiridos polo autor na materia *Recuperación da Información* resultaron ser especialmente útiles para as Tarefas T1 e T3, polo que a implementación foi moito máis rápida do previsto. Ademais, dada a inexperiencia do autor coa xestión de proxectos, a planificación errou de conservadora. Deste modo, para corrixir esta desviación e tratar de manter constante a dedicación dos *Sprints*, modificouse a estimación de Puntos para as Historias 1 e 2 de acordo co ocorrido e incorporouse a Historia 3 ao *Sprint Backlog*. Para esta nova Historia, fíxose a seguinte descomposición en Tarefas:

T4 - Estudo do traballo con RMs.

T4.1 - Estudo e asimilación de aspectos teóricos.

T4.2 - Estudo e asimilación do software de David Otero Freijeiro.

A estimación para a Historia 3 resultou acertada. Os conceptos, aínda que tamén coñecidos da materia *Recuperación da Información*, supoñían unha maior complexidade e o tempo requerido para a súa asimilación foi en concordancia. Ademais, o tratar cun código totalmente descoñecido implicou un proceso de análise tamén custoso en tempo.

Sprint Review

A estimación inicial foi demasiado conservadora en canto aos Puntos de Historia. Por iso, as Historias 1 e 2 completáronse antes da finalización do *Sprint*. Para manter a dedicación constante, incorporouse a Historia 3 ao *Sprint Backlog*, tendo feitas as correspondentes modificacións nas estimacións. Deste xeito, a Figura 7.3 amosa as Historias finalmente completadas no *Sprint* 1, así como os Puntos reais. A Figura 7.4 ilustra o progreso do proxecto ao remate desta iteración, podendo observar como se reflexa nos puntos asignados a modificación realizada. Neste punto, a *velocity* do proxecto é a adecuada, manténdose nos 60 Puntos de Historia por *Sprint*.

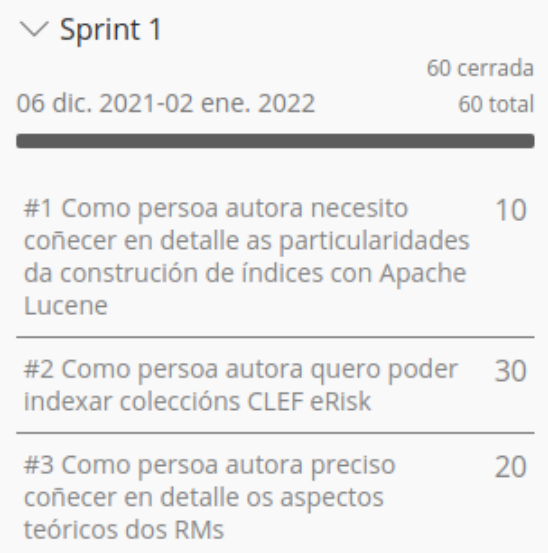

Figura 7.3: Historias finalmente completadas no *Sprint* 1.

7.2.2 *Sprint* 2: Cómputo dos RMs

Superados o proceso inicial de indexación das coleccións e o estudo e asimilación de conceptos teóricos de base, o *Sprint* 2 supuxo a toma de contacto co software de terceiros empregado para este proxecto. En concreto, trátase dun código implementado por David Otero Freijeiro no que se fai uso da libraría Apache Lucene (ver 3.2.3).

Deste xeito, seleccionáronse as Historias 4 e 5 (ver Figura 7.5). Cómpre destacar que

os Puntos estimados para as Historias asignadas non agregan 60, cantidade que se debería acadadar para, de acordo co establecido anteriormente, manter unha *velocity* constante. Isto fíxose de xeito premeditado ao prever un descenso no tempo dispoñible do autor durante as datas do *Sprint*, xa que estas coincidían co período de exames. Así, realizouse a seguinte identificación de Tarefas:

T5 - Adaptación do software de David Otero Freijeiro.

T5.1 - Limpeza das partes non usadas.

T5.2 - Adaptación da estimación de pesos.

T5.2.1 - Axuste *relevance set*.

T5.2.2 - Axuste *query likelyhood*.

T5.3 - Adaptación dos *stats providers*.

T5.4 - Creación dun *main* para a execución de experimentos.

Sprint Review

A planificación inicial tivo éxito na súa premisa, xa que se confirmou a menor dispoñibilidade do autor. Isto fixo que para este *Sprint* tan só se puideran completar 40 Puntos de Historia, o que ocasionou unha desviación na planificación do proxecto. De cara a corrixir a desviación detectada, nos seguintes *Sprints* deberán ser adoptadas medidas para poder minimizar os efectos. Tendo todo isto en conta, a Figura 7.6 amosa as Historias finalmente completadas, que son as previstas inicialmente (Historias 4 e 5). De xeito complementario, a Figura 7.7 amosa o estado do proxecto ao remate do *Sprint* 2, onde se pode observar a desviación no progreso e o descenso nos Puntos/*Sprint*.

7.2.3 *Sprint* 3: Análise e comparativa dos vocabularios de depresión

De xeito similar ao sucedido no *Sprint* 2, quíxose ser realista na previsión do tempo dispoñible para adicar ao proxecto, tratando de adiantarse ás posibles eventualidades derivadas do comezo das prácticas en empresa por parte do autor. Resultaba novamente razoable pensar que non ía ser posible completar os 60 Puntos do caso ideal, polo que se decidiu analizar o *Product Backlog* e seleccionar as Historias 8, 9 e 10 para seren incorporadas ao *Sprint Backlog*. Deste xeito, a Figura 7.8 ilustra a estimación inicial de Historias para o *Sprint 3*. Como é habitual, fíxose a pertinente identificación de Tarefas, definindo as seguintes:

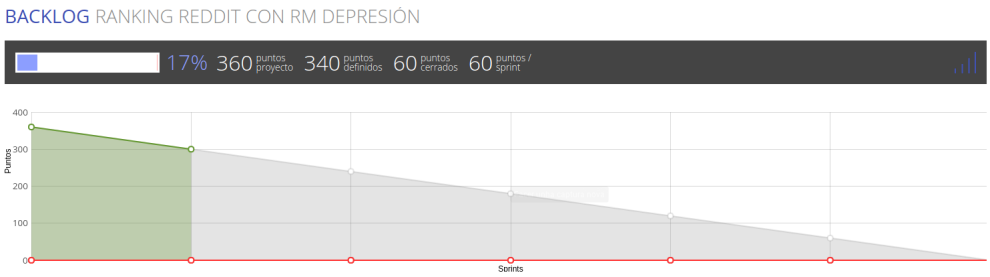

Figura 7.4: Progreso do proxecto ao peche do *Sprint* 1.

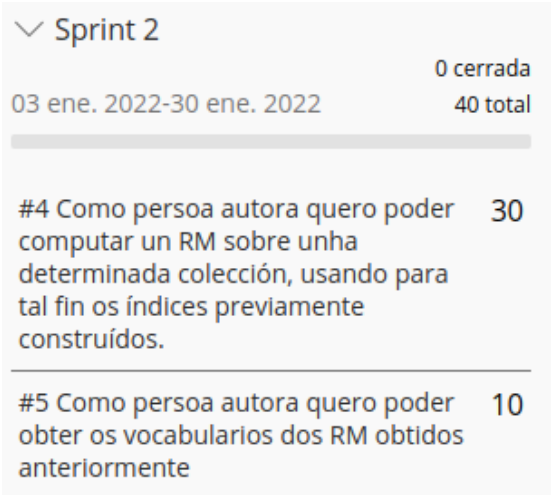

Figura 7.5: Planificación inicial do *Sprint* 2.

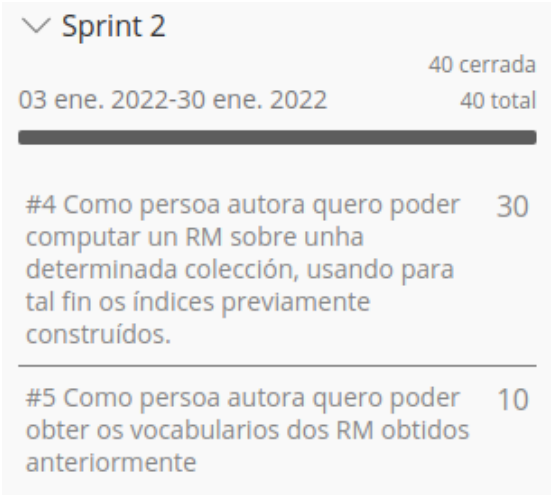

Figura 7.6: Historias finalmente completadas no *Sprint* 2.

T6 - Implementación en Python dun útil para contar as ocorrencias das familias de pronomes nun determinado vocabulario de depresión.

T7 - Comparativa dos vocabularios de depresión con *lexicons* de referencia.

T7.1 - Traballo sobre os *lexicons* de referencia: Pedesis e Choudhury.

