**

*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*

barra direita

DEDICATÓRIA

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

AGRADECIMENTOS

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

A minha família que me apoiou nesse processo de estudos me dando suporte e incentivando a cada dia.  
Ao professor Leonardo Mendonza sempre disponível para orientar me passando muita confiança.

E a todos os professores e coordenadores do curso de BI-Master.

Agradeço também a Harvard Medical School que gentilmente cedeu seu NLP Research Data Set \* para esse trabalho.

\*https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

*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. Você precisa selecionar cuidadosamente.*

*Steve Jobs*

RESUMO

Um ensaio clínico ou do inglês clinical trial é um tipo de [estudo científico](https://pt.wikipedia.org/wiki/M%C3%A9todo_cient%C3%ADfico) utilizado em [medicina](https://pt.wikipedia.org/wiki/Medicina), [psicologia](https://pt.wikipedia.org/wiki/Psicologia) e outras ciências. Trata-se do procedimento utilizado para testar a eficácia de uma dada abordagem [terapêutica](https://pt.wikipedia.org/wiki/Terap%C3%AAutica) em uma população de pacientes, ou para coletar informações sobre [efeitos secundários](https://pt.wikipedia.org/wiki/Efeito_colateral) de um dado [tratamento](https://pt.wikipedia.org/wiki/Tratamento_(sa%C3%BAde)).

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 objetivo desse trabalho a partir da NLP (Natural Language Processing) 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 eficient 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.

The purpose of this work is to use NLP to correlate patient medical records with clinical trials.

SUMÁRIO

1. Introdução

1.1. Motivação ...............................................

1.2. Objetivos do trabalho.....................................

1.3. Descrição do Trabalho....................................

2. Descrição do Problema..................................................

3. Metodologia..............................................................

3.1. Tratamento inicial dos dados.................................

3.2. Pré-processamento de texto..................................

3.2.1. Tokenização......................................

3.2.2. Stop words..........................................

3.2.3. Word Cloud........................................

3.3. Gerar Corpus.................................................

3.4..Estudo comparativos das técnicas de NLP........................

3.4.1. Term frequency-inverse document frequency ..

3.4.2. Cosine Similarity...................................

3.4.3. LDA - Latent Dirichlet allocation.....................

3.4.4. Doc2Vec.............................................

4. Arquitetura do sistema proposto.........................................

5. Resultados............................................................

6. Conclusão e trabalhos futuros............................................

7. Referências Bibliográficas ..................................

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)\*.

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 (Fig 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.



Fig 1. Processo de recrutamento para ensaios clínicos

\* Randomised Clinical Trials

**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.

Por exemplo o ensaio clínico sobre Diabetes Mellitus \*\* no site ClinicalTrials.gov, tem 7 critérios de inclusão e 9 critérios de exclusão (Fig 1.1). Isso gera um enorme gasto de tempo quando falamos em milhares de pacientes com esta doença no mundo inteiro. Eles podem não ser elegíveis, apesar de terem a doença.

