Extração de informação semi-estruturada em bulas de medicamentos

Mini-relatório 3

Proponente(s):
RA 21001116 - Juliane Kristine de Lima
RA 11004813 - Wedeueis Braz da Silva

Santo André, 08 de maio de 2018

Resumo

Atualmente, existe uma grande quantidade de medicamentos em circulação, sendo que vários deles apresentam as mesmas finalidades terapêuticas. Porém, mesmo que tenham a mesma finalidade, cada uma das drogas apresenta características únicas, que são descritas na bula do medicamento. Essas características incluem diferenças em ingredientes, posologia, contraindicações, efeitos colaterais, entre outros. Considera-se que, com a grande disponibilidade de informações técnicas a respeito dos medicamentos, incluindo os mais diferentes fármacos e suas especificidades, mostra-se interessante a criação de um dispositivo automatizado de coleta de informações de bulas de medicamentos, que permita ao usuário a comparação das características de vários fármacos. Dessa forma, o objetivo deste projeto é modelar um dispositivo de busca e comparação de informações em bulas de fármacos, que permita realizar comparações entre diferentes medicamentos que tenham a mesma finalidade e possibilitar ao usuário uma escolha mais bem fundamentada. Nesse sentido, buscaremos implementar um sistema orientado a objetos utilizando a linguagem Python, em que serão calculadas medidas de similaridade de texto, entre determinadas sessões das bulas. Assim, os sistema permitirá que as informações relevantes do corpus sejam resgatadas e comparadas, de forma que a partir de uma certa entrada relacionada a um medicamento, apresentará como saída quais são os fármacos semelhantes, a partir de estimativas relacionadas aos valores de similaridade.

Palavras-chave: Processamento de Linguagem Natural, Recuperação de Informação, Bulas de medicamentos, Similaridade, Grafos.

1 Introdução

Em um contexto de grande informatização de textos e documentos e de sua disponibilização na Internet de forma não estruturada, surge a necessidade de uma melhor organização e sistematização das informações, buscando-se evitar possíveis perdas e alterações de significado. Utilizando conceitos considerados em processamento de linguagem natural, busca-se realizar uma proposta de trabalho relacionada à extração de informações em bulas de medicamentos. Atualmente, estão amplamente disponibilizadas bulas e especificações técnicas de compostos medicamentosos, que podem ser usados como base de dados para um algoritmo de processamento de linguagem natural. A Agência Nacional de Vigilância Sanitária (Anvisa) regulamentou a disponibilização de bulário eletrônico (ANVISA, 2009), sendo que estão disponíveis no site da instituição mais de 7000 bulas de medicamentos (`http://www.anvisa.gov.br/datavisa/fila_bula/index.asp`).

Percebeu-se na literatura que o interesse nesta área ainda é muito recente e insuficientemente explorado. Nogueira (2014), em trabalho de conclusão de curso, buscou utilizar um algoritmo de mineração de textos para extrair padrões em uma base de bulas de medicamentos, utilizando o modelo Cassiopeia. Este trabalho utilizou vários sites que disponibilizam bulas de medicamento como fonte de dados.

Em um trabalho de Silva et al. (2015), ressalta-se a desestruturação das informações contidas no bulário eletrônico direcionado a profissionais, da Anvisa. O objetivo do trabalho foi de propor uma metodologia semiautomática de mineração de textos do bulário eletrônico da Anvisa, juntamente com duas outras bases de dados, uma a respeito de drogas e medicamentos e outra sobre a classificação de doenças. Em sua dissertação de mestrado, o mesmo autor (Silva, 2016), expõe com mais detalhes o sistema de mineração de dados que realizou durante sua formação, trabalho do qual originou-se o site Bula Fácil (`http://facilbula.com.br/`). Nos trabalhos de Silva (2016), buscou-se realizar um algoritmo de auxílio à comunidade médica, e por isso somente foi utilizado como base de dados as bulas para profissionais, e não as bulas simplificadas.

2 Objetivos

2.1 Objetivo Geral

Possibilitar realização de análise das informações contidas nas bulas e comparação entre medicamentos, utilizando ferramentas de Processamento de Linguagem Natural (PLN)

A partir de dados de bulas simplificadas para o paciente, disponíveis publicamante no bulário eletrônico (`http://www.anvisa.gov.br/datavisa/fila_bula/index.asp`) da Anvisa, criar uma ferramenta informatizada de análise de medicamentos, que possibilite buscar e comparar medicamentos por similaridade semântica, utilizando técnicas de PLN para extrais informações contidas nas sessões de indicações e composição.

2.2 Objetivos Específicos

- **Obter dados publicos das bulas de medicamentos**

  Baixar bulas de pacientes disponíveis no bulário eletronico da anvisa que servirão como base de dados para extrair as informações de interesse. As principais informações a serem analisadas estão contidas nas sessões de indicações de uso e componentes. No futuro o sistema poderá ser

expandido para incluir comparações de precauções, contraindicações, interações com outros medicamentos etc.