T7.1.1 - Estudo da elaboración e características.

T7.1.2 - Unión de termos e preprocesado.

T7.2 - Avaliación da semellanza entre os vocabularios de depresión dos RMs e os *lexicons*, efectuando medicións comparativas dentro dun *top* n de termos.

Sprint Review

Tal e como se previu nun comezo, a dedicación durante o *Sprint* non puido ser a óptima, polo que tan só se puideron completar 50 Puntos de Historia. Isto queda reflexado na Figura 7.9, que recolle as Historias finalmente completadas durante este *Sprint*. Complementariamente, a Figura 7.10 amosa o estado do proxecto ao final desta iteración. Pode observarse que a *velocity* continúa por debaixo do desexable e que a desviación se fai cada vez máis patente. Isto tratará de corrixirse nos seguintes *Sprints*, nos que o autor volverá a ter dispoñible o tempo habitual.

7.2.4 *Sprint* 4: Elaboración dun *dataset* propio.

O propósito para este *Sprint* foi a eleboración dun *dataset* propio no contexto de Reddit. Isto fíxose tratando de obter unha nova colección que permitise ter unha mostra aleatoria do dominio co que se está a traballar. Así, o obxectivo deste *Sprint* era crear un conxunto de datos que recolla todas as publicacións de diferentes perfís aleatorios de Reddit, con independencia total da súa condición depresiva. Por tanto, e tras revisar o *Product Backlog*, seleccionáronse as Historias 6 e 7. Tamén, tratando de corrixir a desviación no progreso do proxecto, decidiuse aumentar a dedicación por parte do autor, de modo que se incorporou tamén a Historia 11 ao *Sprint Backlog*. A Figura 7.11 ilustra a estimación inicial do *Sprint*. Tamén, e de acordo co habitual, fíxose a habitual identificación de tarefas:

T8 - Estudo do estado da arte para os *datasets* de Reddit.

T8.1 - Busca de fontes de información.

T8.2 - Estudo e asimilación da información previamente seleccionada.

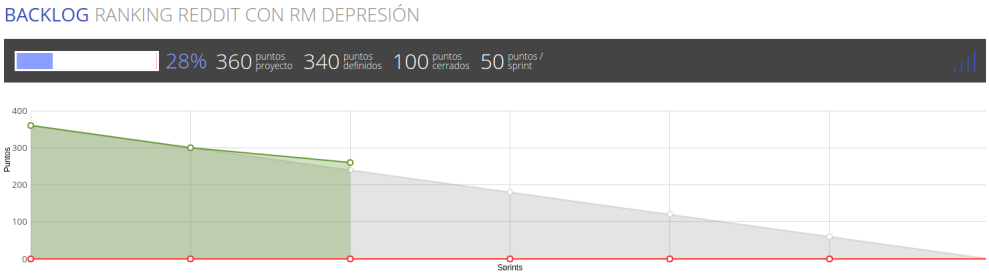

Figura 7.7: Progreso do proxecto ao peche do *Sprint* 2.

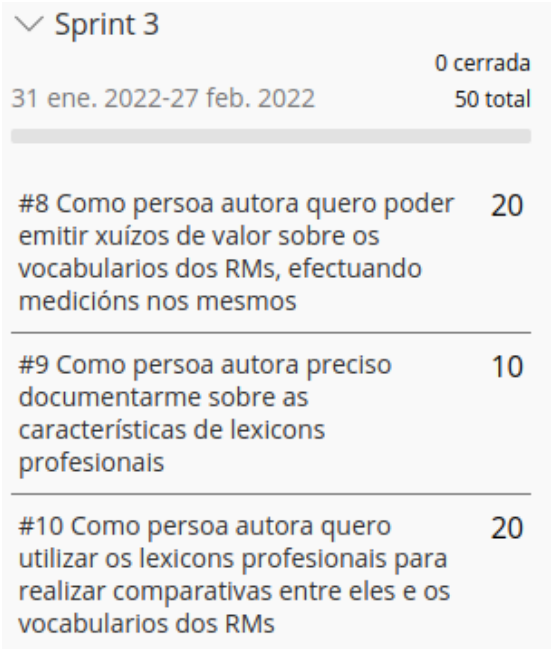

Figura 7.8: Planificación inicial do *Sprint* 3.

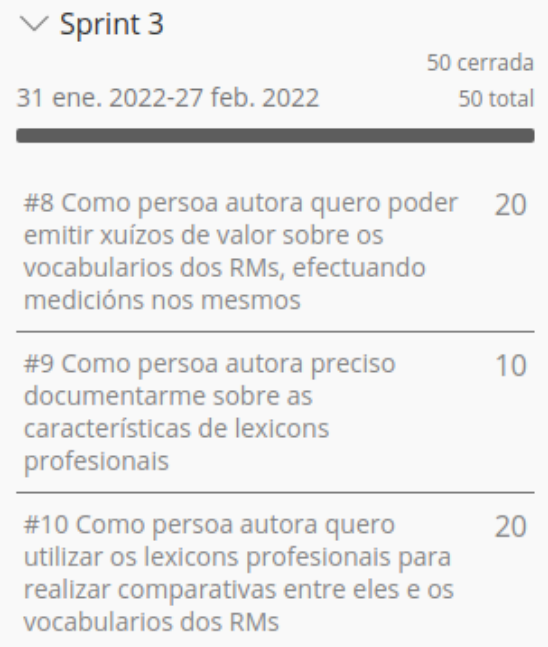

Figura 7.9: Historias finalmente completadas no *Sprint* 3.

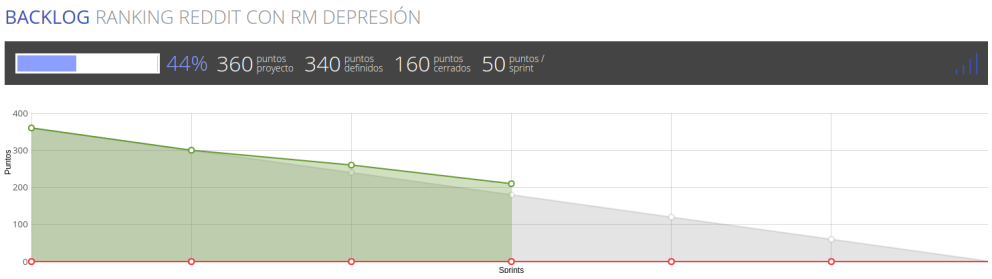

Figura 7.10: Progreso do proxecto ao peche do *Sprint* 3.

T9 - Elaboración da ferramenta para obter o *dataset*.

T9.1 - Deseño.

T9.2 - Implementación.

T9.3 - Probas.

T10 - Ordenación dos termos dos *lexicons* utilizando a información dos vocabularios de depresión dos RMs e os pesos asociados.

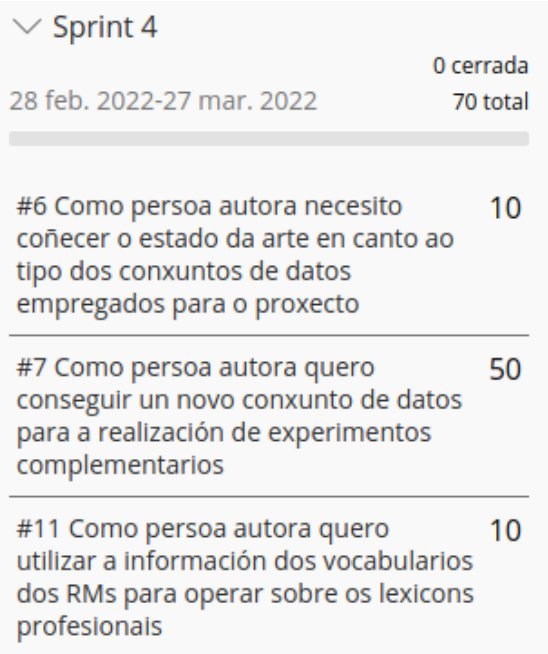

Figura 7.11: Planificación inicial do *Sprint* 4.

Sprint Review

A elaboración do *dataset* resultou ser máis custosa en tempo do previsto. Por isto, e aínda aumentando as horas de traballo durante este *Sprint*, non se puido completar a Historia 11. Deste modo, esta Historia devolveuse ao *Product Backlog* para ser elixible noutro *Sprint* e realizouse a modificación correspondente nos Puntos asignados para a Historia 7. A Figura 7.12 ilustra as Historias finalmente completadas durante o *Sprint* 4, xunto cos Puntos reais. Complementariamente, a figura 7.13 serve para ilustrar o estado do proxecto ao remate deste *Sprint*, podendo ver o aumento nos puntos definidos respecto ao *Sprint* 3 e unha pequena mellora na *velocity*, xa máis próxima aos 60 Puntos de Historia por Sprint.

