Business Intelligence

PUC RIO

Mônica Soares Brandão

Correspondência de registros médicos a ensaios clínicos para elegibilidade de pacientes usando NLP

Monografia de Final de Curso

06/12/2020

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC/Rio como parte dos requisitos para a obtenção do título de Especialização em Business Intelligence.

Orientadores:

Prof. Leonardo Alfredo Forero Mendoza

DEDICATÓRIA

À minha família, minha mãe Dionísia e irmão Marcelo, que me apoiaram, nesse processo de estudos me dando suporte e incentivando a cada dia.

Ao meu pai, Jair Enéas, in memoriam, que foi um grande engenheiro, numa época difícil e venceu na vida com esforço próprio e que me fez seguir seu exemplo de dedicação.

Aos meus amigos e irmãos da Comunidade Católica Shalom que rezaram por mim.

AGRADECIMENTOS

Agradeço primeiramente a Deus que tudo criou e a tudo ama.

À minha família que me apoiou nesse processo de estudos me dando suporte e incentivando a cada dia.

Ao professor Leonardo Mendoza sempre disponível para orientar me passando muita confiança.

Ao professor Cristian Muñoz que gentilmente ajudou em um dos processos desse trabalho.

A todos os coordenadores, professores, monitores e toda equipe de apoio do curso de BI-Master que mesmo em meio a pandemia se desdobraram em dar o melhor para os alunos.

Agradeço também a Harvard Medical School que gentilmente cedeu seu Data Set para este trabalho.

Algumas pessoas acham que foco significa dizer sim para a coisa em que você vai se focar. Mas não é nada disso. Significa dizer não às centenas de outras boas ideias que existem.

Steve Jobs

RESUMO

Um ensaio clínico ou do inglês, Clinical Trials, é um tipo de estudo científico utilizado em medicina, psicologia e outras ciências. Trata-se do procedimento utilizado para testar a eficácia de uma dada abordagem terapêutica em uma população de pacientes, ou para coletar informações sobre efeitos secundários de um dado tratamento.

Uma das fases de um processo de ensaios clínicos é o recrutamento de pacientes, onde eles serão expostos aos testes propostos pelo ensaio em questão.

Cada ensaio clínico possui critérios de elegibilidade ou restrições que pode incluir ou excluir pacientes, tornando os elegíveis ou não para o estudo em questão.

Um bom recrutamento de pacientes é crucial para o bom andamento do ensaio clínico. Mas também é um processo árduo para os recrutadores tentar encontrar pacientes que se enquadram nesses critérios.

O Processamento de linguagem natural é a tecnologia usada para ajudar os computadores a entender a linguagem natural do ser humano.

Não é uma tarefa fácil ensinar as máquinas a entender como nos comunicamos.

Processamento de linguagem natural, geralmente abreviado como NLP, do Inglês Natural Language Processing, é um ramo da inteligência artificial que lida com a interação entre computadores e humanos usando a linguagem natural.

O objetivo final da NLP é ler, decifrar, compreender e dar sentido às linguagens humanas de uma maneira que seja valiosa.

A maioria das técnicas de NLP depende do aprendizado de máquina para derivar significado das linguagens humanas.

O objetivo desse trabalho a partir do Processamento de linguagem natural (NLP) correlacionar registros médicos de pacientes a ensaios clínicos.

ABSTRACT

A clinical trial is a type of scientific study used in medicine, psychology and other sciences. This is the procedure used to test the effectiveness of a given therapeutic approach in a patient population, or to collect information about side effects of a given treatment.

One of the phases of a clinical trial process is the recruitment of patients, where they will be exposed to the tests proposed by the trial.

Each clinical trial has eligibility criteria or restrictions that can include or exclude patients, making them eligible or not for the study.

An efficient patient recruitment is crucial to the smooth running of the clinical trial. But it is also an arduous process for recruiters try to find patients who meet these criteria.

Natural language processing is the technology used to help computers understand the natural language of human beings.

It is not an easy task to teach machines to understand how we communicate.

Natural language processing, usually abbreviated as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using Natural Language.

The ultimate goal of NLP is to read, decipher, understand and make sense of human languages in a way that is valuable.

Most NLP techniques rely on machine learning to derive meaning from human languages.

This work is about, based on Natural Language Processing (NLP), correlate patient medical records with clinical trials.

SUMÁRIO

1. INTRODUÇÃO

A medicina baseada em evidências depende fortemente de ensaios clínicos para traduzir a pesquisa na prática clínica [1][2]. Por isso, dos vários tipos de estudos experimentais, o de uso mais frequente, uma vez que proporciona evidências mais fortes, é o ensaio clínico randomizado (RCTs - Randomised Clinical Trial).

Diferente dos estudos observacionais em que o pesquisador não interfere na exposição, nesse estudo o pesquisador planeja e intervém ativamente nos fatores que influenciam os indivíduos da amostra.

É desafiador para o processo de ensaios clínicos as boas práticas de recrutamento de paciente. Encontrar pacientes que se encaixam nos critérios de elegibilidade tem sido para os recrutadores um processo manual, árduo e demorado conforme o fluxo da Figura 1.

Cada ensaio clínico possui um protocolo no qual descreve claramente quem está elegível para participar do estudo [3].

Neste trabalho nos concentramos em ensaios clínicos que estão em processo de recrutamento.

Figura 1- Processo de recrutamento para ensaios clínicos

Processamento de Linguagem Natural (NLP) é uma área da ciência da computação e inteligência artificial preocupada com as interações entre computadores e linguagens humanas (naturais), em particular como programar computadores para processar com sucesso grandes quantidades de dados de linguagem natural (texto e fala) [4].

As tecnologias baseadas em NLP, Figura2, estão se tornando cada vez mais difundidas. Por exemplo, telefones e computadores portáteis suportam texto previsível e reconhecimento de escrita; mecanismos de pesquisa na web dão acesso a informações encerradas em texto não estruturado; tradutores nos permite recuperar textos escritos em chinês e lê-los em espanhol.