- **Extrair as informações contidas nas bulas**

  Criar um programa "parser" que, aproveitando-se do fato das informações contidas nas bulas serem padronizadas segundo legislação da Anvisa, separe as sessões da bula em diferentes listas que ficarão armazenadas nas variáveis dos objetos Bula do sistema.

- **Classificar os medicamentos segundo sua similaridade**

  Usando os conteúdos das listas referentes às sessões de identificação e indicações da bula, comparar os diversos medicamentos através de técnicas de similaridade semântica atribuindo valores de similaridade a cada par de medicamentos analisado.

- **Criar uma estrutura de grafo para armazenar as informações obtidas**

  Os objetos Bula criados com as informações obtidas ficarão armazenados nos nós de uma estrutura de grafo onde as arestas entre dois elementos contém o valor de similaridade entre eles.

- **Criar um mecanismo de busca**

  Com a estrutura de grafo criada podemos utilizar um mecanismo de busca sobre ela que retorne uma certa quantidade de medicamentos semelhantes utilizando os valores de similaridade.

3 Fundamentos

Serão discutidos a seguir os principais conceitos que serão utilizados neste trabalho.

3.1 Bulas de medicamentos

Bula é um documento legal sanitário que contém informações técnico-científicas e orientadoras sobre os medicamentos para o seu uso racional. Bula para o paciente é uma uma bula destinada ao paciente, aprovada pela Anvisa, com conteúdo sumarizado, em linguagem apropriada e de fácil compreensão ANVISA (2009). Segundo o artigo 6º e o Anexo I da RDC 47 de 2009 que padroniza e regulamenta o conteúdo das bulas, a bula para o paciente deve conter obrigatoriamente as seguintes sessões: "Identificação do Medicamento", "Informações ao Paciente" e "Dizeres Legais". Mais especificamente na sessão "Informações ao Paciente" o primeiro item "1. Para que este Medicamento é Indicado?" Descreve as indicações do medicamento e será a principal fonte de dados para este trabalho.

3.2 Extração de Informação

Processo no qual textos ou partes textuais de interesse são extraídos de documentos e utilizadas pelo sistema para construir um modelo de dados. Podemos realizar a recuperação de informação utilizando informações contidas em metadados do cabeçalho ou, se o documento apresenta estrutura definida e fixa, usar técnicas de casamento de padrões (Pattern Matching) para localizar o texto a ser extraído. No caso das bulas as sessões são padronizadas segundo legislação, o que facilita o processo de recuperação de informação utilizando os títulos da sessão de interesse e da sessão seguinte como

os limites do conteúdo a ser extraído. Para explicações detalhadas sobre técnicas de casamento de padrões em cadeias de caracteres consultar JURAFKSKY (2009).

3.3 Similaridade entre documentos

Em Processamento de Linguagens Naturais existem técnicas para comparar documentos segundo seu conteúdo, para tal podemos usar um tesauro ou usar a distribuição de palavras nos documentos. Na técnica conhecida como Matriz Termo-Documento, definida no capítulo 15 (Vector Semantics) de JURAFKSKY (2009), criamos uma representação matricial onde as colunas representam os diferentes documentos e as linhas as diferentes palavras. Assim uma única coluna apresenta a distribuição de palavras de um documento e pode ser entendida como uma representação vetorial do mesmo. Desta forma podemos considerar dois documentos como similares se seus vetores forem similares e definir uma função de similaridade baseada na diferença vetorial.

3.4 Grafos

Matematicamente um grafo é um par de conjuntos, um conjunto de vértices e um conjunto de arcos ou arestas, onde cada aresta é um par ordenado de vértices. Podemos construir uma estrutura de dados baseada em grafos que armazena elementos nos vértices e as relações entre eles nas arestas e aproveitar os algoritmos, como os apresentados no capítulo VI de CORMEN(2009), desenvolvidos pela Teoria dos Grafos (ramo da matemática que estuda grafos), para estudar as propriedades destas relações. A estrutura de grafo é especialmente útil para estudar relações de distância, pois para cada elemento podemos acessar rapidamente seus "vizinhos" através de suas arestas.

4 Método

Será desenvolvido um sistema orientado a objetos utilizando a linguagem Python que, com uma técnica baseada em comparação de cadeias de caracteres, extrai informações de um arquivo no formato PDF, que contém as informações da bula do paciente obtido no bulário eletrônico da Anvisa, e as separa por sessão em variáveis de um objeto Bula. O objeto Bula deve conter variáveis para armazenar o conteúdo das sessões contendo o nome do medicamento, sua identificação, seus componentes e suas indicações. Com as informações contidas nas sessões de identificação e indicações o sistema deve calcular a similaridade entre duas Bulas utilizando técnicas de similaridade entre documentos e criar um grafo não dirigido contendo todos os objetos Bula em seus nós e o valor de similaridade nas arestas entre eles. O grafo será utilizado para realizar buscas de medicamentos semelhantes utilizando o valor de similaridade contido nas arestas.