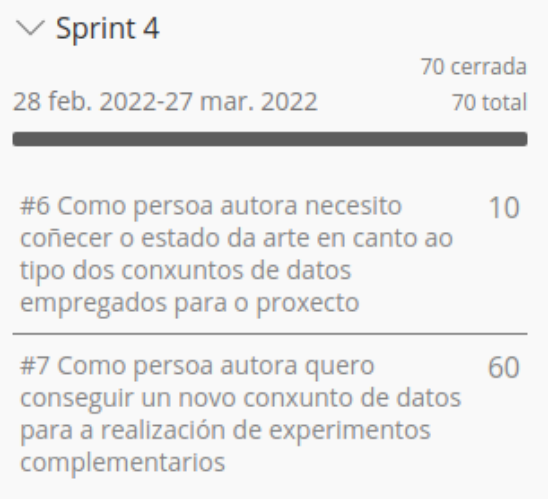

Figura 7.12: Historias finalmente completadas no *Sprint* 4.

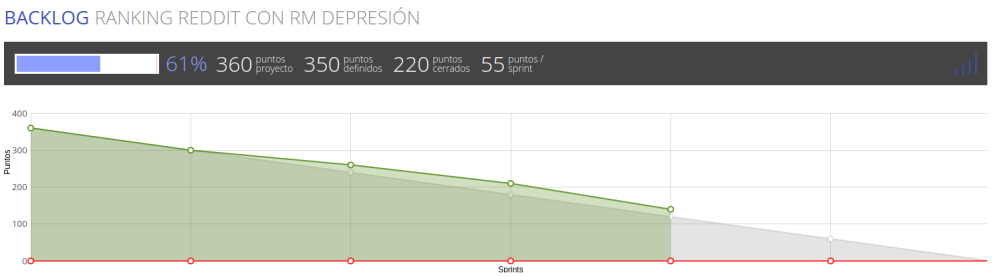

Figura 7.13: Progreso do proxecto ao peche do *Sprint* 4.

7.2.5 *Sprint* 5: Definición de *baselines* e preparación do *ranking*

Chegado o *Sprint* 5, o proxecto encaraba a súa parte final, centrada xa no *ranking*. Continuando coas medidas na procura de corrixir a desviación do proxecto, decidiuse novamente aumentar a dedicación durante o tempo que durara o *Sprint*, polo que tras analizar o *Product Backlog* se seleccionaron as Historias 11, 12 e 13 para o *Sprint Backlog*. Así, a Figura 7.14 amosa a planificación inicial do *Sprint*, para o que se identificaron as seguintes Tarefas:

T10 - Ordenación dos termos dos *lexicons* utilizando a información dos vocabularios de depresión dos RMs e os pesos asociados.

T11 - Optimización e validación de parámetros para as *queries* de *baseline*.

T11.1 - Definición dos termos para as *queries* de *baseline*.

T11.2 - Implementación do código para a optimización de parámetros.

T11.3 - Implementación do código para a validación de parámetros.

T11.4 - Obtención e avaliación dos resultados obtidos, aplicando diversas medicións.

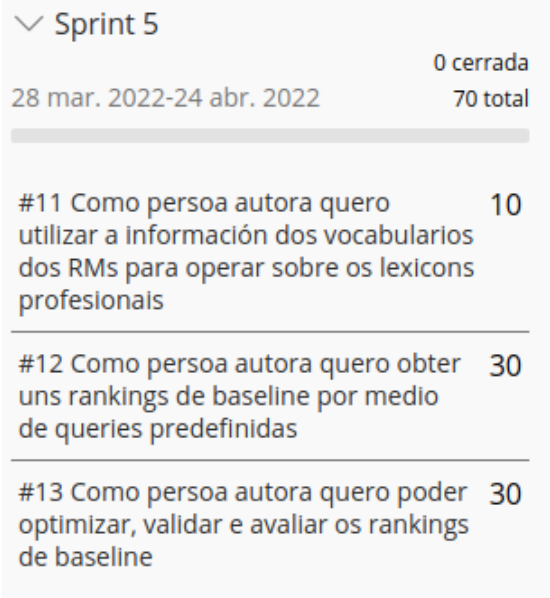

Figura 7.14: Planificación inicial do *Sprint* 5.

Sprint Review

As estimacións feitas na planificación do *Sprint* foron exitosas, de modo que tanto os Puntos das Historias coma o tempo adicado por parte do autor transcorreron conforme o previsto.

Isto permitiu pechar o *Sprint* conforme ao ilustrado na Figura 7.15, e sanear a situación do proxecto tal e como se reflexa na Figura 7.16. Pode verse como o avance do proxecto se aproxima ao ideal, así como a mellora da *velocity*, que con 58 Puntos de Historia por *Sprint* representa un valor próximo ao óptimo, fixado en 60.

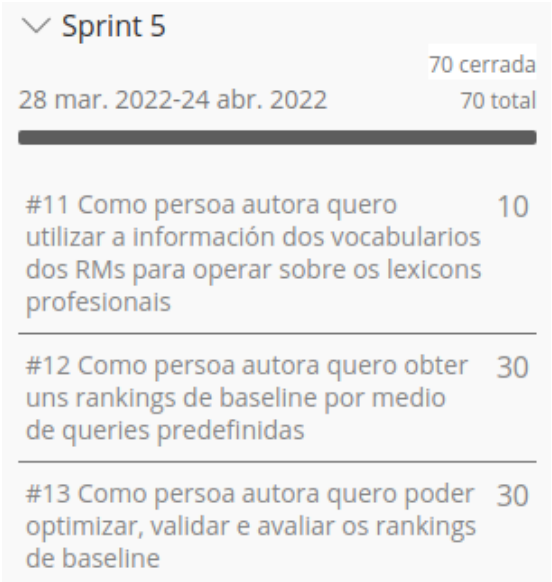

Figura 7.15: Historias finalmente completadas no *Sprint* 5.

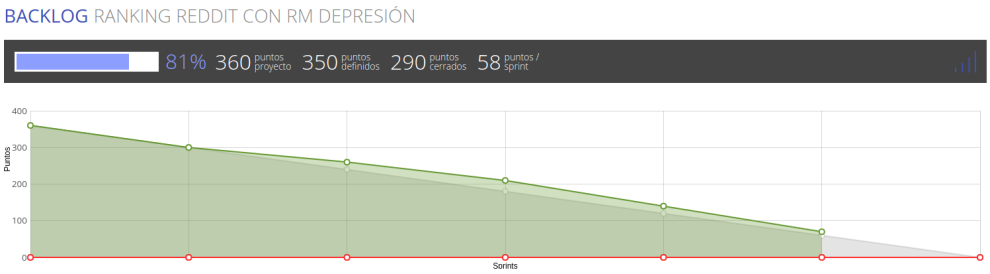

Figura 7.16: Progreso do proxecto ao peche do *Sprint* 5.

7.2.6 *Sprint* 6: *Ranking* e avaliación

O *Sprint* 6 debería ser, de acordo coa planificación, o incremento final do proxecto. Neste punto do desenvolvemento, tan só quedaban as Historias 14 e 15 no *Product Backlog*. Pode apreciarse na Figura 7.17 como foron estas Historias as seleccionadas para formar o *Sprint Backlog*.

T12 - Optimización e validación de parámetros para as *queries* manuais.

T12.1 - Adaptación do código para a optimización de parámetros, tendo en conta as particularidades do experimento.

T12.2 - Adaptación do código para a validación de parámetros, tendo en conta as particularidades do experimento.

T12.3 - Obtención e avaliación dos resultados obtidos, aplicando diversas medicións.

T12.3.1 - Utilizando a ordenación dos *lexicons* feita na tarefa **T10**.

T12.3.2 - Utilizando directamente os vocabularios de depresión dos RMs.

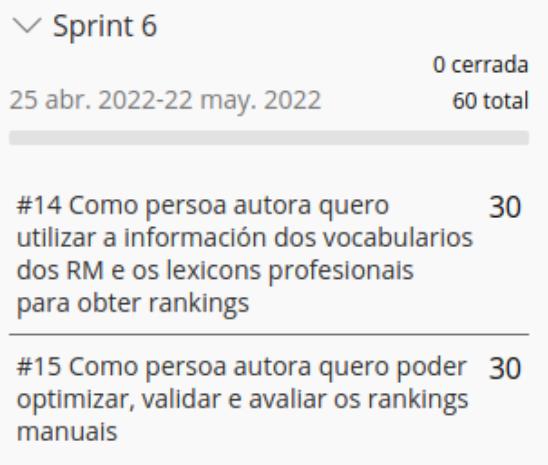

Figura 7.17: Planificación inicial do *Sprint* 6.

Sprint Review

O incremento previo, correspondente ao *Sprint* 5, supuxo unha moi boa base para o traballo desta iteración, polo que o traballo das Historias se completou mediado o transcurso do *Sprint*. Así, fíxose unha asignación de Puntos reais, tal e como se pode ver na Figura 7.18. Estas Historias tan só supuxeron 30 Puntos, pero dado que eran as derradeiras do *Product Backlog*, o proxecto chegou á súa fin e o *Sprint* pechouse con 30 Puntos. A Figura 7.19 ilustra esta situación, podendo apreciar a modificación realizada nos Puntos definidos.