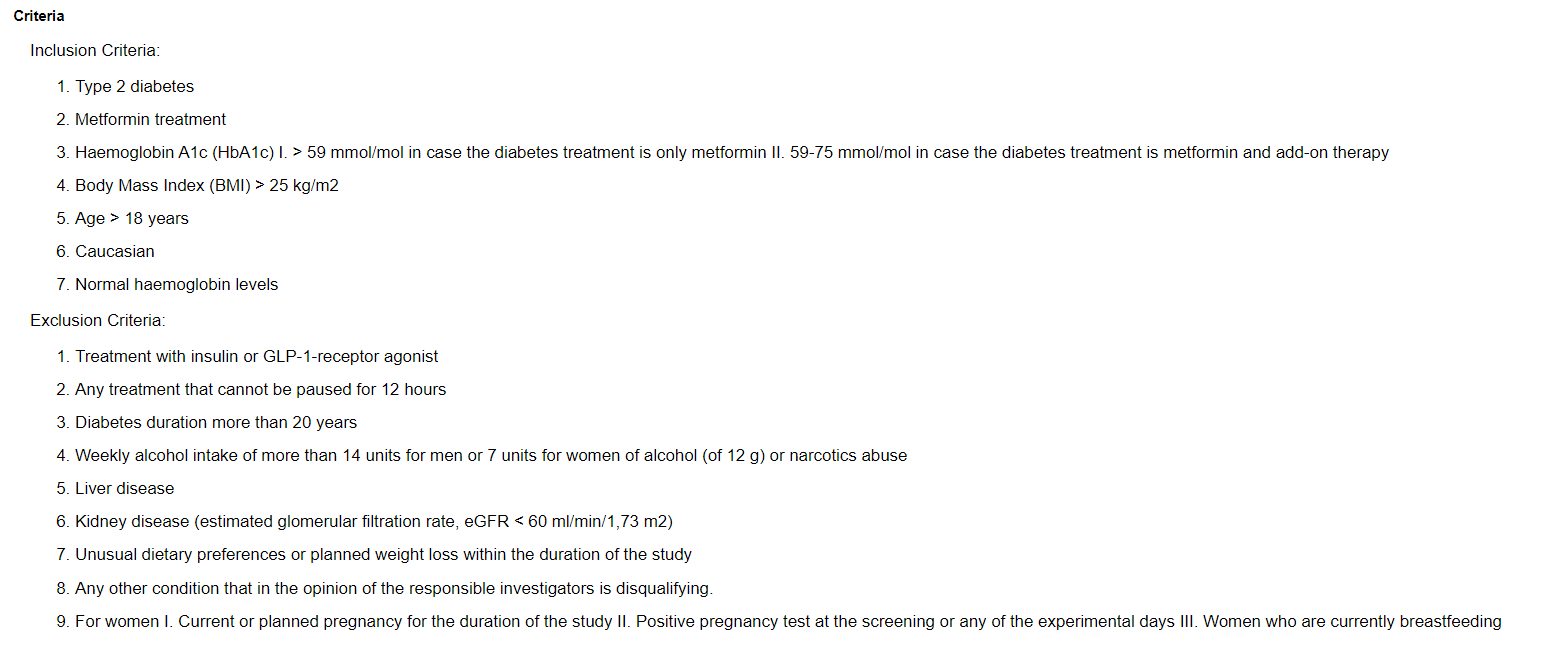


Fig 1.1 – Critérios de inclusão e exclusão

Por isso um sistema que consiga sugerir pacientes compatíveis pode ajudar muito neste processo de recrutamento.

\*\* 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

* 1. **Objetivos do Trabalho**

Com o objetivo de otimizar e agilizar o processo de recrutamento de pacientes para ensaios clínicos foi pensado na Inteligência artificial , que hoje tem solucionado muitos problemas usando aprendizado de máquina (Machine Learning -ML) , 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 gentilmente cedidos pela Hardware Medical School que gentilmente cedeu seu NLP Research Data Set – N2C2 e com apenas um ensaio clínico 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.

* 1. **Descrição do Trabalho**

Este trabalho se desenvolveu em 4 etapas.

Iniciando a primeira etapa com a pesquisa de artigos com o mesmo objetivo, mas com uso de técnicas diferentes. A segunda etapa foi conseguir a base de dados ideal que tivesse um ensaio clínico correlacionado. A base de Dados usada foi cedida após contato pela Harvard Medical School que gentilmente cedeu seu NLP Research Data Set.

Após a primeira análise do Data Set foi feito a terceira etapa que foi pré-processamento dos dados, que será descrita no capítulo 3. A partir do pré-processamentos passamos para a quarta etapa que está na secão 3.4 do capitulo foi um estudo comparativo entre algumas técnicas usadas em NLP. Este trabalho abordou as técnicas de TF-IDF, Similaridade do cosseno, LDA e Doc2Vec.

O capítulo 5 aborda os resultados obtidos neste trabalho, seguindo do capítulo 6, a conclusão do mesmo.

1. **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 (Fig 2).

