Phelipe Gustavo da Silva

Coleta e classificação de publicações científicas relacionadas a depressão

Divinópolis - Brasil

17 de outubro de 2020

Phelipe Gustavo da Silva

Coleta e classificação de publicações científicas relacionadas a depressão

Monografia apresentada ao Curso de Engenharia de Computação da UEMG Unidade Divinópolis, como requisito parcial para obtenção do título de Bacharel em Engenharia da Computação, sob a orientação da Profª. Cristina Maria Valadares de Lima

Universidade do Estado de Minas Gerais - UEMG

Unidade Divinópolis

Curso de Engenharia da Computação

Orientador: Cristina Maria Valadares de Lima

Divinópolis - Brasil

17 de outubro de 2020

Phelipe Gustavo da Silva
Coleta e classificação de publicações científicas relacionadas a depressão/ Phelipe Gustavo da Silva. – Divinópolis - Brasil, 17 de outubro de 2020-
31 p.

Orientador: Cristina Maria Valadares de Lima

Monografia – Universidade do Estado de Minas Gerais - UEMG
Unidade Divinópolis
Curso de Engenharia da Computação, 17 de outubro de 2020.

1. PubMed Central; 2. Mineração de dados; 3. Depressão.

Phelipe Gustavo da Silva

Coleta e classificação de publicações científicas relacionadas a depressão

Monografia apresentada ao Curso de Engenharia de Computação da UEMG Unidade Divinópolis, como requisito parcial para obtenção do título de Bacharel em Engenharia da Computação, sob a orientação da Profª. Cristina Maria Valadares de Lima

---

Cristina Maria Valadares de Lima
Orientador

Professor
Convidado 1

Professor
Convidado 2

Divinópolis - Brasil
17 de outubro de 2020

Dedico este trabalho a Deus, que sempre foi o meio de maior apoio nos momentos mais difíceis.

Agradecimentos

Os agradecimentos principais são direcionados à Deus que me deu oportunidade e força para superar os desafios e dificuldades.

À minha família especialmente meus pais que são minha base, por todo apoio e incentivo aos meus estudos.

Aos meus amigos, por auxiliar no desenvolvimento das atividades acadêmicas.

E por fim, à todos que direta ou indiretamente fizeram parte da minha formação.

"Não desista nas primeiras tentativas,
a persistência é amiga da conquista.
Se você quer chegar aonde a maioria não chega,
faça o que a maioria não faz."
(Bill Gates)

Sumário

Lista de ilustrações

Lista de abreviaturas e siglas

API	Interface de Programação de Aplicações
JSON	Notação de Objetos JavaScript
KDD	Descoberta de Conhecimento em Base de Dados

Resumo

O aumento de dados digitais amplificou o processo de busca por conhecimento, muitas pesquisas acadêmicas são publicadas em repositórios que são acessíveis na *web*, compondo uma extensa base de dados, um exemplo disso são os artigos relacionados a pesquisas sobre o tema depressão, que tende a aumentar ainda mais com a crescente quantidade de casos diagnosticados. O entendimento da doença é fundamental para o diagnóstico precoce e tratamento dos enfermos. Visando contribuir na descoberta de conhecimento das publicações científicas, serão aplicadas técnicas de descoberta de conhecimento em banco de dados para classificação das publicações por país, tornando possível a identificação dos países que mais contribuem com pesquisas relacionadas a doença e os termos mais pesquisados. Para isso, será utilizada a linguagem *Python* com a biblioteca *Entrez* e uma API em Node.JS conectada a um banco de dados MongoDB. Além disso, será desenvolvido um sistema *web* para exibição dos dados coletados.

Palavras-chaves: PubMed Central, mineração de dados, depressão.

Abstract

The increase in digital data has amplified the process of searching for knowledge, many academic researches are published in repositories that are accessible on the web, composing an extensive database, an example of which are the articles related to research of depression, which tends to further increase with the increasing number of diagnosed cases. Understanding the disease is essential for the early diagnosis and treatment of the sick. In order to contribute to the discovery of knowledge of scientific publications, knowledge discovery techniques will be applied in a database to classify publications by country, making it possible to identify the countries that contribute the most with research related to the disease and the most searched terms. For this, the Python language will be used with the Entrez library and an API in Node.JS connected to a MongoDB database. In addition, a web system will be developed to display the collected data.

Key-words: PubMed Central, data mining, depression.

1 Introdução

O crescente aumento de dispositivos digitais como computadores, *smartphones*, *tablets*, entre outros, ocasionou a expansão dos dados digitais. Em 2012, cerca de 2,5 *exabytes* [1] de dados eram criados todos os dias, esse número é dobrado aproximadamente a cada 40 meses. Mais dados são trafegados na *Internet* a cada segundo do que toda a quantidade de dados armazenados em toda a *Internet* há 20 anos (LOHR, 2012).

Em um relatório anual da *internet* elaborado pela Cisco, uma grande companhia de sistemas de redes, é apresentado que no ano de 2017 cerca de 122,000 *petabytes* [2] de dados são trafegados por mês no mundo (Cisco, 2020).

Entre os dados armazenados estão artigos científicos, que são publicados em repositórios acessíveis pela *web*, dentre esses artigos incluem trabalhos acadêmicos relacionados a depressão, uma doença comum em todo o mundo, considerada pela Organização Mundial da Saúde (OMS) como o "mal do século", possui mais de 300 milhões de pessoas afetadas (World Health Organization, 2018).

Segundo (World Health Organization, 2018) a depressão é um distúrbio mental caracterizado pela tristeza persistente e a perda de interesse nas atividades que a pessoa diagnosticada normalmente gosta, acompanhada de uma incapacidade de realizar atividades diárias por pelo menos duas semanas. Pessoas diagnosticadas, normalmente apresentam fatores como perda de energia, mudança de apetite, ansiedade, indecisão e pensamentos de automutilação ou suicídio.

O desenvolvimento de pesquisas científicas relacionadas a doença originou uma grande quantidade de publicações. Em uma busca realizada na base de dados da Pubmed Central através do termo "*depression*" foram encontrados cerca de 500.000 artigos. Nessas condições, a procura por informação se torna uma tarefa extensa e muitas vezes ineficiente.

O estudo da doença é fundamental para melhor identificação de casos e tratamentos. A classificação das publicações por país, e a disponibilidade destes dados categorizados, através da aplicação de técnicas de extração de conhecimento em base de dados é uma das formas de contribuir para para obtenção de informações relevantes para pesquisas futuras.

Diante deste cenário, propõe-se coletar e organizar publicações relacionadas a depressão através de técnicas de descoberta de conhecimento em base de dados (KDD - *Knowledge Data Discovery*), utilizando a linguagem de programação Python e uma API (Interface de Programação de Aplicações) em Node.JS, responsável pela inserção e acesso dos dados em um banco MongoDB. Para disponibilização das publicações e suas

[1] *Exabyte* ou EB, unidade de medida de informação 1 EB = $1 * 10^{18}$ *Bytes*
[2] *Pentabyte* ou PB, unidade de medida de informação 1 PB = $1 * 10^{15}$ *Bytes*

respectivas estatísticas será desenvolvido um sistema *web*.

1.1 Objetivo Geral

Contribuir para a análise e classificação dos países dos autores de artigos científicos relacionados a depressão, através de técnicas de obtenção de conhecimento em base de dados e mapeamento de estudos.

1.2 Objetivos Específicos

- Coleta de publicações científicas da base de dados PubMed Central;
- Implementação de técnicas para descoberta de conhecimento;
- Identificação dos países dos autores;
- Desenvolvimento do sistema *web* para visualização das informações obtidas.

1.3 Justificativa

Através de técnicas de descoberta de conhecimento em base de dados descritas por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) e da linguagem Python em conjunto com uma API em Node.JS, o presente trabalho se justifica por contribuir na organização e obtenção de conhecimento de publicações relacionadas à depressão, tendo em vista o crescente aumento de pessoas diagnosticadas, cerca de 18% entre os anos de 2005 e 2015 (World Health Organization, 2018), identificando os países que estão publicando maior quantidade de artigos.

1.4 Estrutura da Monografia

A presente monografia está dividida em cinco capítulos. No capítulo 2 são apresentados alguns trabalhos relacionados. No capítulo 3 são evidenciadas as ferramentas e conceitos necessárias para realização desse trabalho. No capítulo 4 são apresentadas a metodologia e o desenvolvimento do trabalho. No capítulo 5 são expostos os resultados obtidos. No capítulo 6 são apresentadas as considerações finais.

2 Trabalhos Relacionados

Este capítulo apresenta uma revisão bibliográfica conduzida no presente trabalho. Será apresentado uma breve descrição de alguns trabalhos relacionados aos temas utilizados nesse projeto, tais como, mineração de dados, depressão e também os resultados obtidos por cada um desses autores.

2.1 Coleta, Integração e Caracterização de base de dados de câncer

O trabalho de (FARIA, 2014) foi desenvolvido com o objetivo de construir uma base de conhecimento sobre o câncer com a utilização de técnicas de mineração de dados e uma ferramenta *web* e móvel para absorver os resultados obtidos por meio de mapas, gráficos e tabelas.