7.2.7 Balance final

Dada a pouca experiencia do autor coa xestión de proxectos e as circunstancias académicas nas que se desenvolveu o traballo, existiron certos erros nas estimacións e premisas

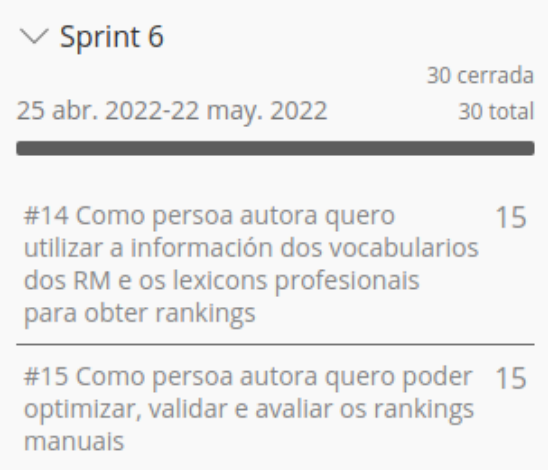

Figura 7.18: Historias finalmente completadas no *Sprint* 6.

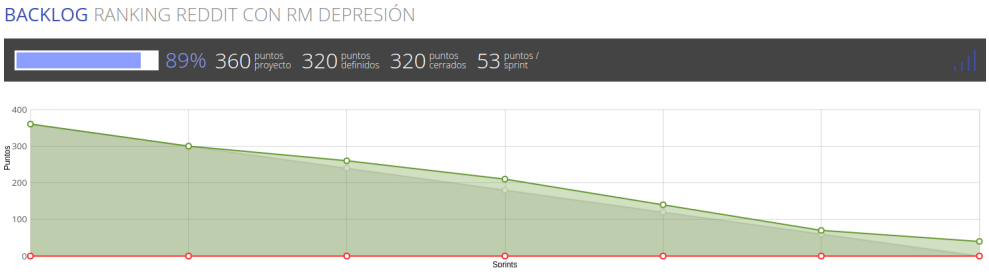

Figura 7.19: Progreso do proxecto ao peche do *Sprint* 6.

establecidas. Non obstante, de acorodo co explicado nos correspondentes *Sprints*, tratouse de corrixir as eventualidades xurdidas. Nesta liña, na Táboa 7.2 pode verse a asignación real de Puntos para cada unha das Historias de Persoa Usuaria. Complementariamente, a Táboa 7.3 detalla a relación existente as Historias, as Tarefas e os *Sprints*. Reflexionando sobre o desenvolvemento do proxecto, cómpre destacar algúns detalles:

- O proxecto completouse no tempo establecido e de acordo cos *Sprints* previstos.
- A duración dos *Sprints* mantívose constante nas 4 semanas fixadas por iteración.
- O proxecto completouse con menos Puntos de Historia definidos (320) dos inicialmente estimados (360).
- A *velocity* non foi constante, pero mantívose sempre próxima ao valor ideal de 60 (Puntos/*Sprint*).

Táboa 7.2: Historias de Persoa Usuaria e puntos reais.

ID	Historia de Persoa Usuaria	Puntos
1	Como *persoa autora* necesito coñecer en detalle as particularidades da construción de índices con Apache Lucene (ver 3.2.3).	20 -> 10
2	Como *persoa autora* quero poder indexar coleccións CLEF eRisk.	40 -> 30
3	Como *persoa autora* preciso coñecer en detalle os aspectos teóricos dos RMs.	20
4	Como *persoa autora* quero poder computar un RM sobre unha determinada colección, usando para tal fin os índices previamente construídos.	30
5	Como *persoa autora* quero poder obter os vocabularios de depresión dos RM obtidos anteriormente.	10
6	Como *persoa autora* necesito coñecer o estado da arte en canto ao tipo dos conxuntos de datos empregados para o proxecto.	10
7	Como *persoa autora* quero conseguir un novo conxunto de datos para a realización de experimentos complementarios.	50 -> 60

. . . . . *(continúa na páxina seguinte)* . . . . . .

Táboa 7.2 – *(vén da páxina anterior)*

ID	Historia de Persoa Usuaria	Puntos
8	Como *persoa autora* quero poder emitir xuízos de valor sobre os vocabularios de depresión dos RMs, efectuando medicións nos mesmos.	20
9	Como *persoa autora* preciso documentarme sobre as características de *lexicons* profesionais.	10
10	Como *persoa autora* quero utilizar os *lexicons* profesionais para realizar comparativas entre eles e os vocabularios de depresión dos RMs.	20
11	Como *persoa autora* quero utilizar a información dos vocabularios de depresión dos RMs para operar sobre os *lexicons* profesionais.	10
12	Como *persoa autora* quero obter uns *rankings* de *baseline* por medio de *queries* predefinidas.	30
13	Como *persoa autora* quero poder optimizar, validar e avaliar os *rankings* de *baseline*.	30
14	Como *persoa autora* quero utilizar a información dos vocabularios de depresión dos RM e os *lexicons* profesionais para obter *rankings*.	30 -> 15
15	Como *persoa autora* quero poder optimizar, validar e avaliar os *rankings* manuais.	30 -> 15

Sprint	Tarefas	Historias
1 (7.2.1)	T1.1, T1.2, T1.3, T2.1, T2.2, T2.3, T3.1, T3.2.1, T3.2.2, T4.1, T4.2	1, 2, 3
2 (7.2.2)	T5.1, T5.2.1, T5.2.2, T5.3, T5.4	4, 5
3 (7.2.3)	T6, T7.1.1, T7.1.2, T7.2	8, 9, 10
4 (7.2.4)	T8.1, T8.2, T9.1, T9.2, T9.3	6, 7
5 (7.2.5)	T10, T11.1, T11.2, T11.3, T11.4	11, 12, 13
6 (7.2.6)	T12.1, T12.2, T12.3.1, T12.3.2	14, 15

Táboa 7.3: Relación entre Historias, Tarefas e *Sprints*.

Capítulo 8

Conclusións e traballo futuro

Este capítulo pretende ser un espazo para a reflexión e a análise do proxecto unha vez este chegou á súa fin. Así, primeiramente, obtéñense unha serie de conclusións, valorando a consecución ou non dos obxectivos establecidos nas fases iniciais. Logo, tamén se fai unha avaliación dos resultados obtidos. Finalmente, establécense unha serie de vías de experimentación futuras.

8.1 Conclusións

Unha boa maneira de tirar conclusións á finalización dun proxecto é revisar o éxito na consecución dos obxectivos fixados. Separarase esta análise tendo en conta, por un lado, os resultados do proxecto e tamén o aspecto formativo.

Resultados do proxecto

- Puideron computarse con éxito os RMs e obter os seus vocabularios de depresión. Complementariamente, estes puideron ser tanto analizados de maneira individual como comparados con *lexicons* de referencia. Esta comparativa resultou sorprendente, pois o *solapamento* entre os *lexicons* e os vocabularios de depresión foi menor do inicialmente agardado. Non obstante, estes vocabularios permitiron despois obter os mellores resultados en *ranking*,
- Foi posible elaborar o *ranking* de suxeitos. Feito isto, tamén se conseguiu avalialo e comparalo cunha serie de medicións de referencia, denominadas *baselines*. Esta comparativa resultou moi exitosa, pois viuse que os resultados de *ranking* obtidos superaban o valor de base.

Aspecto formativo

- Os fundamentos de base e os aspectos teóricos do proxecto, así como as características das tecnoloxías e ferramentas usadas, foron estudados de maneira adecuada nos tempos marcados. Isto permitiu que o autor repasara conceptos xa adquiridos con anterioridade e tamén que se formase en novos eidos.
- Resultou moi enriquecedor o uso dunha metodoloxía, así como a responsabilidade na planificación. O autor, pese a algunhas dificultades iniciais, puido gañar experiencia nun campo totalmente descoñecido para el.
- O autor puido aprender a desenvolverse dun xeito máis autónomo, pautando os ritmos e xestionando problemas que foron aparecendo durante o transcurso do proxecto. Ademais, este traballo serviu como exemplo simplificado dun proxecto de investigación, satisfacendo así a curiosidade do autor.

8.2 Traballo futuro

Durante a realización do proxecto, ocorreu en máis dunha ocasión que xurdiron ideas de cara á posible realización de novos experimentos. Entre estas ideas poderíanse destacar as seguintes, que se consideran novas vías de investigación:

- Aplicación doutro tipo de LMs, principalmente Maximum Entropy Divergence Minimization Model (MEDMM) [41], buscando ver como se comporta este tipo de modelos máis recentes e fortemente caracterizados polos principios estatísticos que os rexen.
- Uso de *contextualized word-embeddings* [42] a través de modelos de linguaxe neuronais como BERT [43], representando os termos do vocabulario como vectores de características. Para isto será necesario reformular a estimación dos RMs para ter estas representacións compactas.