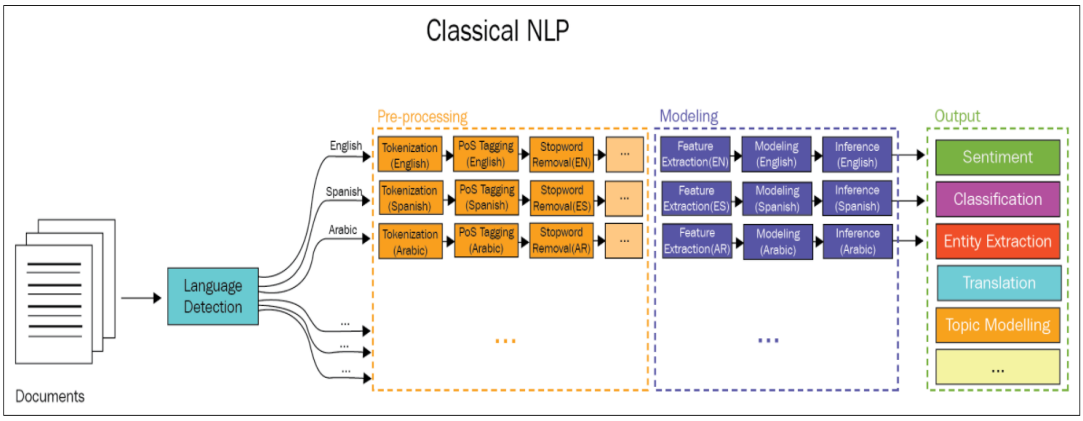


Figura 2- abordagem da NLP clássica

O uso da NLP permite fornecer interfaces homem-máquina mais naturais e acesso sofisticado a informação armazenada, o processamento linguístico passou a desempenhar um papel central na sociedade da informação multilíngue. Também permite relacionar palavras e documentos agrupando em clusters (Figura 3) que é a principal abordagem usadas neste trabalho.

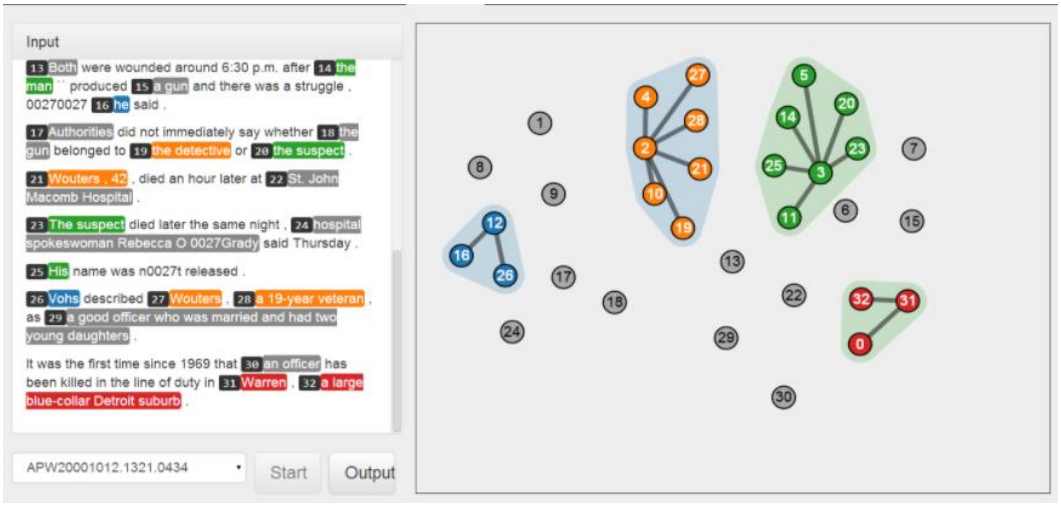


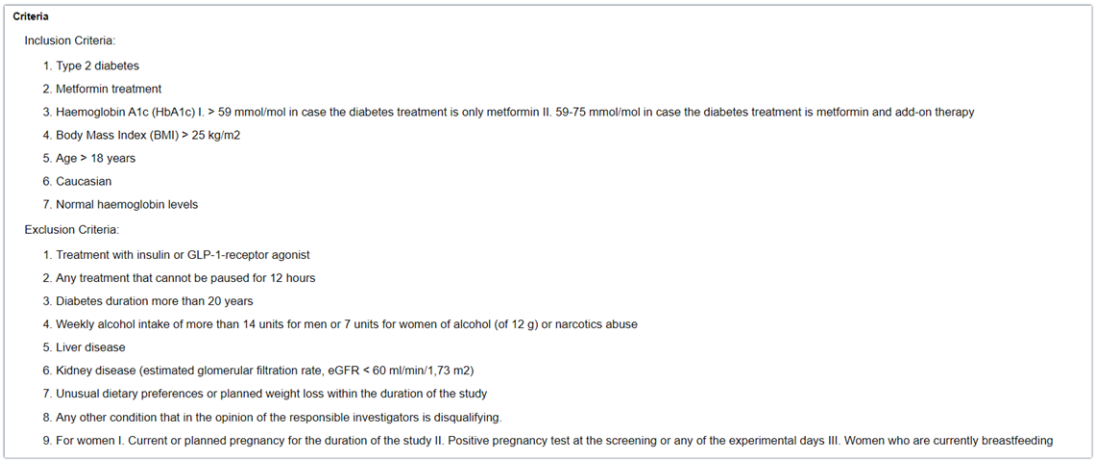
Figura 3- Clusterização de palavras

1.1. MOTIVAÇÃO

O uso da Inteligência Artificial para identificar registros de pacientes (patient records) que são mais próximos de ensaios clínicos em processo de recrutamento ajudaria muito aos recrutadores e pesquisadores a acelerar o processo de pesquisa clínica. Assim como aos pacientes terem uma chance de cura de suas doenças.

Isso gera um enorme gasto de tempo quando falamos em milhares de registro de pacientes no mundo inteiro que precisam ser analisados para verificar a elegibilidade do paciente para o ensaio clínico e assim recrutá-lo.

Por exemplo o ensaio clínico sobre Diabetes Mellitus [11], no site ClinicalTrials.gov, tem 7 critérios de inclusão e 9 critérios de exclusão (Figura 4).