O objetivo geral de definir qual país pesquisa mais sobre determinado tipo de neoplasia foi atingido. Após uma média de 6 execuções do algoritmo, foram aproveitados 755.030 artigos, cerca de de 72% de toda a base coletada.

Este trabalho se difere em relação ao tema no qual os dados serão coletados, o presente trabalho contribui com uma base de conhecimento sobre depressão. Apesar de utilizar a mesma metodologia (KDD), as ferramentas utilizadas se diferem das que foram utilizas no presente trabalho, assim como no trabalho de (FARIA, 2014), o presente trabalho utilizará a linguagem Python em conjunto com o banco de dados MongoDB, será utilizado uma API em Node.JS, tornando possível escalar o sistema, resultando em uma coleta de dados mais rápida.

2.2 As melhores cidades para pesquisa em psicologia no mundo

O trabalho de (BORNMANN; LEYDESDORFF; KRAMPEN, 2012) tem o objetivo de encontrar os melhores centros de excelência em psicologia através de técnicas da cientometria[1]. Foram observadas 214 cidades, com uma produção de pelo menos 50 artigos no ano de 2007. Foi desenvolvida uma interface utilizando *overlays* do Google Maps para exibição da concentração de artigos publicados.

Os resultados obtidos foram satisfatórios, sendo possível visualizar as melhores cidades para pesquisa em psicologia no mundo, além disso, é possível visualizar a relação dos resultados esperados e obtidos de cada cidade.

[1] Ciência da informação que procura estudar aspectos quantitativos da ciência e da produção científica

O que difere o trabalho de (BORNMANN; LEYDESDORFF; KRAMPEN, 2012) do presente trabalho é principalmente o objetivo, o presente trabalho pretende organizar os dados com o objetivo de identificar os países dos autores, já no trabalho de (BORNMANN; LEYDESDORFF; KRAMPEN, 2012) os dados são organizados como objetivo analisar a qualidade dos artigos.

2.3 Contribuições

Os itens abordados nesse capítulo, contribuíram diretamente no desenvolvimento do presente trabalho. (FARIA, 2014) expôs suas dificuldades e soluções durante a identificação de alguns campos nos artigos retornados pelo repositório PubMed. (BORNMANN; LEYDESDORFF; KRAMPEN, 2012) apresentou os dados de forma clara e intuitiva ao usuário utilizando a Api JavaScript do Google Maps, motivando a adoção dessa biblioteca no presente trabalho.

3 Fundamentação Teórica

Este capítulo descreve as ferramentas e conceitos que serão abordados nesse projeto. Serão apresentadas definições básicas sobre a depressão, técnicas para descoberta de conhecimento, banco de dados utilizado e ferramentas auxiliares.

3.1 Depressão

A depressão tem se tornando uma das doenças mais diagnosticadas na atualidade. Dados da Organização Mundial da Saúde (World Health Organization, 2018) indicam que mais de 300 milhões de pessoas sofrem com o transtorno, nos piores desses casos, cerca de 800 mil por ano, chegam ao suicídio, sendo a principal causa de morte entre pessoas com idade entre 15 e 29 anos.

Segundo (PORTO, 1999) o termo depressão pode ser empregado para designar tanto um estado afetivo (a tristeza), quanto um sintoma, uma síndrome e uma (ou várias) doença(s). Enquanto sintoma pode surgir em vários quadros clínicos como transtorno de estresse pós-traumático, demência, esquizofrenia, alcoolismo, doenças clinicas entre outros. Como síndrome a depressão inclui não apenas alterações de humor como tristeza, apatia e irritabilidade, mas também uma gama de outros aspectos, incluindo alterações cognitivas, psicomotoras e vegetativas. Por fim, enquanto doença a depressão tem sido classificada de várias formas. Entre os quadros mencionados na literatura atual encontram-se: transtorno depressivo maior, melancolia, distimia, depressão integrante do transtorno bipolar tipos I e II, depressão como parte da ciclotimia entre outros.

De acordo com (AMORIN, 2014) existe um aumento significativo de pessoas diagnosticadas com depressão, e um número maior de pessoas que estão convivendo com a doença e não buscam atendimento adequado, por compreenderem que não necessitam de ajuda, tornando ainda mais longo e penoso o tempo de espera para um diagnóstico preciso e, logo, o seu prognóstico em que há possibilidade de paciente e médico/psicólogo juntos buscarem uma solução para o problema detectado. Isso acarreta consequências graves, como por exemplo, o suicídio.

A depressão descreve um estado de humor. Segundo (FAVA; KENDLER, 2000) esse estado é transitório sendo vivenciado por praticamente todos os indivíduos em algum momento de sua vida, bem como uma síndrome clínica ou bio-comportamental, geralmente chamada de Transtorno Depressivo Maior (MDD).

Segundo (The National Institute of Mental Health, 2018) existem várias formas de depressão e essas são ligeiramente diferentes podendo se desenvolver em algumas

circunstâncias que serão descritas a seguir.

3.1.1 Transtorno depressivo persistente

Também chamado de distimia, o transtorno depressivo persistente é um humor deprimido caracterizado por apresentar sintomas por pelo menos dois anos. Conforme (The National Institute of Mental Health, 2018) uma pessoa diagnosticada com transtorno depressivo persistente pode ter episódios de depressão juntamente com períodos de sintomas menos graves.

3.1.2 Depressão pós-parto

Algumas mulheres podem apresentar depressão durante a gravidez ou após o parto. Segundo (The National Institute of Mental Health, 2018) a depressão pós-parto é muito mais grave do que os *"baby blues"* (sintomas depressivos e de ansiedade relativamente leves que geralmente desaparecem duas semanas após o parto) que muitas mulheres experimentam. Os sentimentos de extrema tristeza, ansiedade e exaustão que acompanham a depressão pós-parto podem dificultar para essas novas mães a concluir as atividades diárias de cuidado para si e/ou para seus bebês.

3.1.3 Depressão psicótica

Conforme (The National Institute of Mental Health, 2018) a depressão psicótica ocorre quando uma pessoa tem depressão severa com alguma forma de psicose, como ter falsas crenças fixas (ilusões) perturbadoras ou ouvir ou ver coisas perturbadoras que outras pessoas não podem ouvir ou ver (alucinações). Os sintomas psicóticos geralmente têm um "tema"depressivo, como delírios de culpa, pobreza ou doença.

3.1.4 Transtorno afetivo sazonal

O transtorno afetivo sazonal é caracterizado pelo aparecimento de depressão durante os meses de inverno, quando há menos luz solar natural. Conforme (The National Institute of Mental Health, 2018) o transtorno é tipicamente acompanhado de retraimento social, aumento do sono e ganho de peso e retorna previsivelmente a cada ano. A fototerapia (terapia de luz) é um dos tratamentos possíveis para o transtorno. De acordo com (PARTONEN; LÖNNQVIST, 1998) o tratamento com a luz matinal resulta na recuperação em cerca de dois terços dos pacientes com sintomas leves, porém em pacientes em situações mais graves a recuperação ocorre em menos da metade dos casos, esse tratamento pode apresentar resultados em até duas semanas após o início, sendo recomendado a continuidade durante o inverno para evitar possíveis reincidência.

3.1.5 Transtorno bipolar

O transtorno bipolar é diferente da depressão, porem pessoas com transtorno bipolar experimentam episódios de humor extremamente baixo que atendem aos critérios para depressão (denominada "depressão bipolar"). De acordo com (The National Institute of Mental Health, 2018) uma pessoa com transtorno bipolar também experimenta um humor extremamente alto - eufórico ou irritável - chamado "mania" ou uma forma menos grave chamada "hipomania".

Neste trabalho o tema descrito nessa seção será utilizado para coleta de dados dos artigos, para isso, precisa-se entender sobre o processo de descoberta de conhecimento em base de dados, descrito a seguir.

3.2 Descoberta de conhecimento em base de dados (KDD)

Desde o início da era da informação as diversas áreas do conhecimento vem armazenando um grande volume de dados, sendo necessário um tratamento para obtenção de informações significantes.

O método tradicional de análise de dados é feito de forma manual em que especialistas observam novos dados ou tópicos da área de conhecimento e realizam relatórios, que são distribuídos para comunidade ou órgãos superiores. Esse tipo de análise é lento, caro e muito subjetivo.