Apéndices

Apéndice A

Termos dos *lexicons*

A.1 Termos únicos de Pedesis

abandon, absorb, abysmal, acute, afraid, alone, aloof, anxious, appall, arcane, arctic, awful, awkward, bite, bitter, black, bleak, block, blubber, blue, blunt, bore, bottom, bottomless, burn, cavernous, cheerless, chestnut, chilly, cold, colossal, crack, cruel, cry, custodial, dam, dangerous, dark, dead, deep, defenceless, defer, deject, desolate, desperate, despondent, destitute, detach, difficult, dim, dire, dismal, distant, distress, downcast, downhearted, dreadful, dreary, drink, dull, dusk, dusky, embarrass, emotionless, epidemic, evil, excessive, excruciate, extreme, faint, fall, fearful, feeble, fog, forlorn, freeze, friendless, frightful, frosty, fur, ghastly, gloomy, glum, grave, guilty, halt, hard-hearted, heartbreaking, helpless, hidden, hind, hollow, horrendous, horrible, horrific, huge, icy, impend, impenetrable, imperfect, impersonal, incapable, indifferent, ineffective, inferior, inoperative, inscrutable, insufferable, insupportable, intolerable, irritable, joyless, kill, lavish, load, lone, lonely, lonesome, lose, low, melancholy, miserable, miss, monstrous, mourn, multifaceted, murky, nasty, nervous, nightly, nightmarish, nip, nocturnal, numb, obscure, opaque, overpower, painful, penitentiary, pitch-black, poky, poor, powerless, prodigious, quiet, rash, raw, remote, resonant, restless, roar, rotten, run-down, sad, seclude, serious, severe, shadowy, shady, shock, signal, silent, sinister, sink, slow, small, solitary, sombre, sonorous, sore, sorrowful, sour, spooky, stalk, stifle, stony, tedious, tender, terrible, timid, timorous, traumatic, unable, unaccompanied, unbearable, uncomfortable, undulate, uneasy, unendurable, unfathomable, unfeeling, unfortunate, unfriendly, unhappy, unkind, unmitigated, untold, void, vulnerable, waste, weak, weep, wintry, wretched, abyss, ache, agony, aimlessness, anger, annihilation, anteroom, apparition, attic, authority, automaton, bareness, barrage, barrenness, barricade, barrier, basin, bawl, bayou, bear, beast, beckon, bedspread, behemoth, belligerence, bellow, bereavement, bewilderment, birdcage, blacken, blackness, blockade, blur, bog, boulder, break, breaker, burden, burning, cage, capture, casualty, catastrophe, cavity, cell, chamber, chasm, chilliness, choke, cirrocumulus,

cirrus, clink, clog, cloud, clout, coldness, collapse, confinement, confusion, coop, cover, coverlet, crate, crush, cumulonimbus, cumulus, custody, cyclone, damage, darkness, day-dream, daze, death, decease, decline, defeat, dejection, deluge, demise, deprivation, depth, desolation, despair, despondency, destitution, detention, die, dig, dimness, dip, disapproval, disaster, doldrums, doom, downer, downfall, drape, dream, drench, dribble, drill, drizzle, drop, dumper, earth, eclipse, edge, emptiness, enclosure, encumbrance, end, eradication, error, eruption, escape, evening, exclamation, exhaust, expiry, explosion, extermination, extinction, fail, fatality, fate, feather, feel, fen, fence, fiend, filter, fissure, flail, flap, flaw, flicker, flood, floor, flourish, flu, fluff, flutter, foyer, freak, frost, fuddle, funk, fury, futility, fuzz, gag, gap, garrotte, ghost, gloom, gloominess, grief, guilt, gulf, gulp, gumshoe, hail, hall, hallucination, haze, heaviness, heft, hide, hindrance, hole, holler, hollowness, hostility, howl, hurt, iciness, impediment, imprisonment, impulse, incarceration, incubus, indigence, jail, killing, kink, lapse, leak, leviathan, lint, loss, meaninglessness, menace, mesh, miasma, misery, misfortune, mist, monster, monstrosity, morass, mortality, mouthful, movement, muddle, murder, murkiness, nadir, negativeness, night, nightfall, nightmare, night-time, nimbus, obliteration, obscurity, obstacle, obstruction, ogre, ooze, oscillation, outbreak, overlay, overload, overthrow, percolate, pessimism, phantom, plethora, plug, pointlessness, pong, poverty, power, precipitation, prison, privation, puncture, punishment, purposelessness, pursuer, quagmire, quicksand, rage, rain, raindrop, rainfall, rainwater, raven, reformatory, resentment, retreat, riddle, ringlet, ripple, robot, room, ruin, sable, sadness, scream, shack, shade, shadow, shaft, shout, shriek, sieve, sifter, sign, sip, slough, slump, smog, smother, snivel, soak, sob, sorrow, space, sparseness, spate, spectre, spit, spoil, squall, squawk, squeal, stalker, stamp, stem, stench, stink, stomach, suffer, swallow, swamp, tear, tendency, thunder, thundercloud, thunderhead, tornado, tragedy, trickle, tunnel, undulation, unhappiness, upset, upsurge, vacancy, vacuity, vacuum, vapor, vapour, veil, wag, wail, wall, wave, weakness, weight, weightiness, weir, whimper, worthlessness, wraith, wreck, wretchedness, writhe, zombie, abolish, agitate, agonize, alarm, anaesthetize, anguish, annihilate, annoy, asphyxiate, besiege, bewilder, blub, bowl, brandish, bulk, conceal, confuse, congest, consume, darken, deaden, demolish, deplete, descend, desert, destroy, deteriorate, devastate, devour, disburse, dishearten, disorientate, dispirit, drown, dwindle, elude, encase, enclose, endure, engulf, entrench, evade, exclaim, extinguish, frighten, grieve, guzzle, harrow, hinder, horrify, hush, impede, inhibit, inundate, irritate, isolate, limit, lock, loom, mislay, misplace, muffle, mute, nominate, obliterate, obstruct, oppress, overcome, overlie, overshadow, overwhelm, penetrate, permeate, perplex, petrify, purify, pursue, quash, recant, reject, re-press, rescind, restrict, retract, rumble, scare, sift, soothe, spur, startle, suffocate, suppress, terminate, terrify, threaten, traumatize, trouble, worry, worsen, yawn, yell, yelp

A.2 Termos únicos de Choudhury

addictive, adhd, agree, amaze, answer, antidepressant, anxiety, appetite, attack, beautiful, bible, blur, boy, care, chemical, church, clinical, cope, counteract, date, delusion, diagnosis, discuss, dizziness, doctor, dose, drowsiness, drowsy, drug, dysfunction, effective, enjoy, episode, father, fatigue, favourite, friend, fun, game, girl, god, haha, hate, headache, heaven, hell, helpful, help, hold, home, hospitalization, house, imbalance, inhibiter, insomnia, irritability, jesus, leave, life, like, lord, love, man, medication, movie, music, nausea, nervousness, neurotransmitters, party, patient, pill, play, prescribe, prescription, psychosis, psychotherapy, relationship, religion, save, season, sedative, seizure, severe, sexual, side-effects, sleep, social, song, sorry, stimulant, style, suffer, suicidal, swing, talk, therapy, tolerance, toxicity, want, wean, weight, win, withdrawal, woman, young

A.3 Adxectivos non ambiguos de Choudhury

addictive, amaze, helpful, party, sedative, stimulant, suicidal

A.4 Adxectivos non ambiguos de Pedesis

absorb, abysmal, acute, afraid, alone, aloof, anxious, appall, arcane, awful, awkward, bleak, blunt, bottomless, cavernous, cheerless, chilly, colossal, cruel, custodial, dangerous, defenceless, defer, deject, desolate, desperate, despondent, destitute, detach, difficult, dim, dire, dismal, distant, downcast, downhearted, dreadful, dreary, dull, dusk, dusky, embarrass, emotionless, epidemic, excessive, excruciate, faint, fearful, feeble, forlorn, friendless, frightful, frosty, ghastly, gloomy, glum, guilty, hard-hearted, heartbreaking, helpless, hidden, horrendous, horrible, horrific, huge, icy, impend, impenetrable, imperfect, impersonal, incapable, indifferent, ineffective, inoperative, inscrutable, insufferable, insupportable, intolerable, irritable, joyless, lavish, lone, lonely, lonesome, lose, miserable, monstrous, mourn, multifaceted, murky, nasty, nervous, nightly, nightmarish, nocturnal, numb, obscure, opaque, overpower, painful, penitentiary, pitch-black, poky, poor, powerless, prodigious, raw, remote, resonant, restless, rotten, run-down, sad, seclude, serious, severe, shadowy, shady, silent, sinister, slow, sombre, sonorous, sore, sorrowful, spooky, stifle, stony, tedious, terrible, timid, timorous, traumatic, unable, unaccompanied, unbearable, uncomfortable, undulate, uneasy, unendurable, unfathomable, unfeeling, unfortunate, unfriendly, unhappy, unkind, unmitigated, untold, vulnerable, weak, weep, wintry, wretched