Criteria

Inclusion Criteria:

1. Type 2 diabetes
2. Metformin treatment
3. Haemoglobin A1c (HbA1c) I. > 59 mmol/mol in case the diabetes treatment is only metformin II. 59-75 mmol/mol in case the diabetes treatment is metformin and add-on therapy
4. Body Mass Index (BMI) > 25 kg/m2
5. Age > 18 years
6. Caucasian
7. Normal haemoglobin levels

Exclusion Criteria:

1. Treatment with insulin or GLP-1-receptor agonist
2. Any treatment that cannot be paused for 12 hours
3. Diabetes duration more than 20 years
4. Weekly alcohol intake of more than 14 units for men or 7 units for women of alcohol (of 12 g) or narcotics abuse
5. Liver disease
6. Kidney disease (estimated glomerular filtration rate, eGFR < 60 ml/min/1,73 m2)
7. Unusual dietary preferences or planned weight loss within the duration of the study
8. Any other condition that in the opinion of the responsible investigators is disqualifying.
9. For women I. Current or planned pregnancy for the duration of the study II. Positive pregnancy test at the screening or any of the experimental days III. Women who are currently breastfeeding

Figura 4 - Critérios de inclusão e exclusão

Seria de grande auxílio um sistema que consiga sugerir pacientes compatíveis pode ajudar muito neste processo de recrutamento.

Algumas empresas estão investindo em Inteligência Artificial para ajudar nesse processo de ensaio clínicos, como é o caso da IBM que criou o Watson for Clinical Trial Matching [10], específico para a área oncológica, onde se concentra a maior gama de ensaios clínicos.

Por isso a importância de mais estudos nesta área, para ajudar nas pesquisas cientificas e a pacientes a encontrar sua cura.

1.2. OBJETIVOS DO TRABALHO

Este trabalho foi desenvolvido com o propósito de demonstrar que é possível otimizar e agilizar o processo de recrutamento de ensaios clínicos, sugerindo pacientes elegíveis, através da Inteligência Artificial, elencando algumas técnicas que podem atender a este propósito.

Hoje muitos problemas tem sido solucionados usando aprendizado de máquina (Machine Learning -ML) e Redes Neurais (Neural Network) que ajudam a minerar dados de uma forma rápida e eficiente usando também modelos estatísticos que podem ter acurácias altas sendo bastante assertivos.

Para esse trabalho trabalharemos com 288 registros de pacientes do NLP Research Data Set – N2C2, gentilmente cedidos pela Hardware Medical School e com apenas um ensaio clínico relacionado a Diabetes Mellitus, escolhido do site ClinicalTrail.gov.

Os registros de pacientes por serem textos não estruturados, de texto livre escritos de forma narrativa, com terminologias médicas ou ontologias percebemos que o uso do Processamento de Linguagem Natural ou do Inglês, Natural Language Processing (NLP), seria eficaz, concedendo mais rapidez e acurácia.

Esboçaremos nas próximas páginas, técnicas de NLP existentes, mostrando a abordagem de cada uma assim como os resultados que cada uma alcança.

O modelo de aprendizado que mais se encaixa nesse objetivo é o aprendizado não supervisionado. Diferente do aprendizado supervisionado que de maneira simples descrevemos como, aquele que a partir de um conjunto de dados rotulados, previamente definidos, deseja-se encontrar uma função que seja capaz de predizer rótulos desconhecidos, já o aprendizado não supervisionado por não ter um conjunto de dados previamente rotulado, tem como objetivo descobrir similaridades entre os objetos.

Usaremos aqui as seguintes técnicas:

- TF-IDF
- Similaridade do Cosseno
- LDA
- Doc2Vec

1.3. DESCRIÇÃO DO TRABALHO

Este trabalho foi dividido em n capítulos:

O capítulo 2 apresenta a descrição do problema que está sendo apresentado neste trabalho.

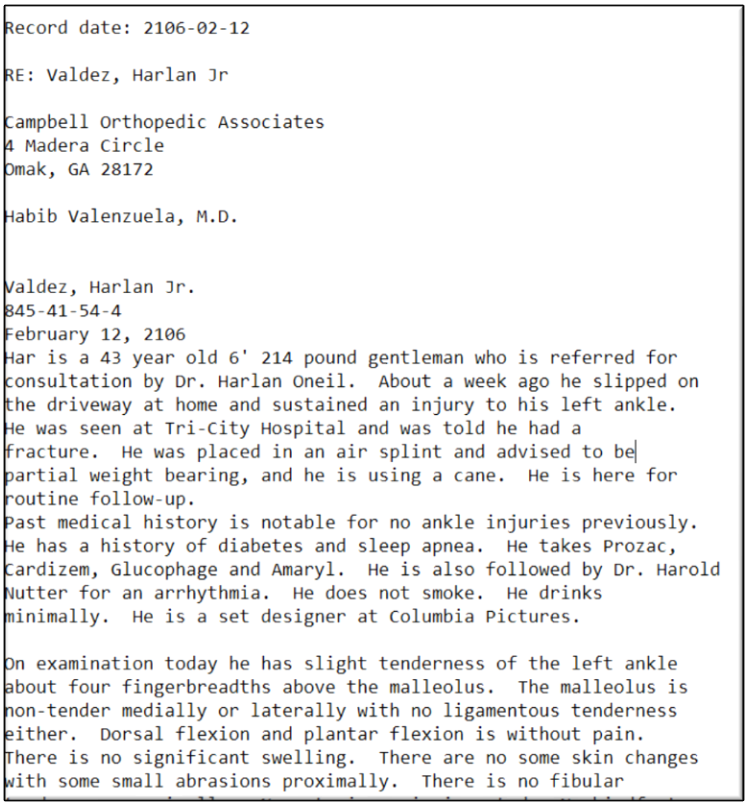
O capítulo 3 apresenta as metodologias aplicadas, descrevendo cada uma e mostrando os resultados alcançados.

O capítulo 4 descreve as conclusões do trabalho e identifica possíveis trabalhos futuros.

2. DESCRIÇÃO DO PROBLEMA

Encontrar pacientes que são elegíveis a partir de milhares de registros médicos em diferentes hospitais parece uma tarefa impossível. Os portuários médicos são de texto livre geralmente são escritos em estilo narrativo, que pode variar de um profissional ou instituição de saúde para outra (Figura 5).