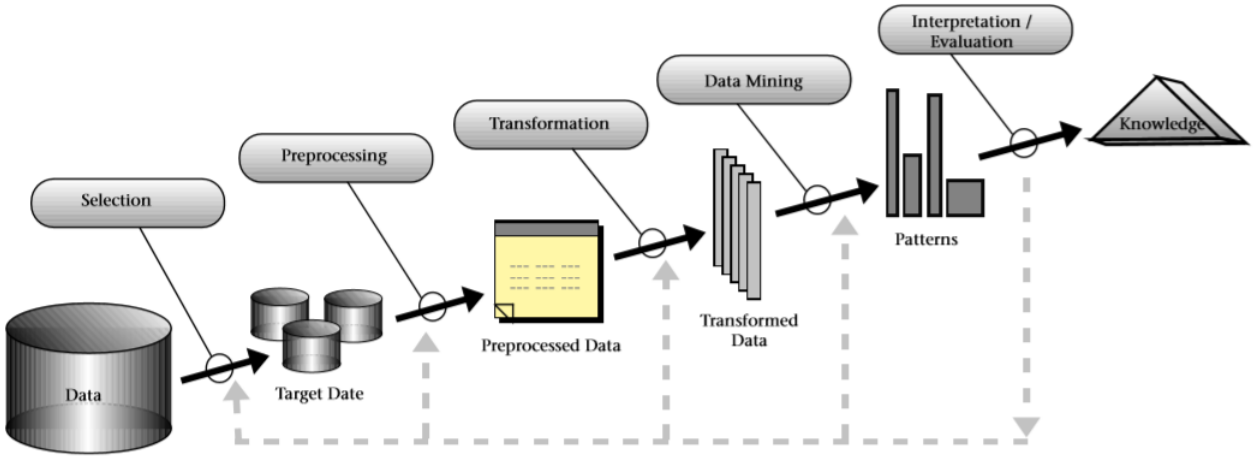
Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) são necessário novas gerações de teorias e ferramentas computacionais para ajudar os humanos a extrair informações úteis (conhecimento) dos volumes crescentes de dados digitais, sendo necessária uma automação minimamente parcial.

O termo descoberta de conhecimento em banco de dados (KDD) foi apresentado em um *workshop* em 1989 para enfatizar que o conhecimento é o produto final de uma descoberta de dados. Sobretudo, a descoberta de conhecimento em base de dados (KDD) lida com um problema específico: sobrecarga de dados, mapeando dados de baixo nível em dados mais compactos, abstratos e úteis. De acordo com (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) o processo KDD pode ser dividido em algumas etapas descritas a seguir e ilustradas na figura 1.

Primeiramente deve-se desenvolver um entendimento da aplicação e os conhecimentos prévios importantes, permitindo a visualização do problema pelo lado do cliente, a fim de criar um conjunto de dados destino, focando em uma determinada variável ou amostra, no qual a descoberta deve ser executada.

Após a obtenção dos dados deve ser realizado a limpeza e pré-processamento, são coletadas as informações necessárias para modelar ou contabilizar os ruídos. De acordo

Figura 1 – Visão geral das etapas que compõem o KDD



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

com (DANTAS et al., 2008) a fase de *Data Cleaning* e Pré-Processamento tem por objetivo assegurar a qualidade dos dados envolvidos no KDD realizando operações básicas como a remoção de ruídos, que podem ser, por exemplo, atributos nulos.

A seguir, deve ser realizado a redução de dados e projeção, procurando recursos úteis para representar os dados de acordo com a necessidade do problema. É importante definir as etapas da mineração de dados, como por exemplo descrição, classificação, regressão, predição, agrupamento e associação.

Posteriormente é realizada a análise e seleção de modelos e hipóteses, escolhendo um modelo de mineração de dados e método a ser utilizado na busca de padrões. A mineração de dados é iniciada, procurando padrões em uma forma representacional tais como regras, árvores de classificação, regressão e agrupamento. Conforme (DANTAS et al., 2008) esta fase é a mais importante do KDD, sendo definido o algoritmo mais compatível com o objetivo da extração, a fim de encontrar padrões nos dados que sirva de subsídios para descobrir conhecimentos ocultos.

Por fim, deve ser realizada uma análise em busca de interpretar padrões dos dados minerados para possíveis ajustes nas etapas anteriores e novas iterações. Assim, é obtido o conhecimento, podendo ser utilizado em outro sistema para tomadas de decisões, ou simplesmente documentá-lo e relatá-lo as partes interessadas. Esse processo inclui a verificação e resolução de conflitos em potencial com o conhecimento anteriormente extraído.

Uma das etapas mais importantes desse processo e que será utilizada neste trabalho é a mineração de dados, sendo melhor descrita a seguir.

3.2.1 Mineração de dados

A mineração de dados é comumente utilizada para referenciar a descoberta de conhecimento, entretanto como citado anteriormente, a mineração de dados é uma das etapas que compõe o processo de KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Segundo (LAROSE; LAROSE, 2014) a mineração de dados é um o processo de descoberta de padrões e tendências úteis em grandes conjuntos de dados.

Em uma definição na perspectiva estatística, (HAND; SMYTH; MANNILA, 2001) define a mineração de dados como a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados.

Conforme (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) a mineração de dados envolve algumas etapas. A primeira delas é a descrição que busca identificar e descrever um modelo bem definido a partir de um grupo de dados. Outro etapa é classificação de acordo com (WEISS S. I., 1991) consiste em definir uma função que mapeia (classifica) um item de dados em uma das várias classes pré-definidas.

Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) a etapa de regressão busca identificar uma função que mapeia um item de dados para uma variável de previsão com valor real. A partir da definição das funções torna-se possível agrupar os dados. De acordo com (JAIN; DUBES, 1988) esse etapa de agrupamento, é uma tarefa descritiva que busca identificar um limite finito de categorias ou grupos que descrevem o dado.

A próxima etapa é a sumarização, conforme (AGRAWAL et al., 1996) consiste em encontrar uma descrição para um subconjunto de dados. Um exemplo simples seria tabular a média e os desvios padrão para todos os campos. Métodos mais sofisticados envolvem a derivação de regras de resumo. Por fim, após todas as etapas de organização dos dados é possível realizar a predição, buscando a partir dos dados previamente coletados, definir o valores de atributos em novas situações.

Ao fim dessas etapas é possível extrair informações que possam ter valor para o tema em questão, encontrando novas correlações, tendências e padrões. Neste projeto as informações serão extraídas a partir do repositório da PubMed Central que será descrito na próxima seção.

3.3 Repositório PubMed Central

O PubMed Central (PMC) é um repositório de artigos científicos completo e gratuito, desenvolvido e mantido pelo centro nacional de biotecnologia da informação (NCBI), possui cerca de cinco milhões de artigos armazenados relacionados a literatura

de biomédicos e de ciências da vida da Biblioteca Nacional de Medicina dos Institutos Nacionais de Saúde dos EUA (NIH / NLM) (NCBI, 2019).

Todo conteúdo mantido no repositório PMC é salvo em arquivos no formato *eXtensible Markup Language*(XML) seguindo o padrão NISO Z39.96-2015 JATS XML, que representa a estrutura e o significado de um documento de forma relativamente simples, facilmente legível (NCBI, 2019).

O repositório foi escolhido para o desenvolvimento do projeto devido a alta disponibilidade de artigos relacionados ao tema (depressão) além da facilidade de integração com a interface de programação Entrez (e-utilities).

3.3.1 Entrez e e-utilities

O Entrez é um sistema de busca que integra 38 bancos de dados incluindo o PMC (NCBI, 2006). O sistema é composto uma interface gráfica com várias opções de configuração de busca, como filtros e campos de busca, permitindo a construção de buscas precisas.

O Utilitário de programação do Entrez (e-utilities) é uma Interface de programação de aplicações (API) constituída de oito programas que proveem acesso a consultas no banco de dados da NCBI (NCBI, 2008).

Neste projeto, os utilitários de programação disponibilizados pela Entrez serão consumidos para a busca dos artigos do repositório PMC através de um *script* implementado com a linguagem de programação Python seguindo o paradigma de programação orientada a objetos.

3.4 Programação Orientada a Objetos

Programação Orientada a Objetos (POO), é um paradigma de programação que utiliza instâncias de classes (objetos) que determinam qual informação um objeto contém e como ele pode manipulá-la (RICARTE, 2001).

Um dos conceitos presentes na POO é a herança, que permite a extensão de definições já existentes. Além da herança o paradigma permite selecionar funcionalidades que um programa ira utilizar de forma dinâmica durante sua execução através do polimorfismo (RICARTE, 2001).

Segundo (RICARTE, 2001), classes são como gabaritos para a definição de objetos, através de uma definição de classe, descreve-se que propriedades ou atributos o objeto terá. Além da especificação de atributos, classes descrevem qual será o comportamento de objetos da classe, ou seja, quais funcionalidades podem ser aplicadas aos objetos da

classe. Um método é o equivalente a um procedimento ou função, com a restrição que ele manipula apenas suas variáveis locais e atributos definidos pela classe.