A.5 Adxectivos non ambiguos de Choudhury expandido co método *Distribucional*

addictive, amaze, helpful, party, sedative, stimulant, suicidal, club, dazzle, strongholds, stun, surprise, troop

A.6 Adxectivos non ambiguos de Choudhury expandido co método *WordNet*

addictive, amaze, ataractic, ataraxic, habit-forming, helpful, party, sedative, self-destructive, stimulant, stimulating, suicidal, tranquilising, tranquilizing, tranquillising, tranquillizing

A.7 Adxectivos non ambiguos de Pedesis expandido co método *Distribucional*

absorb, abysmal, accelerate, acute, adsorb, affect, afraid, alleviate, alone, aloof, anger, anxious, appall, arcane, ask, avoid, awful, awkward, beat, bestow, bleak, blotched, blunt, bottomless, bruise, cancel, capture, carry, cause, cavernous, cdot, characterise, characterize, cheerless, chilly, clinch, collapse, colossal, colour, confront, conquer, convert, convince, cruel, cry, custodial, dangerous, decline, defeat, defenceless, defer, define, deject, delay, denote, depopulate, derive, desolate, desperate, despondent, destitute, destroy, detach, detect, devastate, devote, difficult, dim, diminish, dire, disappear, disappoint, dismal, distant, divide, downcast, downhearted, dreadful, dreary, dull, dusk, dusky, elongate, embarrass, emit, emotionless, encircle, enclose, encourage, enlarge, epidemic, erode, evaporate, evoke, evolve, exacerbate, excessive, exclude, excruciate, exercise, extract, facilitate, fade, faint, fearful, feeble, fill, finish, flank, flatten, fleck, focus, foil, forlorn, forward, friendless, frightful, frosty, ghastly, gloomy, glum, grab, grieve, guilty, halt, hamper, hard-hearted, hawthorn, heal, heartbreaking, helpless, hidden, hinder, hope, horrendous, horrible, horrific, huge, icy, impede, impend, impenetrable, imperfect, impersonal, imply, impress, incapable, indifferent, induce, ineffective, infuse, inject, innervate, inoperative, inscrutable, insufferable, insupportable, intolerable, invade, ionize, irritable, isolate, joyless, kill, lavish, leach, lone, lonely, lonesome, lose, metabolize, minimize, miserable, monstrous, mourn, multifaceted, murky, nasty, nervous, nightly, nightmarish, nocturnal, numb, obscure, opaque, opt, orange-red, outflank, outrage, overhang, overpower, owe, oxidise, oxidize, pacify, painful, peasantry, penetrate, penitentiary, pertain, pitch-black, plan, poky, poor, postpone, powerless, pray, prepare, present, prevent, prodigious, protrude, ravage, raw, react, refer, relate, remote, remove, repel, repulse, reschedule, resonant, respond, restless, revere, reward, rotten, run-down, sad, satisfy, schedule, seclude,

seedling, seep, send, separate, serious, severe, shadowy, shady, sharpen, shock, shower, silent, sinister, slate, slow, sombre, sonorous, soothe, sore, sorrowful, speckle, spooky, stifle, stony, stop, streak, strive, subdue, subjugate, submit, surprise, surround, swell, taper, tedious, tell, terrible, thwart, timid, timorous, ting, transform, traumatic, traverse, treat, tremble, turn, unable, unaccompanied, unbearable, uncomfortable, undulate, uneasy, unendurable, unfathomable, unfeeling, unfortunate, unfriendly, unhappy, unkind, unmitigated, untold, urinate, vaporize, venerate, vine, vomit, vulnerable, wait, wane, weak, weep, wield, win, wintry, wish, worship, wretched, yearn

A.8 Adxectivos non ambiguos de Pedesis expandido co método *WordNet*

abject, abominable, absorb, abysmal, abyssal, acuate, acute, afflictive, aflutter, afraid, alone, aloof, amazing, anxious, apathetic, appall, arcane, arctic, atrocious, austere, awed, awe-inspiring, awesome, awful, awing, awkward, baleful, barbarous, bare-assed, bare-ass, bare, barren, benumbed, black, bleak, bloodcurdling, blue, blunt, boring, bottomless, bouldered, bouldery, Brobdingnagian, brutal, bunglesome, candid, cavernous, charnel, cheerless, chilly, cloudy, clumsy, colossal, concealed, cranky, crappy, creaky, crisp, cruddy, crude, cruel, cryptical, cryptic, cumbersome, custodial, cutting, dangerous, dark, dark-skinned, deadening, debile, decayed, decrepit, deep, defenceless, defenseless, defer, deject, dense, deplorable, depressed, derelict, desolate, despairing, desperate, despicable, despondent, destitute, detach, devoid, difficult, dilatory, dim, dingy, direful, dire, disconsolate, discriminating, dismal, dispirited, distant, distressed, distressing, dour, downcast, downhearted, drab, dreaded, dreadful, drear, dreary, dull, dumb, dusk, dusky, embarrassing, embarrass, emotionless, epidemic, erectile, evocative, exceeding, exceptional, excessive, excitable, excruciate, execrable, extravagant, exuberant, faint-hearted, fainthearted, faint, fallible, fearful, fearsome, feeble, fell, filthy, fishy, flagitious, flea-bitten, flighty, flint, flinty, fly-by-night, forbidding, forlorn, forthright, foul, fractious, frail, frank, free, free-spoken, friendless, frightening, frightful, frigid, frosty, frozen, funny, gelid, ghastly, glacial, gloomful, glooming, gloomy, glowering, glum, good, granitelike, granitic, grave, grievous, grim, grisly, grotesque, gruesome, guilty, hair-raising, hangdog, hapless, hard-hearted, hardhearted, hard, heartbreaking, heartless, heartrending, heartsick, heavy, helpless, heroic, hidden, hideous, ho-hum, horrendous, horrible, horrid, horrific, horrifying, hostile, huffy, huge, icky, icy, ill-chosen, immaterial, immense, impend, impenetrable, imperfect, impersonal, impossible, impoverished, inadequate, inapt, incapable, incapacitated, incisive, incompetent, indefensible, indifferent, indigent, ineffective, ineffectual, inefficient, inept, inert, infelicitous, infirm, inimical, innocent, inoperative, inordinate, inscrutable, insufferable, insupportable, intense, intolerable, irksome, irritable, irritating, iso-

lated, jerkwater, joyless, keen, knifelike, knockout, laggard, lame, lamentable, lavish, leaden, life-threatening, light-headed, lightheaded, light, lone, lonely, lonesome, long-winded, lose, lost, louche, lousy, low-down, low, low-spirited, lucullan, lush, macabre, mad, many-sided, measly, melancholy, menacing, minacious, minatory, mirky, miscellaneous, miserable, misfortunate, monstrous, moody, morose, mourn, muddy, muffled, multifaceted, multifarious, mum, munificent, murky, muted, mute, mysterious, mystifying, naked, nasty, natural, necessitous, needlelike, needy, nerveless, nervous, nettlesome, neural, neutral, new, nightly, nightmarish, nipping, nippy, nocturnal, numb, obdurate, obscure, obtuse, olympian, ominous, one-horse, only, opaque, outcast, outrageous, outside, outspoken, overgenerous, overpower, overweening, painful, paltry, parky, passionless, pathetic, peckish, peeled, peevish, penetrating, penetrative, penitential, penitentiary, pettish, petulant, piercing, pitch-black, pitch-dark, piteous, pitiable, pitiful, pitiless, plainspoken, plush, plushy, point-blank, pokey, poky, polar, poor, portentous, poverty-stricken, powerless, prodigious, queasy, raw, redolent, remindful, reminiscent, remote, removed, resonant, resonating, resounding, restless, reverberating, reverberative, rickety, rimed, rimy, rocklike, rocky, rotted, rotten, roughshod, run-down, sad, sapless, saturnine, savage, scratchy, scummy, scurvy, seclude, secret, sensitive, sepulchral, serious, severe, shadowed, shadowy, shady, shamed, shamefaced, sharp, shitty, short, shy, sick, silent, sinister, skittish, slimy, slow, sluggish, smutty, snappy, sober, softened, solitary, somber, sombre, sonorous, sore, sorrowful, sorry, soundless, sour, spartan, spooky, stark, stern, sticky, stifle, still, stinking, stinky, stonyhearted, stony, straight-from-the-shoulder, stupendous, subdued, suffering, sulky, sullen, surpassing, suspect, suspicious, swarthy, swart, swooning, tacit, techy, tedious, tender, terrible, testy, tetchy, threatening, thudding, tight, timid, timorous, tiresome, too-generous, traumatic, tremendous, trepid, turbid, tutelar, tutelary, twilit, ugly, umbrageous, unable, unaccented, unacceptable, unaccompanied, unbearable, unbiased, unbiassed, uncheerful, uncomfortable, uncongenial, understood, undue, undulate, uneasy, uneffective, unendurable, unenviable, unequaled, unequalled, unfathomable, unfeeling, unfortunate, unfrequented, unfriendly, ungainly, ungratified, unhappy, unintelligible, unique, unjustifiable, unkind, unknown, unmitigated, unnoticeable, unparalleled, unquiet, unreasonable, unsafe, unsanded, unsatisfied, unsounded, unsparing, unspeakable, unstinted, unstinting, unsufferable, unsung, unsure, untold, unwarrantable, unwarranted, unworthy, upstage, vague, vast, verbose, vicious, vile, vulnerable, washy, watery, weak, weakly, wearisome, weep, wicked, windy, wintery, wintry, wispy, woebegone, woeful, wordy, worthless, wraithlike, wretched