Com isso foram criadas terminologias médicas e ontologia padronizar e normalizar os termos médicos em prontuários clínicos.



```
Record date: 2106-02-12

RE: Valdez, Harlan Jr

Campbell Orthopedic Associates
4 Madera Circle
Omak, GA 28172

Habib Valenzuela, M.D.


Valdez, Harlan Jr.
845-41-54-4
February 12, 2106
Har is a 43 year old 6' 214 pound gentleman who is referred for
consultation by Dr. Harlan Oneil.  About a week ago he slipped on
the driveway at home and sustained an injury to his left ankle.
He was seen at Tri-City Hospital and was told he had a
fracture.  He was placed in an air splint and advised to be
partial weight bearing, and he is using a cane.  He is here for
routine follow-up.
Past medical history is notable for no ankle injuries previously.
He has a history of diabetes and sleep apnea.  He takes Prozac,
Cardizem, Glucophage and Amaryl.  He is also followed by Dr. Harold
Nutter for an arrhythmia.  He does not smoke.  He drinks
minimally.  He is a set designer at Columbia Pictures.

On examination today he has slight tenderness of the left ankle
about four fingerbreadths above the malleolus.  The malleolus is
non-tender medially or laterally with no ligamentous tenderness
either.  Dorsal flexion and plantar flexion is without pain.
There is no significant swelling.  There are no some skin changes
with some small abrasions proximally.  There is no fibular
```

Figura 5 - Exemplo de prontuário de paciente do Harward Medical School Data Set

O ensaio clínico escolhido em clinicaltrials.gov foi o NCT03986073, relacionado a Diabetes Mellitus tipo 2 [5].

Todos os 288 registros de pacientes cedidos pelo Harward Medical School N2C2 NLP Research Data Sets cedidos foram analisados neste trabalho e estavam no formato XML.

Então a primeira decisão foi converter o ensaio clínico NCT03986073 de HTML para o mesmo formato XML para facilitar o trabalho.

A ideia neste trabalho é, a partir de um ensaio clínico previamente escolhido em ClinicalTrial.org, usar algumas técnicas de NLP para sugerir qual os registros de pacientes do Harward Medical School N2C2 NLP Research Data Sets, que são notas não estruturadas do repositório de dados de pacientes de pesquisa da Partners Healthcare[4], mais se correlacionam com o estudo clínico e a nível de serem usados como sugestão de elegibilidade para recrutamento para o ensaio clinico em questão.

Em primeira análise por não haver a definição de uma varável preditora estamos diante de um problema de modelo não supervisionado e este será o caminho adotado neste trabalho.

Além disso por ser um problema de origem textual, o Processamento de Linguagem Natural (NLP) se aplica perfeitamente. Em linhas gerais a NLP é uma interseção de vários campos: Ciência da Computação, Inteligência Artificial e Linguística e permite aos computadores analisar e compreender a linguagem humana.

A linguagem utilizada foi Phyton e a plataforma o Google Colaboratory.

3. METODOLOGIAS

Por ser um modelo não supervisionado é possível neste trabalho usar quatro abordagens:

- Clusterização T-SNE por vizinhança
- Cosine Similarity (Similiaridade do Cosseno)
- Treinamento Doc2Vec
- Tópicos LDA

Também serão usadas as representações Term Frequency-Inverse Document Frequency (Tf-Idf) e bag of words.

A finalidade de explorar as diversas abordagens é mostrar suas diferenças e as aplicabilidades de cada uma no processo de elegibilidade de pacientes para um ensaio clínico.

3.1. TRATAMENTO INICIAL DOS DADOS

Os registros de paciente são arquivos no formato XML e o ensaio clínico um arquivo web que foi convertido para um arquivo no mesmo formato.

Todos os arquivos foram importados e organizados em data frame. Alguns dados como nome do arquivo e tipos de documento, se é um registro de paciente (PR – Patient Record) ou ensaios clínicos (CL – Clinical Trial), foram separados em colunas para facilitar a identificação dos arquivos (Figura 6).

	description	file_name	doc_type	patient_name
0	xml record date care cen...	101.xml	PR	Russell Donna
1	xml record date dr ...	105.xml	PR	TUTTLE Sandy
2	xml record date g obrya...	124.xml	PR	Edwin Workman
3	xml record date washingt...	146.xml	PR	Francis Lydia
4	xml record date you...	161.xml	PR	Jaquante Xue

Figura 6 - Organização dos dados

3.2. PRÉ-PROCESSAMENTO DE TEXTO

Os registros médicos são textos narrativos e por isso foi necessária uma limpeza para remover palavras ou símbolos que não são relevantes para o modelo, assim como letras maiúsculas para evitar qualquer sensitive-case, pontuação, números e as stopwords que veremos mais adiante. Por isso para todos os registros médicos foi feita a limpeza do texto.

3.2.1. TOKENIZAÇÃO

Faz parte do pré-processamento e tem como finalidade seccionar um documento textual em unidades mínimas, que exprimam a mesma semântica original do texto. O termo token (Figura 7) é utilizado para designar estas unidades, que em muitas vezes correspondem a somente uma palavra do texto.

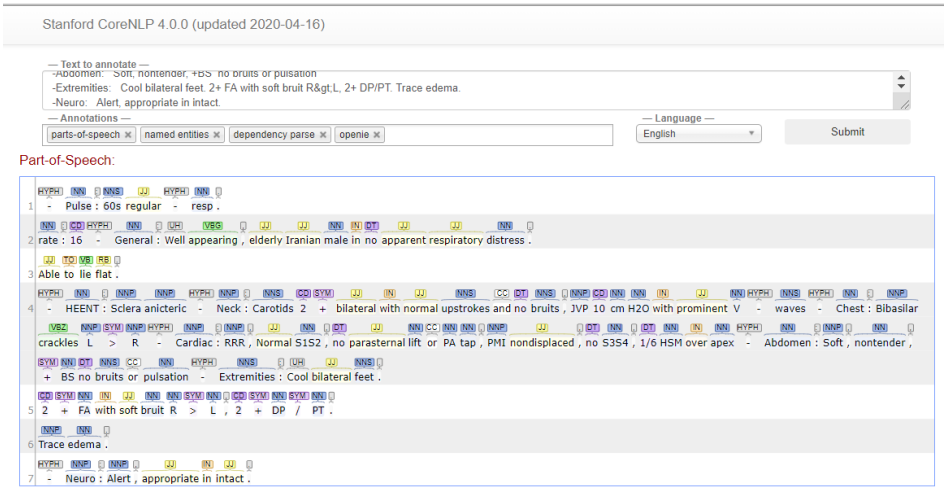


Figura 7 – saída do site corenlp.run

Após a limpeza do texto foi feita a tokenização, separando as palavras e gerando uma lista de palavras.