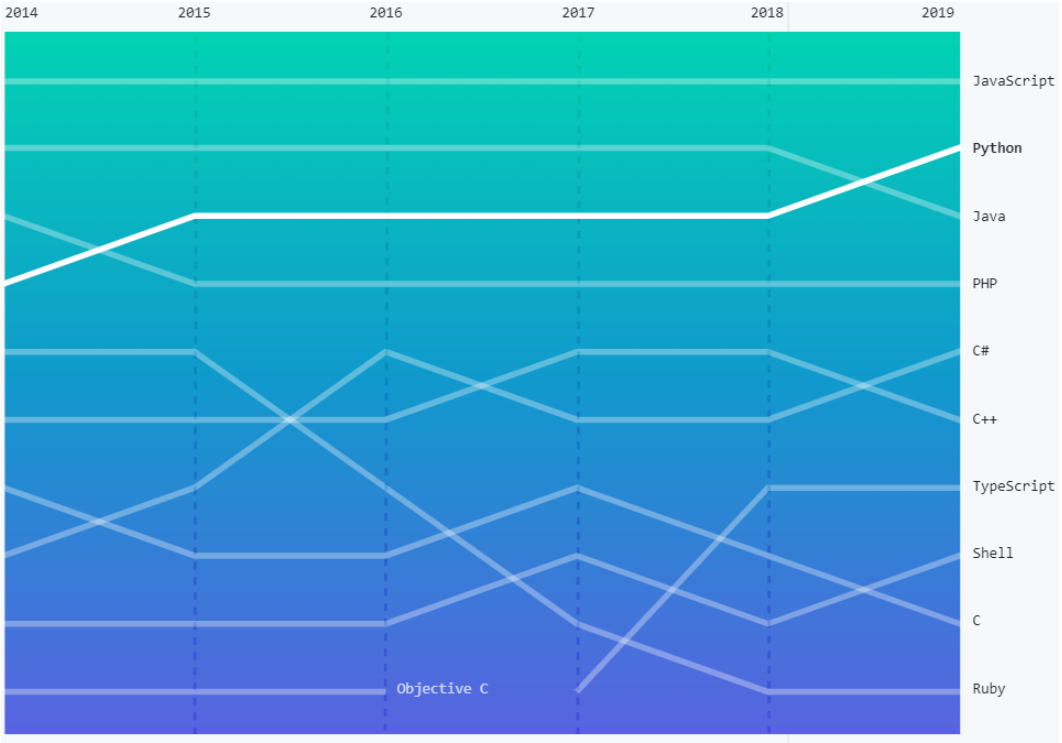
Neste projeto, o paradigma será utilizado em conjunto com a linguagem Python para definir as classes responsáveis pela mineração e manipulação dos artigos, nessas classes serão descritas os métodos que serão responsáveis pela coleta e tratamento dos dados.

3.5 Python

Criada em 1990 por Guido van Rossum na Holand, Python é uma linguagem de código aberto com licença GPL (*General Public License*), com uma sintaxe simples e grande poder de processamento de texto. Segundo (BORGES, 2014) Python é uma linguagem de altíssimo nível orientada a objetos, de tipagem dinâmica e forte, interpretada e interativa.

De acordo com (LAYTON, 2015) a alta popularidade da linguagem Python é devido à alta flexibilidade que ela dá ao programador, portando grande número de módulos que permitem executar diferentes tipos de tarefas, além disso, o código escrito em Python geralmente é mais legível e conciso do que em outra linguagem. Segundo (LAYTON, 2015) Python possui uma grande comunidade ativa de pesquisadores e iniciantes que utilizam a linguagem para mineração de dados.

Figura 2 – Principais Linguagens de programação



Fonte: (GitHub, 2019)

Em um relatório apresentado pelo GitHub, uma das principais plataformas de desenvolvimento, indicou um aumento do número de projetos utilizando majoritariamente linguagem Python, os dados apresentados na figura 2 incluem o período de 2014 a 2019 sendo possível observar que a linguagem ocupa em 2019 a segunda posição do *ranking*, liderado pelo JavaScript.

A linguagem Python possui uma biblioteca XML que será de grande importância para a manipulação dos artigos coletados, a seguir serão descritas as principais funcionalidades oferecidas pela biblioteca.

3.5.1 Biblioteca XML

A biblioteca XML provê alguns métodos que facilitam no tratamento de arquivos XML. O método parse da classe *xml.etree.ElementTree* retorna uma árvore de elementos *xml.etree.ElementTree* a partir de um arquivo XML.

A partir de um objeto *ElementTree* é possível realizar operações como:

- *find:* Realiza uma busca na árvore de elementos filhos;
- *findtext:* Busca o texto contido no primeiro elemento filho;
- *getroot:* Retorna o elemento pai (Python, 2019).

Essas operações auxiliam no processo de busca dos elementos de um arquivo XML.

3.6 Banco de dados

Segundo (DATE, 2004) banco de dados é um sistema computadorizado de manutenção de registros, um equivalente eletrônico de um armário de arquivamento, sendo possível realizar operações como inserção, busca, remoção, e alteração de dados.

Um banco de dados pode ser relacional onde os dados são armazenados em forma de tabelas, ou não relacional geralmente agrupados por uma coleção. Bancos de dados não relacionais oferecem melhor performance e maior escalabilidade, entretanto bancos relacionais possuem forte consistência de dados, seguindo as propriedades da ACID (Atomicidade, Consistência, Isolamento e Durabilidade).

Alguns bancos de dados possuem um conjunto de características que definem o termo NOSQL, algumas dessas características são essenciais para o desenvolvimento desse projeto e serão descritas a seguir.

3.6.1 NOSQL

Os Bancos de dados NOSQL surgiram da necessidade de escalar o armazenamento e processamento de grandes volumes de dados, o movimento surgiu em 2009 e cresceu rapidamente.

No início, grandes empresas enfrentando esse tipo de problema criaram suas próprias soluções, e publicaram alguns artigos científicos descrevendo diversas soluções ligadas ao gerenciamento de dados distribuído em larga escala, mas sem usar ainda o nome NOSQL (DECANDIA et al., 2007).

O termo NOSQL possui o significado "Não apenas SQL" e representa bancos de dados com características como: não relacional, distribuído, de código aberto e escalável horizontalmente, ausência de esquema ou esquema flexível, suporte a replicação nativo e acesso via APIs simples (NOSQL, 2019).

Um dos motivos que levaram a utilização desse tipo de banco no presente trabalho é a possibilidade de escalar, ou seja, a capacidade de acrescentar mais recursos de *hardware* viabilizando a manipulação de uma crescente quantidade de dados. Segundo (CHODOROW, 2013) o principal motivo para migrar de um banco relacional para uma banco relacional é a facilidade para escalar. Um exemplo de um sistema gerenciador de banco de dados NOSQL é o MongoDB que será abordado no próximo tópico.

3.6.2 MongoDB

MongoDB é um banco de dados de alta performance, escalável horizontalmente, distribuído e baseado em documentos JSON (Notação de Objetos JavaScript).

Alguns dos conceitos básicos do banco apresentados por (CHODOROW, 2013), envolvem o documento, que é a unidade básica de um dado do MongoDB e é similar a uma linha em um banco de dados relacional, porém mais expressivo, as coleções quem podem ser comparadas como uma tabela com um esquema dinâmico.

Segundo (CHODOROW, 2013) uma instância de MongoDB pode servir vários bancos de dados independentes, onde cada um pode ter suas próprias coleções. Além disso, todo documento possui um chave especial, *"_id"*, que é único em uma coleção.

MongoDB se destaca pela versatilidade e pela ampla comunidade de usuários ativos, sendo utilizado por milhões de desenvolvedores e empresas como Facebook, Google, Ebay e Adobe.

Outra característica do banco, é que ele possui um *shell* JavaScript, que pode ser utilizado para gerenciamento de instâncias de MongoDB e manipulação de dados, muito útil para o aprendizado, sendo possível executar as consultas e obter a respostas rapidamente.

Neste trabalho, a aplicação responsável pela manipulação dos dados no banco MongoDB será construída em um interpretador JavaScript, descrito no próximo item.

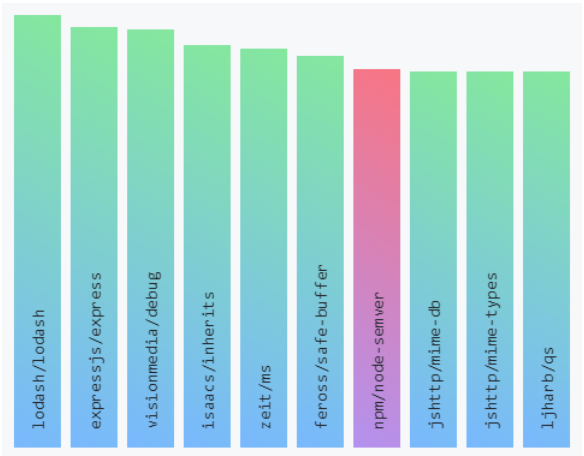
3.7 Node.js

Node.js é um interpretador de JavaScript de código aberto criado por Ryan Dahl em 2009 a partir da *engine* V8 do Google Chrome, foi projetado para sanar o problema de gargalos em servidores, que executavam aplicações que operavam de forma bloqueante, onde as requisições são enfileiradas e executadas ao final de outras tarefas I/O (entrada e saída).

Segundo (IHRIG, 2014) o modelo assíncrono do Node.js permite alta escalabilidade e baixa sobrecarga. Esse modelo trabalha em uma arquitetura completamente não bloqueante, cada processo em Node.js possui uma *thread* prevenindo possíveis *dead-locks* [1].

Em um relatório apresentado pelo GitHub, foram apresentadas as maiores dependências dos projetos de código aberto, como destacado na figura 3 o *node server* é exibido na sétima posição do *ranking*, comprovando a grande adoção dos desenvolvedores pelo *framework*.

Figura 3 – Maiores dependências de projetos de código aberto



Fonte: (GitHub, 2019)

Considerando a alta escalabilidade e a baixa sobrecarga de processamento, o Node.js será utilizado em conjunto com a biblioteca mongoose durante a etapa de armazenamento e disponibilização dos dados.