Apéndice B

Inventario BDI-II

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel. If several statements in the group seem to apply equally well, choose the highest number for that group.

1. Sadness

1. I do not feel sad.
2. I feel sad much of the time.
3. I am sad all the time.
4. I am so sad or unhappy that I can't stand it.

2. Pessimism

1. I am not discouraged about my future.
2. I feel more discouraged about my future than I used to be.
3. I do not expect things to work out for me.
4. I feel my future is hopeless and will only get worse.

3. Past Failure

1. I do not feel like a failure.
2. I have failed more than I should have.

3. As I look back, I see a lot of failures.
4. I feel I am a total failure as a person.

4. Loss of Pleasure

1. I get as much pleasure as I ever did from the things I enjoy.
2. I don't enjoy things as much as I used to.
3. I get very little pleasure from the things I used to enjoy.
4. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings

1. I don't feel particularly guilty.
2. I feel guilty over many things I have done or should have done.
3. I feel quite guilty most of the time.
4. I feel guilty all of the time.

6. Punishment Feelings

1. I don't feel I am being punished.
2. I feel I may be punished.
3. I expect to be punished.
4. I feel I am being punished.

7. Self-Dislike

1. I feel the same about myself as ever.
2. I have lost confidence in myself.
3. I am disappointed in myself.
4. I dislike myself.

8. Self-Criticalness

1. I don't criticize or blame myself more than usual.
2. I am more critical of myself than I used to be.
3. I criticize myself for all of my faults.
4. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes

1. I don't have any thoughts of killing myself.
2. I have thoughts of killing myself, but I would not carry them out.
3. I would like to kill myself.
4. I would kill myself if I had the chance.

10. Crying

1. I don't cry anymore than I used to.
2. I cry more than I used to.
3. I cry over every little thing.
4. I feel like crying, but I can't.

11. Agitation

1. I am no more restless or wound up than usual.
2. I feel more restless or wound up than usual.
3. I am so restless or agitated that it's hard to stay still.
4. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest

1. I have not lost interest in other people or activities.
2. I am less interested in other people or things than before.
3. I have lost most of my interest in other people or things.
4. It's hard to get interested in anything.

13. Indecisiveness

1. I make decisions about as well as ever.
2. I find it more difficult to make decisions than usual.
3. I have much greater difficulty in making decisions than I used to.
4. I have trouble making any decisions.

14. Worthlessness

1. I do not feel I am worthless.
2. I don't consider myself as worthwhile and useful as I used to.
3. I feel more worthless as compared to other people.
4. I feel utterly worthless.

15. Loss of Energy

1. I have as much energy as ever.
2. I have less energy than I used to have.
3. I don't have enough energy to do very much.
4. I don't have enough energy to do anything.

16. Changes in Sleeping Pattern

0. I have not experienced any change in my sleeping pattern.

1a. I sleep somewhat more than usual.

1b. I sleep somewhat less than usual.

2a. I sleep a lot more than usual.

2b. I sleep a Iot less than usual.

3a. I sleep most of the day.

3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability

1. I am no more irritable than usual.
2. I am more irritable than usual.
3. I am much more irritable than usual.
4. I am irritable all the time.

18. Changes in Appetite

0. I have not experienced any change in my appetite.

1a. My appetite is somewhat less than usual.

1b. My appetite is somewhat greater than usual.

2a. My appetite is much less than before.

2b. My appetite is much greater than usual.

3a. I have no appetite at all.

3b. I crave food all the time.

19. Concentration Difficulty

1. I can concentrate as well as ever.
2. I can't concentrate as well as usual.
3. It's hard to keep my mind on anything for very long.
4. I find I can't concentrate on anything.

20. Tiredness or Fatigue

1. I am no more tired or fatigued than usual.
2. I get more tired or fatigued more easily than usual.
3. I am too tired or fatigued to do a lot of the things I used to do.
4. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex

1. I have not noticed any recent change in my interest in sex.
2. I am less interested in sex than I used to be.
3. I am much less interested in sex now.
4. I have lost interest in sex completely.

Relación de Acrónimos

API Application Programming Interface. 14

BSD Berkeley Software Distribution. 14

EPL Eclipse Public License. 15

eRisk Early Risk Prediction on the Internet. 3

GPL General Public License. 14

HTML HyperText Markup Language. 14

IDE Integrated Development Environment. 15

IR Information Retrieval. 6, 7, 11, 23, 50

JVM Java Virtual Machine. 13, 14

LM Language Model. 7, 10, 68

MEDMM Maximum Entropy Divergence Minimization Model. 68

MIT Massachusetts Institute of Technology. 15, 16

MLE Maximum Likelihood Estimate. 8

OMS Organización Mundial da Saúde. 1

POM Project Object Model. 16

PRAW Python Reddit API Wrapper. 14

PRP Probability Ranking Principle. 7

PyPI Python Package Index. 15

RM Relevance-Based Language Models. 10, 11, 21–25, 45–49, 51, 55, 62, 64, 65, 67, 68

RM1 Relevance Model 1. 10, 11

RM3 Relevance Model 3. 11

SCM Source Code Management. 17

WWW World Wide Web. 6

XML eXtensible Markup Language. 14, 16

Glosario

ad hoc Solución elaborada especificamente para un problema ou fin preciso e, por tanto, non xeneralizable nin utilizable para outros propósitos. 7

baseline Valor coñecido ou inicial a partires do cal se poden comparar valores posteriores do que se está a medir. 3, 49, 60, 65, 67

branching No contexto do control de versións, replicación de obxectos para que se poida traballar sobre estes de forma separada e en paralelo. 17

bytecode Código independente da máquina que xeran compiladores de determinadas linguaxes (Java, Erlang,...) e que é executado polo correspondente intérprete. 13

kernel Parte fundamental e básica de calquera sistema operativo. 17

lexicon Serie ordenada de palabras dunha lingua, persoa, rexión, materia ou época determinada. v, 3, 10, 19–21, 25–27, 30–32, 47, 49, 55, 62, 65, 67

macro Serie de instrucións que se almacenan para poderen ser executadas de maneira secuencial mediante unha única chamada ou orde de execución. 15

plugin Aplicacións que permite extender as funcionalidades doutra aplicación ou programa sen modificar o seu código. Pode verse coma un complemento. 15

scrapping Proceso que polo cal se transforman datos sen estructura da web en datos estructurados que poden ser almacenados e analizados nunha base de datos central ou nalgunha outra fonte de almacenamento. 14

software Sistema formal dun sistema informático, que comprende o conxunto das compoñentes lóxicas necesarias para a realización de tarefas específicas, tipicamente en contraposición ás componentes físicas. v, 4, 14–16, 37, 42–44, 51–53

toolkit Conxunto de ferramentas que tradicionalmente se distribúen xuntas dado que gardan algunha relación entre elas. 14

Bibliografía

[1] H. R. Saloni Dattani and M. Roser, "Mental health," *Our World in Data*, 2021, https://ourworldindata.org/mental-health.