3.2.2. STOP WORDS

Aplicar Stop words seria tirar termos comuns ou tudo que pode ser desconsiderado do texto, sem mudar o significado dele. No caso deste trabalho precisamos aplicar as stop words para retirar as palavras sem relevância do nosso “corpus” (bag of words) (Figura 8).

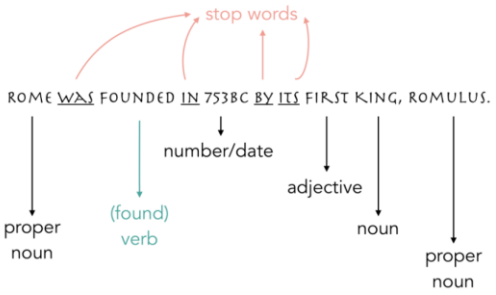


Figura 8 - exemplo da aplicação de Stopwords

3.2.3. WORD CLOUD

A Word Cloud é uma ferramenta que cria a nuvem de palavras a partir de um texto fornecido pelo usuário. Essas nuvens dão maior destaque às palavras que aparecem com mais frequência no texto de origem. Com a lista de palavras limpa geramos uma word cloud para analisar a frequência das palavras. Assim pode-se analisar as palavras mais repetidas em todos os registros de paciente conforme Figura 9 abaixo.

Figura 9 - Word cloud dos registros médicos

3.3. GERAR CORPUS

Mas, antes de aprofundarmos, vamos entender algumas terminologias da NLP.

- Um 'token' normalmente significa uma 'palavra'.
- Um 'documento' normalmente pode se referir a uma 'frase' ou 'parágrafo'.
- Um 'corpus' é normalmente uma 'coleção de documentos como um pacote de palavras'.

Ou seja, para cada documento, um corpus contém o id de cada palavra e sua contagem de frequência nesse documento.

O pacote que usaremos para o NLP é o Gensim. Para trabalhar em documentos de texto, o Gensim exige que as palavras (também conhecidas como tokens) sejam convertidas em ids únicos. Para conseguir isso, Gensim permite criar um objeto Dicionário que mapeia cada palavra para um id único. Os objetos de dicionário são normalmente usados para criar um Corpus de "bag of words". O dicionário contém um mapa de todas as palavras (tokens) para seu id único. É este Dicionário e o "bag of words" (Corpus) que são usados como entradas para a modelagem de tópicos e outros modelos nos quais a Gensim é especializada [7].

Com isso foi criada uma matriz de termos de documento vectorizado retirando as stop words, configurando o CountVectorizer convertendo as palavras em minúsculas.

4. ARQUITETURA DO SISTEMA PROPOSTO

O objetivo deste trabalho é demostrar as diversas técnicas de NLP possíveis para sugerir pacientes possivelmente elegíveis para ao ensaio clínico escolhido.

Neste capítulo explicitaremos as técnicas utilizadas e no capítulo 5 os resultados obtidos para cada uma delas.

4.1. TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

O peso TF-IDF é um peso frequentemente usado na recuperação de informações e mineração de texto. Esse peso é uma medida estatística usada para avaliar a importância de uma palavra para um documento em uma coleção ou corpus. A importância aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento, mas é compensada pela frequência da palavra no corpus[8].

Tf-IDF é calculado multiplicando um componente local como frequência de termo (TF) com um componente global, isto é, frequência de documento inversa (IDF) e opcionalmente normalizando o resultado para comprimento de unidade.

Como resultado disso, as palavras que ocorrem com frequência nos documentos serão reduzidas [8].

Usamos o TF-IDF para comparar o arquivo do ensaio clínico contra os 288 registros de pacientes para tentar achar os registros mais similares ao ensaio clínico. Sendo cada registro de um único paciente, achando os registros mais similares ao ensaio clínicos podemos sugerir que esses pacientes têm grande chance de elegibilidade.

4.2. COSINE SIMILARITY

Similaridade do Cosseno (Cosine Similarity) é uma métrica usada para determinar o quão semelhantes os documentos são, independentemente de seu tamanho.

Matematicamente, ele mede o cosseno do ângulo entre dois vetores projetados diferentes de zero em um espaço multidimensional que mede o cosseno do ângulo entre eles. O cosseno de 0 ° é 1 e é menor que 1 para qualquer ângulo no intervalo $(0, \pi]$ radianos.

Nesse contexto, os dois vetores são matrizes que contêm a contagem de palavras de dois documentos.

Quando plotado em um espaço multidimensional, onde cada dimensão corresponde a uma palavra no documento, a similaridade do cosseno captura a orientação, o ângulo, dos documentos e não a magnitude (para a magnitude deve se calcular a distância euclidiana).

A semelhança de cosseno é vantajosa porque mesmo que os dois documentos semelhantes estejam distantes pela distância euclidiana por causa do tamanho, eles ainda poderiam ter um ângulo menor entre eles. Quanto menor o ângulo, maior a similaridade. Conforme o exemplo da figura 10, a palavra “críquete” pode aparecer 50 vezes em um documento e 10 vezes em outro e eles ainda serem similares.

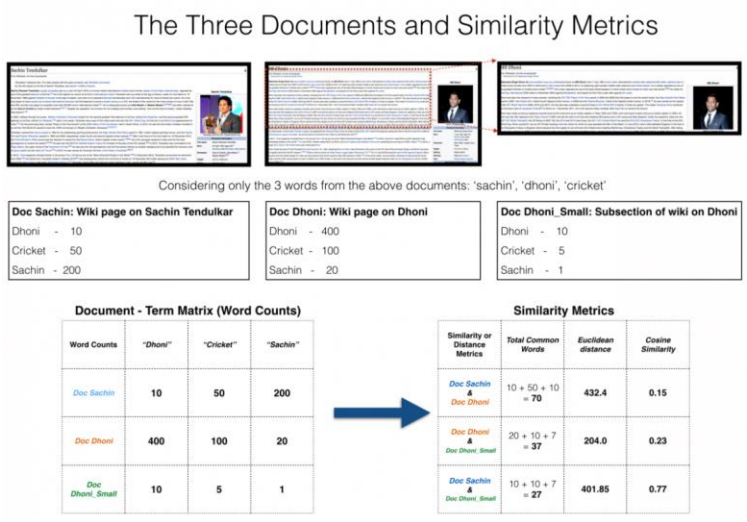


Figura 10- Exemplo de Similaridade do Cosseno, onde três documentos estão conectados por um tema comum - o jogo de críquete.