5 Resultados

O método desenvolvido foi implementado em um programa orientado à objetos composto por três classes que dividem as responsabilidades de armazenamento e processamento dos dados extraídos das bulas.

5.1 Classe Bula

- Extrai as informações das seções de interesse utilizando expressões regulares e as rmazena. Por exemplo a **E.R.** *INDICADO\?(?s)(.*?)2*. extrai todo conteúdo presente entre o fim de frase "INDICADO?" e o número de início de seção "2.".
- Processa o texto extraído em *Tokens*, conjunto de símbolos com significado coletivo, e retira as *Stopwords*, palavras sem relevância semântica para o problema. O resultado fica armazenado em variáveis do tipo lista do objeto. Armazena em ordem decrescente uma lista de bulas similares que podem ser adicionadas utilizando o método *inserirSimilar()*.

5.2 Classe Medicamentos

- Cria e mantém uma lista de nomes e um dicionário de medicamentos, contendo os objetos bula como valor e o nome do medicamento como chave.
- Computa e mantém dois vocabulários, um contendo o conjunto das palavras das seções de indicação e outro das seções de composição analisadas até então.

5.3 Classe GrafoSimilaridade

- Guarda uma referência para o objeto *Medicamentos* e, usando seus vocabulários, realiza o cálculo de frequência de palavras para cada bula para então computar uma matriz termo-documento para cada seção analisada.
- Com as matrizes termo-documento calcula a distância entre os documentos para cada seção usando a distância do cosseno. Consolida as distâncias em valores de distância total usando a equação $distancia\_total = \alpha * distancia\_indicacao + (1 - \alpha) * distancia\_composicao$, onde α é um parâmetro que controla a importância dada à cada seção.
- Calcula o valor de similaridade entre documentos como $similaridade = 1–distancia\_total$ e constrói o grafo de similaridade utilizando os objetos bula como nós e a similaridade calculada entre eles como arestas.
- Realiza buscas no grafo de similaridade através do nome do medicamento retornando as seções de texto extraídas para o objeto com aquele nome e seu conjunto de similares ordenado decrescentemente pelo valor de similaridade.

5.4 Programa Principal

O programa quando executado apresenta um menu simples oferecendo a busca por medicamentos por nome ou a impressão do grafo de similaridade. Esse grafo pode ser visualizado no site `https://dreampuf.github.io/GraphvizOnline/`

6 Considerações Finais

6.1 Sobre o método proposto

O método é capaz de realizar a extração de informação dos arquivos de bula em formato .txt e usá-la para encontrar medicamentos similares baseado apenas na semântica do texto extraído de forma satisfatória.

Acreditamos que a técnica possa ser expandida para considerar outras seções da bula, podendo gerar classificação de medicamentos, comparações de contraindicações e efeitos colaterais, verificação de interações medicamentosas, entre outros, com relativa facilidade.

6.2 Sobre os resultados

A seção de Composição não é tão padronizada como as outras parte do texto, o que dificultou sua extração. Se refinarmos o método de extração para esta seção os resultados ficarão ainda melhores.

O download e conversão das bulas pode ser automatizado dentro do próprio programa mas, por limitações de tempo, não pôde ser implementado.

Os resultados se mostraram promissores e é de certa forma surpreendente não conseguirmos encontrar muitas aplicações que explorem técnicas semelhantes.

6.3 Divisão do trabalho

Este trabalho foi realizado colaborativamente entre os autores, incluindo a idealização do projeto, a elaboração dos objetivos e a metodologia. Houve divisão de tarefas em algumas partes do projeto, em que Juliane ficou responsável pela revisão de outros trabalhos e Wedeueis realizou a implementação do algoritmo para a obtenção dos resultados.

7 Referências

ANVISA. Resolução-rdc nº 47, de 8 de setembro de 2009, 2009.

NOGUEIRA, T. C. Mineração de texto em bulas de medicamentos. 2014. Trabalho de conclusão de curso (Bacharelado em Sistema de Informação), Universidade Federal dos Vales do Jequitinhonha e Mucuri, Minas Gerais.

SILVA, J. V. F.; SILLA, C. N.; KASHIWABARA, A. Y. Adicionando informações estruturadas ao Bulário Eletrônico da ANVISA, 2015. In: XI Brazialian Symposium on Information System, Goiânia, Goiás, Brasil, 2015, 517-24.

SILVA, J. V. F. Fácil Bula: Sistema que estrutura o bulário eletrônico da Anvisa. 2016. Dissertação de mestrado (Programa de Pós-Graduação em Informática), Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná.

JURAFKSKY, D.; MARTIN, J. H. (2009) Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2ª ed., 2009, Prentice-Hall, Inc.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. (2009) Introduction to Algorithms, Third Edition, 3ª ed., The MIT Press