[2] Global Burden of Disease Collaborative Network, "Global burden of disease study 2019 (gbd 2019) reference life table," 2021. [En liña]. Dispoñible en: http://ghdx.healthdata.org/record/ihme-data/global-burden-disease-study-2019-gbd-2019-reference-life-table

[3] S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. T. Chiu, S. Florescu, G. de Girolamo, O. Gureje, J. M. Haro, Y. He, C. Hu, E. G. Karam, N. Kawakami, S. Lee, C. Lund, V. Kovess-Masfety, D. Levinson, F. Navarro-Mateu, B. E. Pennell, N. A. Sampson, K. M. Scott, H. Tachimori, M. ten Have, M. C. Viana, D. R. Williams, B. J. Wojtyniak, Z. Zarkov, R. C. Kessler, S. Chatterji, and G. Thornicroft, "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO world mental health (WMH) surveys," *Psychological Medicine*, vol. 48, no. 9, pp. 1560–1571, Nov. 2017. [En liña]. Dispoñible en: https://doi.org/10.1017/s0033291717003336

[4] F. Charlson, M. van Ommeren, A. Flaxman, J. Cornett, H. Whiteford, and S. Saxena, "New WHO prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis," *The Lancet*, vol. 394, no. 10194, pp. 240–248, Jul. 2019. [En liña]. Dispoñible en: https://doi.org/10.1016/s0140-6736(19)30934-1

[5] J. Bueno-Notivol, P. Gracia-García, B. Olaya, I. Lasheras, R. López-Antón, and J. Santabárbara, "Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies," *International Journal of Clinical and Health Psychology*, vol. 21, no. 1, p. 100196, Jan. 2021. [En liña]. Dispoñible en: https://doi.org/10.1016/j.ijchp.2020.07.007

[6] D. Talevi, V. Socci, M. Carai, G. Carnaghi, S. Faleri, E. Trebbi, A. di Bernardo, F. Capelli, and F. Pacitti, "Mental health outcomes of the covid-19 pandemic,"

Rivista di Psichiatria, no. 2020May-June, May 2020. [En liña]. Dispoñible en: https://doi.org/10.1708/3382.33569

[7] R. Mojtabai, M. Olfson, and B. Han, "National trends in the prevalence and treatment of depression in adolescents and young adults," *Pediatrics*, vol. 138, no. 6, Dec. 2016. [En liña]. Dispoñible en: https://doi.org/10.1542/peds.2016-1878

[8] A. Barak and J. M. Grohol, "Current and future trends in internet-supported mental health interventions," *Journal of Technology in Human Services*, vol. 29, no. 3, pp. 155–196, Jul. 2011. [En liña]. Dispoñible en: https://doi.org/10.1080/15228835.2011.616939

[9] A. Halfin, "Depression: The benefits of early and appropriate treatment," *The American journal of managed care*, vol. 13, pp. S92–7, 12 2007.

[10] A. Picardi, I. Lega, L. Tarsitani, M. Caredda, G. Matteucci, M. Zerella, R. Miglio, A. Gigantesco, M. Cerbo, A. Gaddini, F. Spandonaro, M. Biondi, and T. S.-D. Group, "A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care," *Journal of Affective Disorders*, vol. 198, pp. 96–101, Jul. 2016. [En liña]. Dispoñible en: https://doi.org/10.1016/j.jad.2016.03.025

[11] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," 2015. [En liña]. Dispoñible en: http://hdl.handle.net/2152/31333

[12] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, vol. 54, no. 1, pp. 547–577, Feb. 2003. [En liña]. Dispoñible en: https://doi.org/10.1146/annurev.psych.54.101601.145041

[13] R. A. CALVO, D. N. MILNE, M. S. HUSSAIN, and H. CHRISTENSEN, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, Jan. 2017. [En liña]. Dispoñible en: https://doi.org/10.1017/s1351324916000383

[14] N. Colineau and C. Paris, "Talking about your health to strangers: understanding the use of online social networks by patients," *New Review of Hypermedia and Multimedia*, vol. 16, no. 1-2, pp. 141–160, Apr. 2010. [En liña]. Dispoñible en: https://doi.org/10.1080/13614568.2010.496131

[15] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, 2nd ed. USA: Addison-Wesley Publishing Company, 2011.

[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* USA: Cambridge University Press, 2008.

[17] F. Lancaster and E. Fayen, *Information Retrieval: On-line*, ser. A Wiley-Becker & Hayes series book. Melville Publishing Company, 1973.

[18] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975. [En liña]. Dispoñible en: https://doi.org/10.1145/361219.361220

[19] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, no. 3, pp. 129–146, May 1976. [En liña]. Dispoñible en: https://doi.org/10.1002/asi.4630270302

[20] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98.* ACM Press, 1998. [En liña]. Dispoñible en: https://doi.org/10.1145/290941.291008

[21] C. Zhai, "Statistical language models for information retrieval a critical review," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 3, pp. 137–213, 2007. [En liña]. Dispoñible en: https://doi.org/10.1561/1500000008

[22] S. Robertson, "The Probability Ranking Principle in IR," *Journal of Documentation*, vol. 33, no. 4, pp. 294–304, Apr. 1977. [En liña]. Dispoñible en: https://doi.org/10.1108/eb026647

[23] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 179–214, Apr. 2004. [En liña]. Dispoñible en: https://doi.org/10.1145/984321.984322

[24] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01.* ACM Press, 2001. [En liña]. Dispoñible en: https://doi.org/10.1145/383952.383972

[25] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. D. Diaz, L. S. Larkey, X. Li, D. Metzler, M. D. Smucker, T. Strohman, H. R. Turtle, and C. Wade, "Umass at trec 2004: Notebook," 2004.

[26] "index | tiobe - the software quality company," 2022. [En liña]. Dispoñible en: https://www.tiobe.com/tiobe-index/

[27] "Stack overflow developer survey 2021," 2021, consultado o 2022-06-27. [En liña]. Dispoñible en: https://insights.stackoverflow.com/survey/2021#section-most-popular-technologies-integrated-development-environment

[28] D. E. Losada, F. Crestani, and J. Parapar, "eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations," in *Lecture Notes in Computer Science*. Springer International Publishing, 2017, pp. 346–360. [En liña]. Dispoñible en: https://doi.org/10.1007/978-3-319-65813-1_30

[29] ——, "Overview of eRisk: Early risk prediction on the internet," in *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 343–361. [En liña]. Dispoñible en: https://doi.org/10.1007/978-3-319-98932-7_30

[30] ——, "Overview of eRisk 2019 early risk prediction on the internet," in *Lecture Notes in Computer Science*. Springer International Publishing, 2019, pp. 340–357. [En liña]. Dispoñible en: https://doi.org/10.1007/978-3-030-28577-7_27

[31] ——, "Overview of eRisk 2020: Early risk prediction on the internet," in *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 272–287. [En liña]. Dispoñible en: https://doi.org/10.1007/978-3-030-58219-7_20

[32] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of eRisk 2021: Early risk prediction on the internet," in *Lecture Notes in Computer Science*. Springer International Publishing, 2021, pp. 324–344. [En liña]. Dispoñible en: https://doi.org/10.1007/978-3-030-85251-1_22

[33] A. T. Beck, R. A. Steer, and G. Brown, "Beck depression inventory–II," 1996. [En liña]. Dispoñible en: https://doi.org/10.1037/t00742-000

[34] D. Losada and P. Gamallo, "Evaluating and improving lexical resources for detecting signs of depression in text," *Language Resources and Evaluation*, vol. 54, pp. 1–24, 03 2020.

[35] Y. Neuman, Y. Cohen, D. Assaf, and G. Kedma, "Proactive screening for depression through metaphorical and automatic text analysis," *Artificial Intelligence in Medicine*, vol. 56, no. 1, pp. 19–25, Sep. 2012. [En liña]. Dispoñible en: https://doi.org/10.1016/j.artmed.2012.06.001

[36] M. Gamon, M. Choudhury, S. Counts, and E. Horvitz, "Predicting depression via social media," in *ICWSM*, 07 2013.

[37] E. Zhang and Y. Zhang, "Average precision," in *Encyclopedia of Database Systems*. Springer US, 2009, pp. 192–193. [En liña]. Dispoñible en: https://doi.org/10.1007/978-0-387-39940-9_482

[38] R. M. Ortega-Mendoza, D. I. Hernández-Farías, M. M. y Gómez, and L. Villaseñor-Pineda, "Revealing traces of depression through personal statements analysis in social media," *Artificial Intelligence in Medicine*, vol. 123, p. 102202, Jan. 2022. [En liña]. Dispoñible en: https://doi.org/10.1016/j.artmed.2021.102202

[39] J. Sutherland and K. Schwaber, "Scrum guide," consultado o 2022-06-27. [En liña]. Dispoñible en: https://scrumguides.org/scrum-guide.html

[40] "Guia salarial sector tic galicia 2015-2016," consultado o 2022-06-27. [En liña]. Dispoñible en: https://www.scribd.com/document/288511179/Guia-Salarial-Sector-TI-Galicia-2015-2016

[41] Y. Lv and C. Zhai, "Revisiting the divergence minimization feedback model," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, Nov. 2014. [En liña]. Dispoñible en: https://doi.org/10.1145/2661829.2661900

[42] S. Naseri, J. Dalton, A. Yates, and J. Allan, "Ceqe: Contextualized embeddings for query expansion," 2021. [En liña]. Dispoñible en: https://arxiv.org/abs/2103.05256

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [En liña]. Dispoñible en: https://aclanthology.org/N19-1423