4.3. LDA - LATENT DIRICHLET ALLOCATION

No processamento de linguagem natural, a Alocação Latente de Dirichlet do Inglês Latent Dirichlet Allocation (LDA) é um modelo que assume que os documentos nada mais são do que uma distribuição de probabilidade de tópicos e os tópicos nada mais são do que uma distribuição de probabilidade de palavras, o LDA calcula a probabilidade de que um documento seja principalmente este tópico ou aquele tópico (por exemplo, documento N é 77% tópico 1 , 10% tópico 2, 8% tópico 5 e 5% tópico 7) com base nas palavras que contém [6].

LDAvis são ferramentas para criar uma visualização interativa usando página web, de um modelo de tópico que foi ajustado a um corpus de dados de texto usando a Alocação de Dirichlet Latente (LDA). Dado os parâmetros estimados do modelo de tópico, ele calcula várias estatísticas de resumo como entrada para uma visualização interativa criada com o D3.js que é acessada por meio de um navegador. O objetivo é ajudar os usuários a interpretar os tópicos em seu modelo de tópico LDA.

O algoritmo de modelo de tópico LDA requer uma matriz de palavras do documento como entrada principal que pode ser criada anteriormente no CountVectorizer.

Após usar LDA o resultado ficou muito bastante pulverizado a nível de tópicos e não pareceu um modelo adequado (Figura11).

Figura11 - Primeiro modelo com LDA

Foi usado o GridSearch para encontrar o melhor modelo LDA. O parâmetro de ajuste mais importante para o LDA é o número de tópicos (n_components), como

não sabemos quantos tópicos fornecemos a ele uma lista de diferentes valores supostos que podemos definir como n_compnents.

Outro parâmetro que foi alterado foi a taxa de aprendizado (learning_decay), essa taxa também não tem um valor padrão.

Antes de aplicar o Grid Search foi feito ajustes no vetor de dados (CountVectorizer) alterando os parâmetros token_patern para considerar como tokens apenas palavras alfanuméricas de maior que 5 caracteres.

A Grid Search constrói vários modelos LDA para todas as combinações possíveis de valores de parâmetros informados. Como ele testa cada combinação ele leva um grande tempo processando uma saída.

O resultado do Grid Search aponta os melhores parâmetos para o modelo LDA que no nosso caso ficou em 5 tópicos (Figura 12).

```
Melhores parâmetros:  {'learning_decay': 0.7, 'n_components': 5}
Melhor score de probabilidade logarítmica:  -608551.0734770491
Perplexidade do modelo:  4576.9691309491145
```

Figura 12 - Primeiro modelo com LDA

Também foi feita uma comparação entre os scores de performance dos modelos (Figura13).

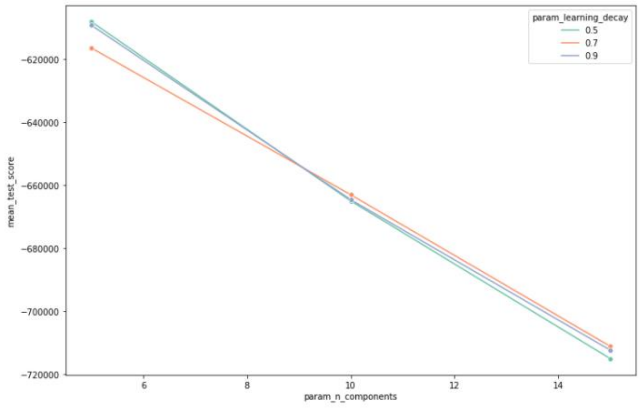


Figura 13 - Análise de sensibilidade do modelo LDA quanto ao número de tópicos utilizados.

Uma abordagem que a LDA nos mostra é que para classificar um documento como pertencente a um tópico específico, é ver qual tópico tem a maior contribuição para esse documento e atribuí-lo.

Podemos ver na figura abaixo, destacado em verde, todos os principais tópicos de um documento e identificarmos o tópico dominante (dominant_topic) (Figura 14).

	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic
Doc0	0.680000	0.000000	0.000000	0.000000	0.320000	1
Doc1	0.000000	0.000000	0.000000	0.000000	1.000000	5
Doc2	0.000000	0.000000	0.900000	0.000000	0.100000	3
Doc3	0.000000	0.730000	0.000000	0.000000	0.270000	2
Doc4	0.680000	0.000000	0.000000	0.000000	0.320000	1
Doc5	0.000000	0.000000	0.000000	1.000000	0.000000	4
Doc6	0.000000	0.000000	0.000000	0.000000	1.000000	5
Doc7	0.000000	0.000000	0.000000	0.000000	1.000000	5
Doc8	1.000000	0.000000	0.000000	0.000000	0.000000	1
Doc9	0.650000	0.000000	0.000000	0.000000	0.350000	1

	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	file_name	patient_name
Doc0	0.68	0.00	0.0	0.0	0.32	1	101.xml	Russell Donna
Doc1	0.00	0.00	0.0	0.0	1.00	5	105.xml	TUTTLE Sandy
Doc2	0.00	0.00	0.9	0.0	0.10	3	124.xml	Edwin Workman
Doc3	0.00	0.73	0.0	0.0	0.27	2	146.xml	Francis Lydia
Doc4	0.68	0.00	0.0	0.0	0.32	1	161.xml	Jaquante Xue
Doc5	0.00	0.00	0.0	1.0	0.00	4	163.xml	Gerald Marlon
Doc6	0.00	0.00	0.0	0.0	1.00	5	183.xml	Roberta Vincen
Doc7	0.00	0.00	0.0	0.0	1.00	5	186.xml	URQUHART WEND
Doc8	1.00	0.00	0.0	0.0	0.00	1	200.xml	HEATH QUINTEN
Doc9	0.65	0.00	0.0	0.0	0.35	1	209.xml	Parrish Chris

Figura 14 - Análise de tópicos dominantes (objeto style x dataframe), correlacionado os documentos aos registros de paciente

Outra análise que pode ser feita é a quantidade de documentos por tópico (Figura 15)

	Topic Num	Num Documents
0	5	79
1	1	65
2	2	58
3	3	45
4	4	42

Figura 15 - Análise de tópicos dominantes

Com esses ajustes vejamos como ficou o modelo ajustado usando o pyLDA (Figura 16).