[1] Operações bloqueantes, onde um processo fica aguardando outro para continuar a execução

3.7.1 Mongoose

Mongoose é uma biblioteca do Node.js que provê uma solução baseada em esquemas, para modelar os dados de uma aplicação, é semelhante a alguns ORMs (*Object Relatinal Mapping*) ou seja, é uma interface para "tradução" dos dados do banco de dados para objetos de uma linguagem, como JavaScript.

Mongoose oferece alguns métodos que facilitam no gerenciamento dos dados como validação, criação, atualização, remoção e consulta de dados (Mongoose, 2019). A biblioteca também possui alguns métodos que implementam *queries* mais avançadas para tratamento de dados em massa, como por exemplo a atualização de vários registros.

Além das funções já disponibilizadas pela biblioteca é possível adicionar *plug-ins* que permitem estender as funcionalidades do Mogoose.

3.8 JavaScript

JavaScript é uma linguagem de programação que permite implementar funcionalidades mais complexas em páginas *web* (MOZILLA, 2019).

A linguagem foi criada na década dos anos 90 por Brendan Eich na empresa Netscape a fim de tornar a web mais dinâmica. De acordo com (FLANAGAN, 2004) JavaScript é uma linguagem de alto nível, dinâmica, interpretada e não tipada, conveniente para estilos de programação orientados a objeto e funcionais.

A evolução da linguagem e o desenvolvimento de *frameworks* e bibliotecas pela comunidade possibilitou outras aplicações para o JavaScript como REST APIs, testes unitários, banco de dados, gerenciadores de pacotes, desenvolvimento de jogos, aplicativos moveis, *desktops* e para *smartTVs*.

Neste projeto a linguagem será utilizada no backend com o Node.js, e no frontend com o React.js que será descrito a seguir.

3.8.1 React

React é uma biblioteca JavaScript para construção de interfaces de usuário, tornando mais simples a criação de sistemas interativos. O React realiza a atualização e renderização de componentes de uma página conforme alterações realizados no estado da aplicação, poupando recarregamento de paginas e requisições, tornando uma experiencia mais fluida durante a navegação do usuário (React, 2020).

A criação de componentes é uma característica da biblioteca importante para o desenvolvimento desse projeto, sendo possível combinar vários componentes para montagem

de uma interface, separando-os em pequenas blocos, que podem ser reutilizados e que tornam o processo de manutenção mais simples.

Durante o desenvolvimento da aplicação será utilizado o *framework* material ui que possui um conjunto de componentes React, estilizados com características comuns na maioria das aplicações tornando o desenvolvimento mais ágil e fácil. Material UI possui uma documentação ampla e um blog com exemplos de implementações e boas praticas que podem ser seguidas com o *framework* (Material-UI, 2020).

3.8.2 Google Maps Javascript API

Desenvolvido em 2005 pela empresa Google, o Google Maps é um serviço gratuito de pesquisa e visualização de mapas e imagens de satélite da Terra (MAPS, 2019).

O serviço oferece mapas com nomes de ruas, rodovias, com possibilidades de traçar rotas com e destino, além disso o Google Maps engloba outros serviços como a descoberta de locais turísticos, hotéis, bares, restaurantes e a visualização de imagens terrestres ao nível do solo com o Google Street View.

A Google Maps JavaScript API permite a customização do Google Maps, para exibição em páginas web e dispositivos móveis. A API apresenta quatro tipos básicos de mapas (roteiro, satélite, híbrido, e terreno) que pode ser modificado usando camadas e estilos, controles e eventos e vários outros serviços da biblioteca (MAPS, 2019).

Um dos recursos disponibilizados pela Google Maps Javascript API são os *Markers* (marcadores) que permitem identificar uma localização no mapa através da sua latitude e longitude. Os marcadores serão utilizados nesse trabalho para identificação dos países referentes as publicações coletadas.

4 Metodologia e Desenvolvimento

A metodologia adotada para o desenvolvimento desse trabalho segue algumas das etapas do KDD definidas por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A seguir, a implementação de cada etapa será descrita, conforme os objetivos do presente trabalho.

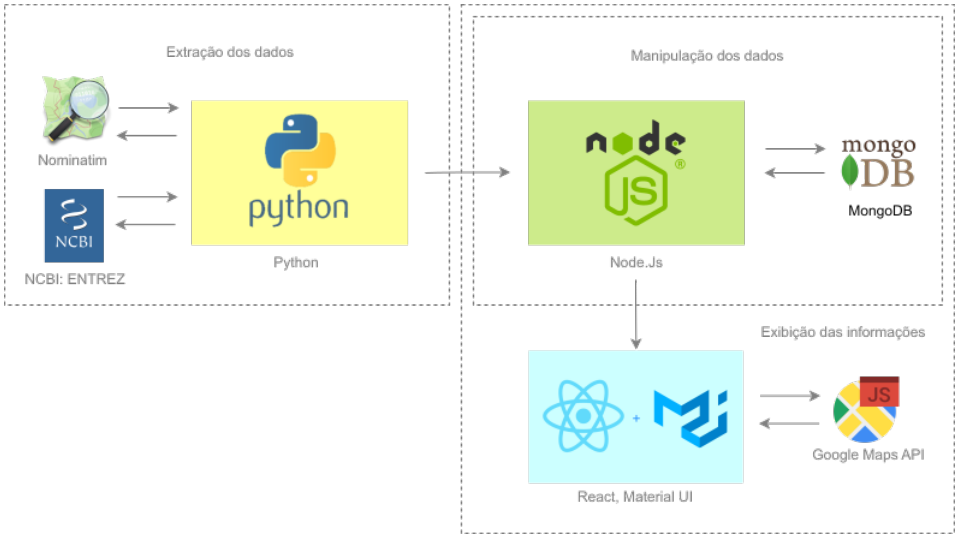
4.1 Preparação

Primeiramente, foram realizadas pesquisas para entendimento do termo "depressão", compreendendo melhor a "visão do cliente". Nessa etapa, foram identificadas e selecionadas as palavras-chave relacionadas ao tema.

Em seguida, foi realizada a seleção, uma amostra de dez artigos foi coletada através de um algoritmo desenvolvido em Python que realiza o *download* dos artigos no formato XML através da ferramenta efetch da biblioteca Entrez e-utils.

As amostras obtidas permitiram iniciar a próxima etapa definida por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996): limpeza e pré-processamento, onde foram encontrados os padrões de estrutura dos artigos. Para as próximas etapas foi elaborado um diagrama (figura 4) com a visão geral da aplicação, assim foi possível visualizar o fluxo de dados e as ferramentas que se comunicam diretamente.

Figura 4 – Visão geral da aplicação



Fonte: Próprio autor

4.2 Extração dos dados

Na extração de dados, todo conteúdo obtido da base PMC através da biblioteca Entrez E-utils é tratado e modelado para inserção no banco.

4.2.1 Mineração

Para coleta dos dados essenciais para o desenvolvimento do projeto como: título do artigo, palavras-chave, autores, instituição, país, entre outras, foi desenvolvido um *script* em Python, utilizando o paradigma de POO (Programação orientada a objetos). Foi desenvolvida uma classe nomeada de *Mining* que é responsável pela mineração dos dados.

A classe *Mining* é construída informando o termo que será utilizado para a busca. O método *start* é responsável por iniciar o processo de mineração, nele é chamado o método *search* que faz uma chamada para a rota *esearch* da API Entrez *eutils*. A resposta contém uma lista de IDs e informações como um contador do número de páginas encontradas para o termo informado, um exemplo de resposta pode ser visualizado no apêndice A - 1, figura 11.

A lista de IDs obtida é percorrida, e para cada item é chamado a rota *efetch* da Entrez *eutils*, que retorna o XML com as informações do artigo como o corpo do artigo, editora, revista, categorias, entre outros.

4.2.2 Padronização

Alguns campos podem não ser retornados para determinados artigos, e muitos não possuem um padrão quanto a estrutura de marcação do XML, como por exemplo informações de endereço. Para tanto é necessário tratar cada exceção, os dados coletados são padronizados a fim de obter um dicionário[1] que será transformado em objeto javascript JSON, conforme o apêndice B - 2, figura 12. A classe responsável por esse processo foi nomeada de *Parser*.

Para identificação do país foi utilizado um banco de dados *open source*, que possui as expressões regulares e as variantes em diversos idiomas dos nomes dos países, além de dados de latitude e longitude de cada país. São efetuadas duas tentativas, primeiramente é realizado um teste com a expressão regular de cada país registrado no banco de dados, caso não ocorra uma combinação com o texto preenchido no campo país do artigo, outra tentativa é realizada buscando o texto na API Nominatim, caso seja encontrado um ou vários resultados o primeiro é definido como o país do artigo. Quando o campo país não é preenchido, o campo instituição do autor é utilizado para essa busca.

[1] Em Python, um dicionário é uma coleção desordenada, mutável, indexada e possuem chaves e valores

4.3 Manipulação dos dados

Após a obtenção dos dados, a aplicação Python realiza uma requisição na API Node.js para inserção dos dados no banco de dados Mongo. Ao receber a requisição, a aplicação Node.js faz uma busca no banco através do *id* PMC. Caso o artigo não seja encontrado, é criado um novo registro, caso contrario é realizado apenas uma atualização.