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

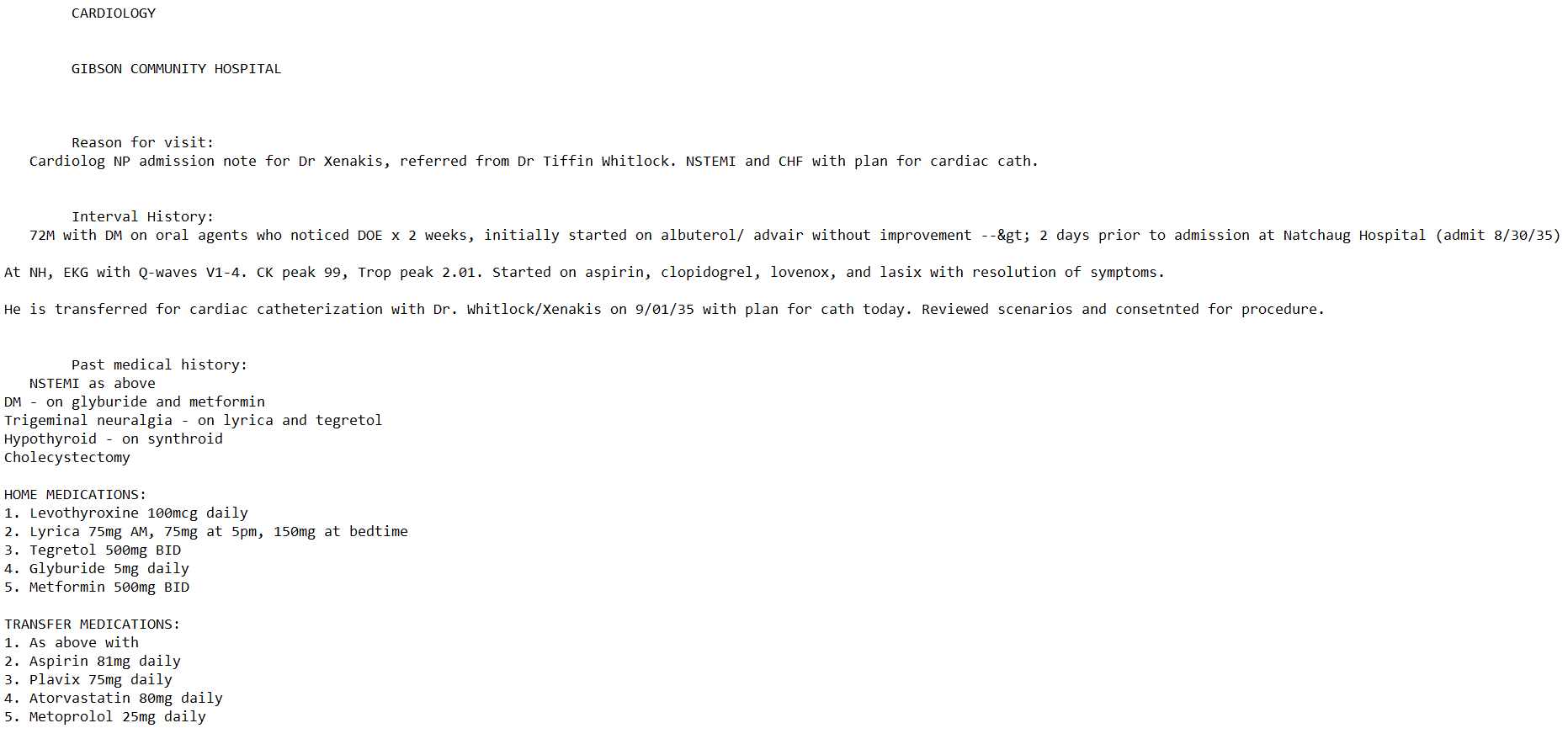


Fig 2. Exemplo de prontuário de paciente do Harward Medical School Data Set

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.

1. **Metodologia**

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]. A NLP contém um vasto conjunto de sintaxe, semântica,etc., formando os níveis linguístico (Fig 3.0).