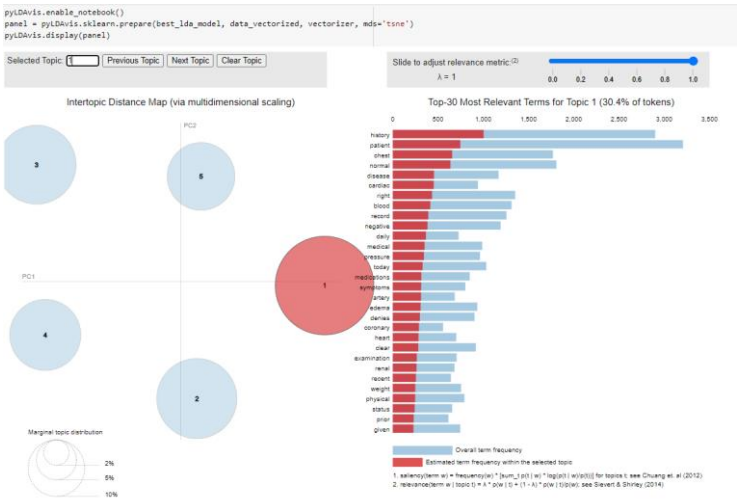


Figura 16 - Modelo ajustado destacando o tópico 1

4.4. DOC2VEC

Uma das técnicas mais eficientes para representar uma palavra é o Word2Vec. Word2vec é um modelo preditivo computacionalmente eficiente para aprender embeddings de palavras a partir de texto bruto. Ele plota as palavras em um espaço vetorial multidimensional, onde palavras semelhantes tendem a estar próximas umas das outras.

As palavras ao redor de uma palavra fornecem o contexto para essa palavra. Doc2Vec é outra técnica amplamente usada que cria a incorporação de um documento independentemente de seu comprimento. Enquanto o Word2Vec calcula um vetor de recurso para cada palavra no corpus, Doc2Vec calcula um vetor de recurso para cada documento no corpus. O modelo Doc2vec é baseado no Word2Vec, com apenas a adição de outro vetor (ID do parágrafo) à entrada [9].

O objetivo do Doc2vec é criar uma representação numérica de um documento, independentemente do seu comprimento.

A técnica que usamos para a saída do modelo foi rankear os documentos mais similares ao ensaio clínico.

Concluímos essa técnica aplicando o TensorBoard Embeddings Projector, Figura 22, que é uma ótima ferramenta para analisar seus dados e ver os valores incorporados entre si. O painel permite a pesquisa de termos específicos e destaca palavras que estão próximas no espaço de incorporação. O T-SNE foi o método utilizado por ser útil para explorar vizinhos e encontrar clusters.

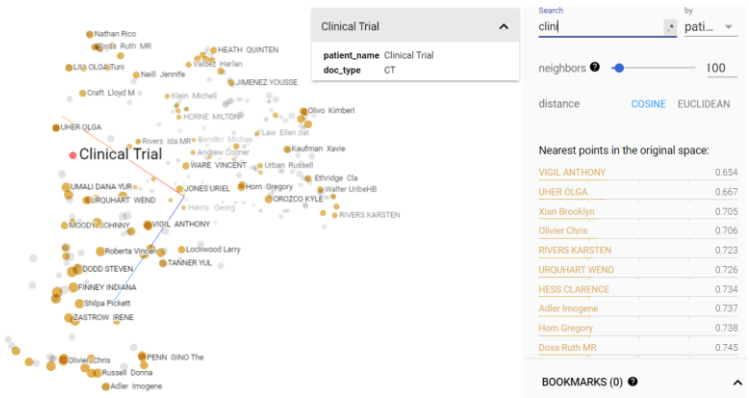


Figura 22 – Tensor Board baseado em T-SNE

5. RESULTADOS

Neste capítulo será mostrado os resultados obtidos pelas diversas metodologia ou técnicas aplicadas sobre a base de dados NLP Research Data Set – N2C2.

Verificamos que a aplicação em TF-IDF, Figura 17, obteve um resultado similar a aplicação da Similaridade do Cosseno, Figura 18, apesar de terem abordagens diferentes, já que o TF-IDF mede a frequência de termo ao invés da distância entre vetores que contêm a contagem de palavras, como é o caso da Similaridade do Cosseno.

```
[(0.12029476556175747, '          283.xml', ' HOLCOMB DENNIS'),
 (0.1186357162264522, '          365.xml', ' RIVERS KARSTEN'),
 (0.11134372047497697, '          323.xml', ' Xian  Brooklyn'),
 (0.10858972151963334, '          202.xml', ' Jesus  Jarome '),
 (0.10808094288418249, '          157.xml', ' Jonathan Oswal'),
 (0.10719611870989881, '          125.xml', ' Fair  Bill MR '),
 (0.10521948535304923, '          137.xml', ' MOODY  JOHNNY '),
 (0.10289014479485269, '          320.xml', ' Iles  Louise M')]
```

Figura 17 - Resultado dos 10 registros mais similares aplicando TF-IDF

```
[(0.16858176840254288, '          365.xml', ' RIVERS KARSTEN'),
 (0.1542013235472454, '          125.xml', ' Fair  Bill MR '),
 (0.1538150590990304, '          202.xml', ' Jesus  Jarome '),
 (0.15301270968413172, '          283.xml', ' HOLCOMB DENNIS'),
 (0.15274496001150814, '          132.xml', ' Jorgenson  Viv'),
 (0.1448782937637705, '          323.xml', ' Xian  Brooklyn'),
 (0.14365502380782572, '          320.xml', ' Iles  Louise M'),
 (0.1383773091900867, '          354.xml', ' FAY  BROOK    ')]
```

Figura 18 - Resultado dos 10 registros mais similares aplicando Similaridade do Cosseno

A outra técnica utilizada, o LDA nos mostra a abordagem por tópicos.