Em seguida, é iniciado um processo para inserção das palavras-chave do artigo, esse processo não é bloqueante, ou seja, a resposta da requisição é enviada para a aplicação Python antes do término do processamento das palavras-chave, otimizando o processo de extração de dados.

Essa etapa tende a consumir grande esforço computacional devido a manipulação dos dados e inserção no banco, a solução adotada de processar os dados em segundo plano foi fundamental para reduzir o tempo gasto durante a mineração dos dados.

4.4 Exibição das informações

Para exibição das informações obtidas para o usuário, foi desenvolvido um sistema *web* com o *framework* React e a biblioteca Material UI além da API JavaScript do Google Maps, o sistema é constituído de três telas: Mapa, Gráficos e Assuntos Frequentes.

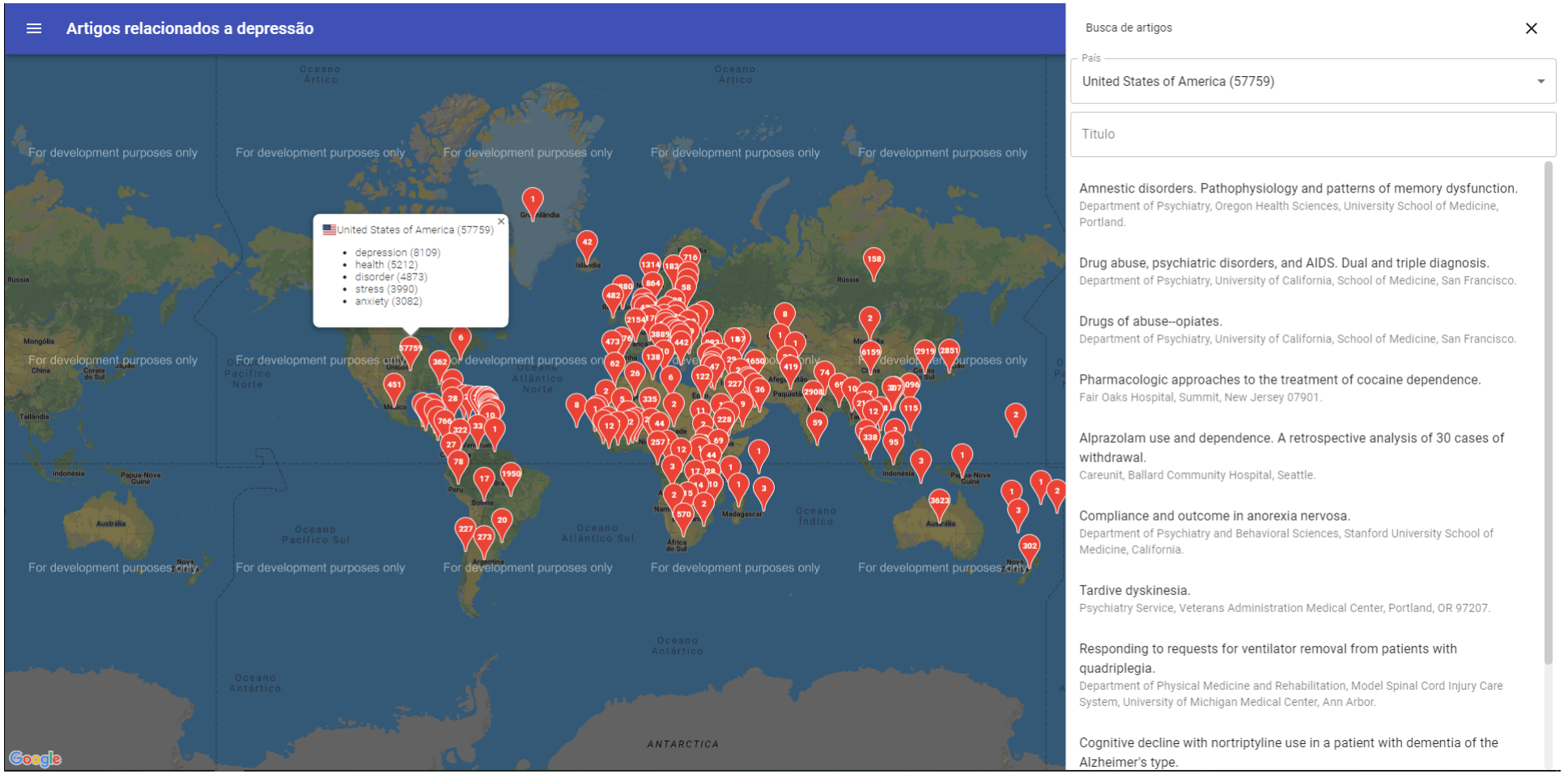
O mapa é a tela principal do sistema. Nele são exibidos todos os Países com a contagem de publicações, conforme a figura 5. Ao clicar em um ponto no mapa são exibidas as palavras-chave mais utilizadas nas publicações do país e suas respectivas contagens.

Na barra lateral são exibidas as publicações, sendo possível acessá-las ao clicar no título. Além disso, é possível adicionar filtro por país e realizar buscas pelo título da publicação.

Durante o desenvolvimento do *frontend*, foram isolados componentes que possuem regras de funcionamento comuns, para que possam ser utilizados por outros componentes, como por exemplo uma lista de itens, onde é possível carregar mais itens conforme o usuário realiza a ação de rolagem sobre a lista. Essa é uma prática comumente adotada durante o desenvolvimento com React ou outros frameworks *frontend*.

Ao fim dessa etapa, foi possível visualizar todas as informações que foram coletadas. No próximo capitulo serão apresentados os resultados e uma breve análise dos dados.

Figura 5 – Sistema web - Mapa



Fonte: Próprio autor

5 Resultados e Análise

Neste capítulo serão apresentados os resultados obtidos no projeto, além de estatísticas coletadas. Por fim é exposta uma análise sobre algumas informações observadas.

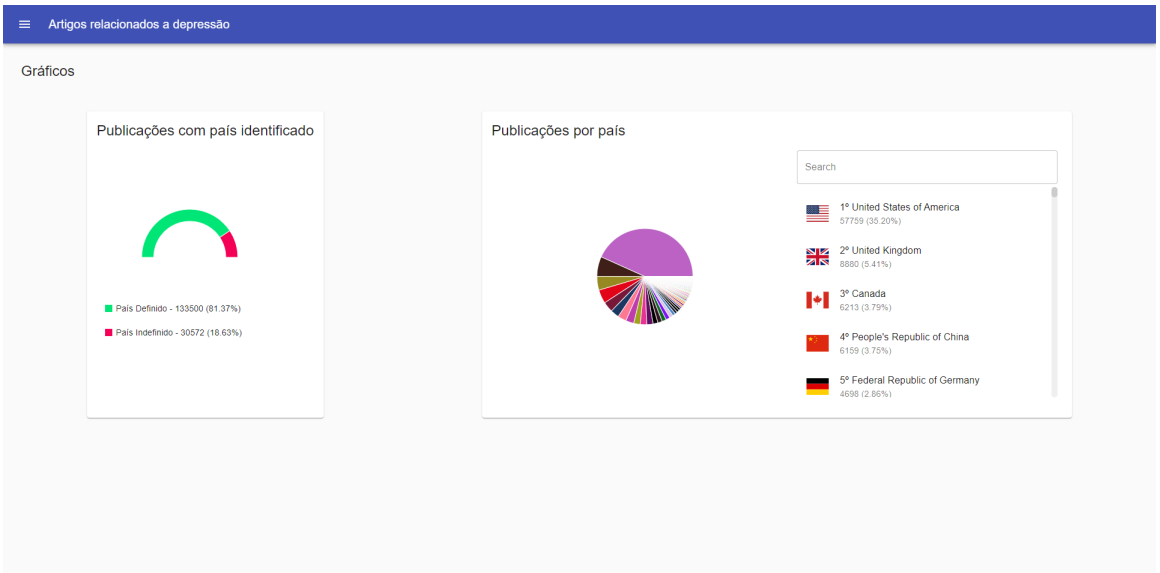
5.1 Gráficos

Os gráficos estão disponíveis no sistema web (figura 6) e pode ser acessado através do menu. Por ele estatísticas dos dados coletados são apresentados de forma visual.

No primeiro gráfico (Publicações com País identificado) observa-se que em cerca de 80% dos artigos foi possível identificar o País de origem.

No segundo gráfico (Publicações por País) observa-se que os Estados Unidos da América foi o País que mais apresentou artigos relacionados a depressão 57.759, cerca de 35,2% em relação a quantidade total coletada, em seguida são listados Reino Unido 5,41%, Canada 3,79%, China 3,75% e Alemanha 2,86%.