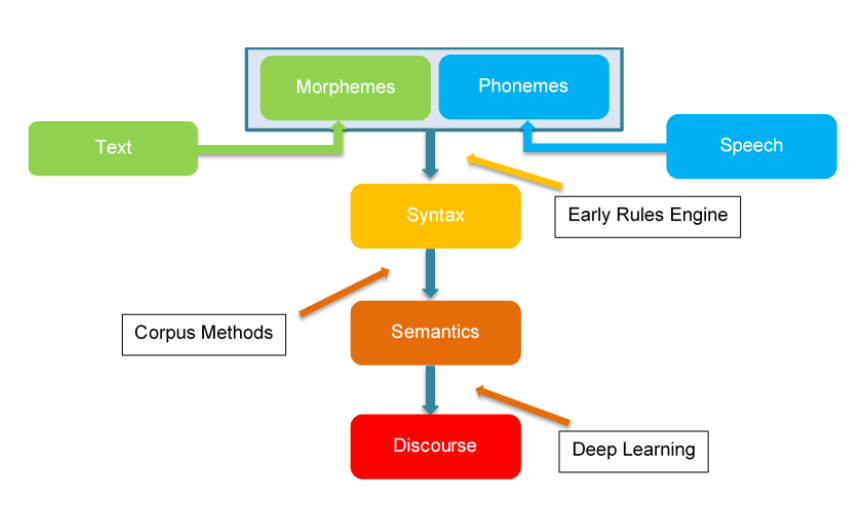


Fig 3– Niveis da NLP

Por ser um modelo não supervisionado é possível neste trabalho usar três abordagens: Clusterização T-SNE, Similiaridade (cosseno) ou treinamento word2vector e tópicos LDA.  
Utilizaremos todas as abordagens com a finalidade de 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 estão em arquivos no formato XML e o ensaio clínico foi convertido para um arquivo no mesmo formato.

Todos os arquivos foram importados e organizados em um dataframe. 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 (Fig 3).

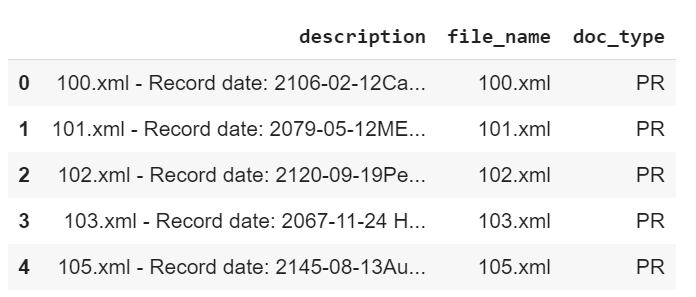


Fig 3.1– Organização dos dados

**3.2. Pré-processamento de texto**

Os registros médicos é um texto narrativo 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 (Fig 3.2.1)* é utilizado para designar estas unidades, que em muitas vezes correspondem a somente uma palavra do texto. [7]

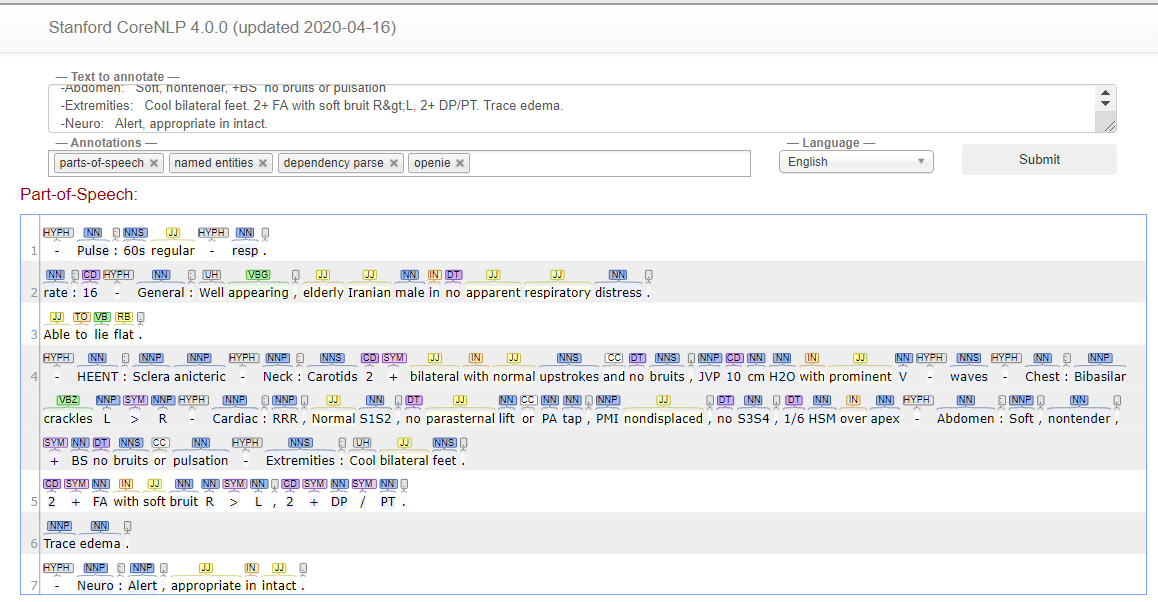


Fig 3.2.1– saída do site corenlp.run

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) (Fig 3.2.2).