Podemos ver que por esse modelo o nosso ensaio clínico foi classificado como tópico 1 (Figura 19).

	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	file_name	patient_name
Doc288	1.0	0.0	0.0	0.0	0.0	1	CT_NCT03986073.xml	Clinical Trial

Figura 19 - Tópico para o Clinical Trial

Verificando que o ensaio clínico foi classificado no tópico 1, buscamos a lista de 10 documentos do tópico 1 que seriam possíveis registros elegíveis (Figura 20).

	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic	file_name	patient_name
Doc266	1.0	0.0	0.0	0.0	0.0	1	CT_NCT03986073.xml	Clinical Trial
Doc21	1.0	0.0	0.0	0.0	0.0	1	399.xml	Nathan Rico
Doc197	1.0	0.0	0.0	0.0	0.0	1	396.xml	AARON JEAN
Doc100	1.0	0.0	0.0	0.0	0.0	1	384.xml	UHRICH KARSON
Doc93	1.0	0.0	0.0	0.0	0.0	1	380.xml	HEATH HANNAH
Doc262	1.0	0.0	0.0	0.0	0.0	1	370.xml	UMALI DANA YUR
Doc200	1.0	0.0	0.0	0.0	0.0	1	359.xml	Hill Owen S
Doc190	1.0	0.0	0.0	0.0	0.0	1	355.xml	DALEY WADE MR
Doc204	1.0	0.0	0.0	0.0	0.0	1	354.xml	FAY BROOK
Doc287	1.0	0.0	0.0	0.0	0.0	1	351.xml	GLENN OLIVIA

Figura 20 – 10 registros do tópico 1 onde encontramos nosso ensaio clínico

Doc2Vec foi outra técnica usada neste trabalho para indicar registros de pacientes similares ao ensaio clínico, podemos ver através do Ranking que alguns documentos foram eleitos mais simulares.

O resultado foi mostrado abaixo com os 5 documentos mais similares pelo modelo Doc2vec (Figura 21).

```
Doc2vec Ranking

Clinical Trial - Document number: 266 - Record Number: CT_NCT03986073.xml -  Patient Name: Clinical Trial

Similiarity of the documents per model using Word2ve Doc2Vec(dm/m,d50,n5,w5,mc2,s0.001,t3):

* MOST SIMILAR        =>  (157, 0.5628523230552673) - file_name : 280.xml  - Patient Name: «ZASTROW  IRENE»

* SECOND-MOST SIMILAR =>  (57, 0.5561850666999817) - file_name : 162.xml  - Patient Name: «Quijano  Bayle»

* THIRD-MOST SIMILAR  =>  (40, 0.5350175499916077) - file_name : 283.xml  - Patient Name: «HOLCOMB DENNIS»

* MEDIAN              =>  (108, 0.27954474091529846) - file_name : 392.xml  - Patient Name: «Adair  HelenMR»

* LEAST SIMILAR       =>  (41, -0.014131680130958557) - file_name : 268.xml  - Patient Name: «Garrison Sexto»
```

Figura 21 - resultado do racking

6. CONCLUSÃO E TRABALHOS FUTUROS

Apesar da eficácia desses modelos depender de uma avaliação de um especialista em recrutamento, Clinical Trials Recruiter, é algo fora do escopo deste trabalho que tem o intuito de mostrar que é possível facilitar o processo de recrutamento usando a Inteligência Artificial e os modelos hoje disponíveis. Cada técnica tem uma abordagem diferente e um dos modelos pode atender melhor ao processo de recrutamento descrito neste trabalho.

Acredito que mais pesquisas podem ser feitas nesta área para que o processo seja cada vez mais assertivo e ajude tanto profissionais a acharem pacientes para suas pesquisas, como principalmente ajudem a pacientes acharem a cura para suas patologias.

Para trabalhos futuros pretende usar uma abordagem diferente a nível de classificação dos ensaios clínicos, ou seja, uma abordagem supervisionada, em LSTM com NLP, para classificar os mesmos que tem critérios de elegibilidade similares.

A ideia seria uma base de dados de ensaios clínicos de uma patologia específica, extraída via webscrapping, com os dados tratados e classificados por elegibilidade.

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. Matching patients to clinical trials using semantically enriched document representation

2. References [1] P.M. Spieth, A.S. Kubasch, A.I. Penzlin, B.M. Illigens, K. Barlinn, T. Siepmann, Randomized controlled trials – a matter of design, Neuropsychiatric Dis. Treat. 12 (2016) 1341–1349
Hamed Hassanzadeha, Sarvnaz Karimib, Anthony Nguyena.

3. C.A. Umscheid, D.J. Margolis, C.E. Grossman, Key concepts of clinical trials: a narrative review, Postgrad. Med. 123 (5) (2011) 194–204.

4. **The majority of these Clinical Natural Language Processing** (NLP) data sets were originally created at a former NIH-funded National Center for Biomedical Computing (NCBC) known as i2b2: Informatics for Integrating Biology and the Bedside. Based at Partners HealthCare System in Boston from 2004 to 2014.

5. **Study of TQ-F3083 Capsules in Subjects With Type 2 Diabetes Mellitus** (https://clinicaltrials.gov/ct2/show/NCT03986073?recrs=a&cond=Diabetes+Mellitus&draw=2&rank=42)

6. https://medium.com/swlh/latent-dirichlet-allocation-lda-eff969bda284

7. https://www.machinelearningplus.com/nlp/gensim-tutorial/#2whatisadictionaryandcorpus

8. http://www.tfidf.com/

9. **NLP: Word Embedding Techniques Demystified**. Rabeh Ayari, PhD

10. https://www.ibm.com/products/clinical-trial-matching-oncology

11. **NCT04228484 - The Insulin Response to the Gut Hormone GIP After Near-normalisation of Plasma Glucose in Patients With Type 2 Diabetes (GA-16)** - https://clinicaltrals.gov/ct2/show/NCT04228484?recrs=a&cond=Diabete+Type+2&draw=3&rank