Figura 6 – Sistema web - Gráficos



Fonte: Próprio autor

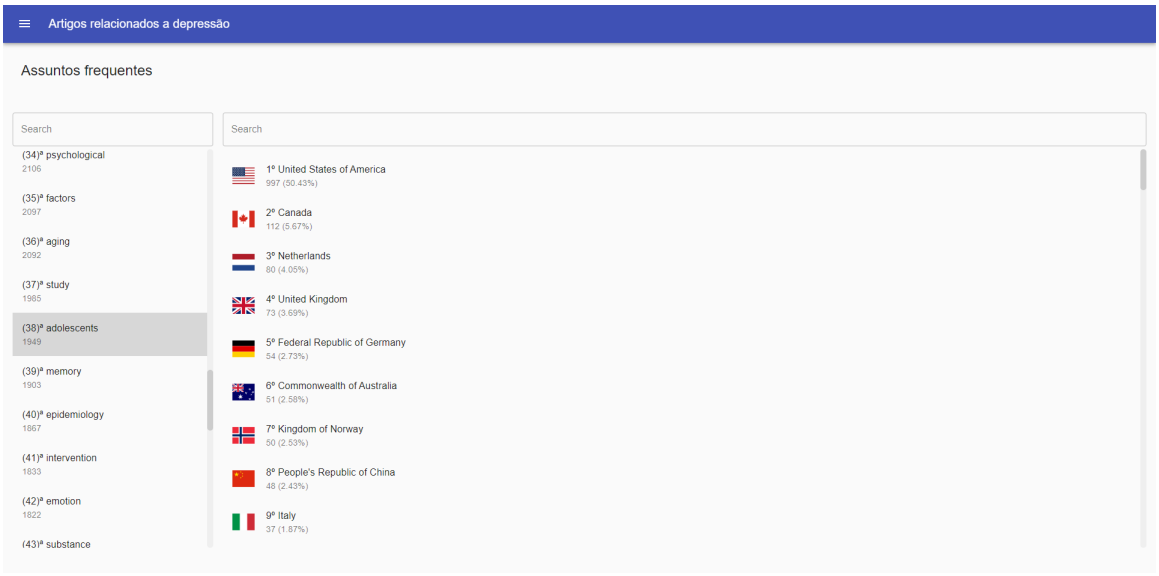
5.2 Assuntos Frequentes

Na página de assuntos frequentes é exibida uma listagem com as palavras-chave mais utilizadas nos artigos, ao selecionar uma palavra é exibida uma lista com os Países

que estão utilizando a palavra em seus artigos, sendo possível identificar qual País está pesquisando em maior quantidade determinado assunto.

Na figura 7 observa-se que o termo adolescentes foi citado em 1.949 artigos. Os países que mais possuem publicações com esse termo são Estados Unidos 50,43%, Canadá 5,67%, Holanda 4,05% e Reino Unido 3,69%.

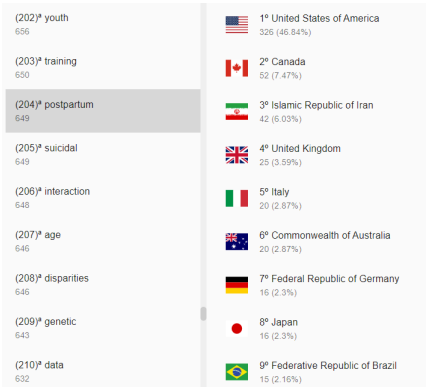
Figura 7 – Sistema web - Assuntos Frequentes



Fonte: Próprio autor

Analisando os tipos de depressão, foi observado que em alguns países há um domínio maior sobre alguns tipos. Um exemplo é o Irã, possui 1.660 artigos na base de dados coletada, sendo o 16º país no *ranking* de quantidade de artigos publicados, porém para o termo *postpartum* é o terceiro país com mais publicações (figura 8).

Figura 8 – Artigos relacionados a depressão pós-parto

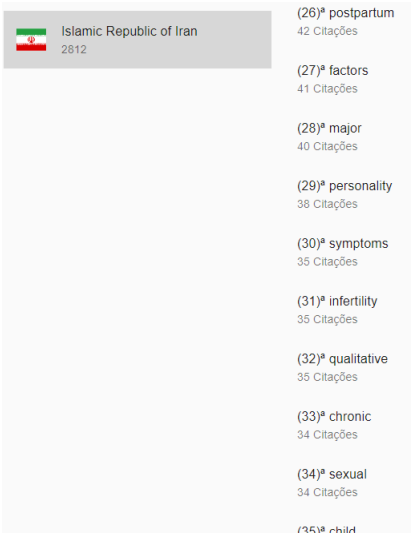


Fonte: Próprio autor

Visualizando os assuntos pesquisados do país (figura 9), é observado que termos

como "infertilidade", "sexual"e "criança"são muito citados, o que pode confirmar a existência de muitas pesquisas relacionadas a depressão pós-parto.

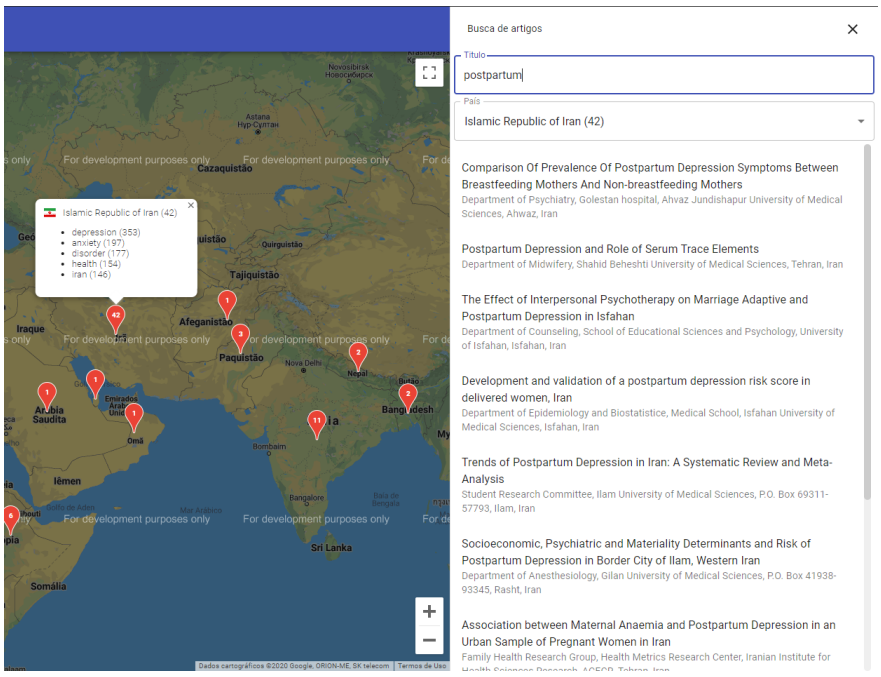
Figura 9 – Assuntos frequentes no Irã



Fonte: Próprio autor

Através do mapa, filtrando pelo país Irã e pelo termo *postpartum* é possível visualizar o título e acessar o artigo completo clicando sobre ele. Na figura 10 são exibidos alguns artigos que analisam tendências da doença e associações a fatores socioeconômicos.

Figura 10 – Artigos relacionados a depressão pós-parto no Irã



Fonte: Próprio autor

No Brasil, os assuntos predominantes estão relacionados ao envelhecimento. Termos como "memória", "envelhecida", "enfraquecimento"e "idosos"aparecem entre os trinta primeiros itens da listagem de termos frequentes, indicando que possivelmente existem muitas pesquisas relacionada a depressão em idosos no País.

Essa análise se tornou possível após a mineração e extração de informações do país de origem das publicações coletadas, processo apresentado no capítulo 4, com os dados classificados a busca se torna mais rápida além de incluir publicações que não apresentavam informações do país de forma explícita. Na figura 10 para realizar a busca, a aplicação *web* consulta a API Node.js que tem a simples responsabilidade de executar um consulta no banco na coleção de publicações com os parâmetros país e título correspondentes ao país Iran e ao termo "postpartum".

Como exemplificado nessa seção, o presente trabalho oferece de maneira simplificada ferramentas para a extração de conhecimento sobre pesquisas já realizadas relacionadas a depressão por determinado país.

6 Considerações Finais e Trabalhos Futuros

O estudo realizado nesse trabalho possibilita a análise dos termos frequentes e classificação dos países de publicações científicas relacionadas a depressão, agregando informações a partir dos dados coletados da base de dados da PubMed Central através das ferramentas de busca e transferência da biblioteca Entrez.

Conclui-se que o objetivo geral de definir o país das publicações coletadas foi atingido. Com uma definição dos países em cerca de 80% dos artigos coletados, foi possível exibir estatísticas de cada país e termos mais frequentes. O sistema *web* desenvolvido apresenta as principais informações coletadas de forma intuitiva ao usuário, respondendo rapidamente as requisições solicitadas.

Em meio a grande quantidade de dados disponíveis no repositório da PubMed, o trabalho desenvolvido agrega ao meio acadêmico oferecendo ferramentas que auxiliam no processo de pesquisa de informações relacionadas a depressão.

Os dados obtidos nesse trabalho, possibilitaram a identificação de quais assuntos relacionados a depressão determinado país está pesquisando, permitindo identificar defasagens de estudos, análises e correlação desses dados a outras estatísticas, como por exemplo o número de casos diagnosticados.

Em projetos futuros, espera-se analisar os dados obtidos, colher mais informações, disponibilizando estatísticas detalhadas de cada país. Uma possibilidade para ampliação do projeto é a coleta de informações sobre outros temas, tendo em vista que com poucos ajustes é possível generalizar os *scripts* desenvolvidos. Outro possibilidade é a implementação de redes neurais artificias (RNAs) associadas a repositórios de artigos científico viabilizando o aprendizado de máquina e identificação automatizada de padrões nos artigos.

Por fim, é necessário disponibilizar a ferramenta desenvolvida *online* e apresentar para pesquisadores, psicólogos e psiquiatras especializados no tema (depressão) com o objetivo de mapear, selecionar e complementar os dados coletados, gerando novas informações sobre os artigos.

Referências

AGRAWAL, R. et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, AAAI/MIT Press Menlo Park, CA, v. 12, n. 1, p. 307–328, 1996. Citado na página 9.

AMORIN, V. A depressão na atualidade: diagnóstico e tratamento. 2014. Citado na página 5.

BORGES, L. E. *Python para desenvolvedores: aborda Python 3.3*. [S.l.]: Novatec Editora, 2014. Citado na página 11.

BORNMANN, L.; LEYDESDORFF, L.; KRAMPEN, G. Which are the "best" cities for psychology research worldwide? PsychOpen, 2012. Citado 2 vezes nas páginas 3 e 4.

CHODOROW, K. *MongoDB: the definitive guide: powerful and scalable data storage*. [S.l.]: "O'Reilly Media, Inc.", 2013. Citado na página 13.

Cisco. *Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper*. Cisco, 2020. Disponível em: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>. Citado na página 1.

DANTAS, E. R. G. et al. O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisões. *V Simpósio de Excelência em Gestão e Tecnologia*, p. 1–10, 2008. Citado na página 8.

DATE, C. J. *Introdução a sistemas de bancos de dados*. [S.l.]: Elsevier Brasil, 2004. Citado na página 12.

DECANDIA, G. et al. Dynamo: amazon's highly available key-value store. In: ACM. *ACM SIGOPS operating systems review*. [S.l.], 2007. v. 41, n. 6, p. 205–220. Citado na página 13.

FARIA, A. A. Coleta, integração e caracterização de base de dados de câncer. 11 2014. Citado 2 vezes nas páginas 3 e 4.

FAVA, M.; KENDLER, K. S. Major depressive disorder. *Neuron*, Elsevier, v. 28, n. 2, p. 335–341, 2000. Citado na página 5.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996. Citado 5 vezes nas páginas 2, 7, 8, 9 e 17.

FLANAGAN, D. *JavaScript: o guia definitivo*. [S.l.]: Bookman Editora, 2004. Citado na página 15.

GitHub. *The State of the Octoverse*. 2019. Disponível em: <https://octoverse.github.com/>. Citado 2 vezes nas páginas 11 e 14.

HAND, D. J.; SMYTH, P.; MANNILA, H. *Principles of Data Mining*. Cambridge, MA, USA: MIT Press, 2001. ISBN 0-262-08290-X, 9780262082907. Citado na página 9.

IHRIG, C. J. *Pro node. js for developers*. [S.l.]: Apress, 2014. Citado na página 14.

JAIN, A. K.; DUBES, R. C. Algorithms for clustering data. *Englewood Cliffs: Prentice Hall, 1988*, 1988. Citado na página 9.

LAROSE, D. T.; LAROSE, C. D. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014. Citado na página 9.

LAYTON, R. *Learning data mining with python*. [S.l.]: Packt Publishing Ltd, 2015. Citado na página 11.

LOHR, S. The age of big data. *New York Times*, v. 11, n. 2012, 2012. Citado na página 1.

MAPS. *Maps JavaScript Api. Google Maps Platform*. 2019. Disponível em: <https://developers.google.com/maps/documentation/javascript>. Citado na página 16.

Material-UI. *Material-UI: Um framework popular de React UI*. 2020. Disponível em: <https://material-ui.com/pt/>. Citado na página 16.

Mongoose. *Mongoose ODM*. 2019. Disponível em: <https://mongoosejs.com/>. Citado na página 15.

MOZILLA. *JavaScript. MDN web docs*. 2019. Disponível em: <https://developer.mozilla.org/pt-BR/docs/Aprender/JavaScript>. Citado na página 15.

NCBI. *Entrez Help*. U.S. National Library of Medicine, 2006. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK3837/>. Citado na página 10.

NCBI. *Entrez Programming Utilities Help - NCBI Bookshelf*. U.S. National Library of Medicine, 2008. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK25501/>. Citado na página 10.

NCBI. *PMC - NCBI*. U.S. National Library of Medicine, 2019. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/>. Citado na página 10.

NOSQL. *NOSQL Database Management Systems*. 2019. Disponível em: <http://nosql-database.org>. Citado na página 13.

PARTONEN, T.; LÖNNQVIST, J. Seasonal affective disorder. *CNS Drugs*, v. 9, n. 3, p. 203–212, Mar 1998. ISSN 1179-1934. Disponível em: <https://doi.org/10.2165/00023210-199809030-00004>. Citado na página 6.

PORTO, J. A. A. D. Conceito e diagnÃ. *Brazilian Journal of Psychiatry*, scielo, v. 21, p. 06 – 11, 05 1999. ISSN 1516-4446. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-44461999000500003&nrm=iso>. Citado na página 5.

Python. *The ElementTree XML API. Python documentation*. 2019. Disponível em: <https://docs.python.org/2/library/xml.etree.elementtree.html>. Citado na página 12.

React. *React – Uma biblioteca JavaScript para criar interfaces de usuário*. 2020. Disponível em: <https://pt-br.reactjs.org/>. Citado na página 15.

RICARTE, I. L. M. Programação orientada a objetos: uma abordagem com java. *http://www.dca.fee.unicamp.br/cursos/PooJava/Aulas/poojava.pdf> Acesso em*, v. 29, n. 10, p. 2014, 2001. Citado na página 10.

The National Institute of Mental Health. *Depression*. NIMH, 2018. Disponível em: <https://www.nimh.nih.gov/health/topics/depression/index.shtml>. Citado 3 vezes nas páginas 5, 6 e 7.

WEISS S. I., K. C. Computer systems that learn: Classification and prediction methods from statistics, neural networks, machine learning, and expert systems. 1991. Citado na página 9.

World Health Organization. *Depression*. World Health Organization, 2018. Disponível em: <https://who.int/news-room/fact-sheets/detail/depression>. Citado 3 vezes nas páginas 1, 2 e 5.

Apêndices

APÊNDICE A – 1

A figura 11 corresponde a uma resposta de busca realizada pelo utilitário *e-search* da ferramenta Entrez, a resposta recebida é um XML composto por uma lista de IDs de publicações encontradas para o termo pesquisado.

Figura 11 – Exemplo de resposta esearch


```
<eSearchResult>
  <Count>37</Count>
  <RetMax>20</RetMax>
  <RetStart>0</RetStart>
  <IdList>
    <Id>7249014</Id>
    <Id>7145631</Id>
    <Id>7162489</Id>
    <Id>7135272</Id>
    <Id>7077981</Id>
    <Id>7016367</Id>
    <Id>6868504</Id>
    <Id>6835476</Id>
    <Id>6731011</Id>
    <Id>6694732</Id>
    <Id>6688443</Id>
    <Id>6658306</Id>
    <Id>6542324</Id>
    <Id>6371893</Id>
    <Id>6407012</Id>
    <Id>6305677</Id>
    <Id>6308717</Id>
    <Id>6326255</Id>
    <Id>5996178</Id>
    <Id>5838108</Id>
  </IdList>
  <TranslationSet/>
  <TranslationStack>
    <TermSet>
      <Term>depressive[All Fields]</Term>
      <Field>All Fields</Field>
      <Count>186196</Count>
      <Explode>N</Explode>
    </TermSet>
    <TermSet>
      <Term>JSON[All Fields]</Term>
      <Field>All Fields</Field>
      <Count>3083</Count>
      <Explode>N</Explode>
    </TermSet>
    <OP>AND</OP>
    <OP>GROUP</OP>
  </TranslationStack>
  <QueryTranslation>depressive[All Fields] AND JSON[All Fields]</QueryTranslation>
</eSearchResult>
```


Fonte: Próprio autor

APÊNDICE B – 2

A figura 12 representa o *schema* de uma publicação que é salvo no banco de dados, apenas os campos *title* (titulo), *pmc* (id referente ao cadastro na PubMed) e *pubDate* (data de publicação) são definidos como obrigatórios.

Figura 12 – Modelo Publicação

```
title: {
    type: String,
    required: true,
},
abstract: {
    type: Object,
    required: false,
},
pmid: {
    type: Number,
    required: false,
    index: true,
},
pmc: {
    type: Number,
    required: false,
    index: true,
    unique: true
},
doi: {
    type: String,
    required: false,
},
publisherId: {
    type: String,
    required: false,
},
elocationId: {
    type: String,
    required: false,
},
contributors: {
    type: Array,
    required: false,
},
journal: {
    type: Object,
    required: false,
},
pubDate: {
    type: Array,
    required: true,
},
keywords: {
    type: Array,
    required: false,
},
categories: {
    type: Array,
    required: false,
},
affiliations: {
    type: Array,
    required: false,
},
country: {
    type: Schema.Types.ObjectId,
    ref: 'Country',
    required: false,
},
```

Fonte: Próprio autor