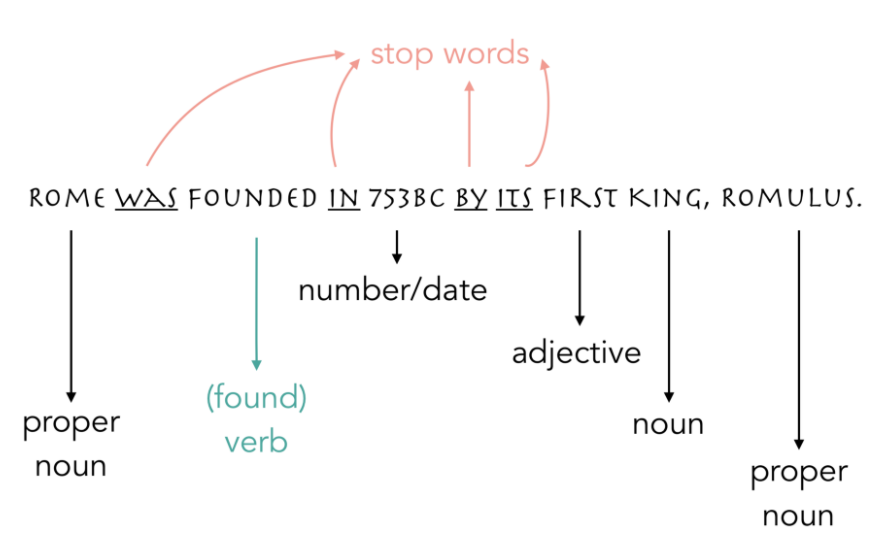


Fig 3.2.2– exemplo da aplicação de Stopwords

**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.



Fig 3.2.3 – 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 [8].

Com isso foi criada uma matriz de termos de documento vectorizado retirando as stop words , configurando o CountVectorizer convertendo as palavras em minúsculas (Fig 3.2.4).

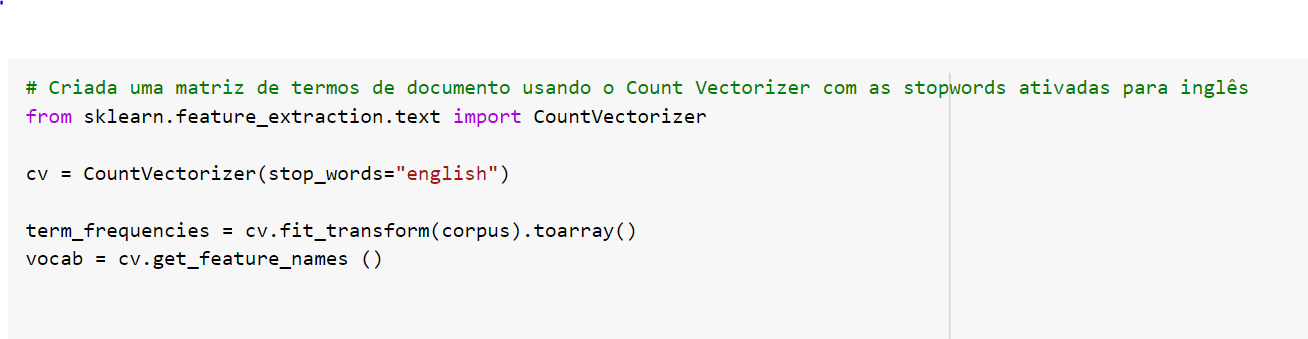


Fig 3.2.4. Transformando em vetores o documento

**3.4. Estudo comparativos das técnicas de NLP**

**3.4.1. 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.[9]

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 (Fig 3.2.5) 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.

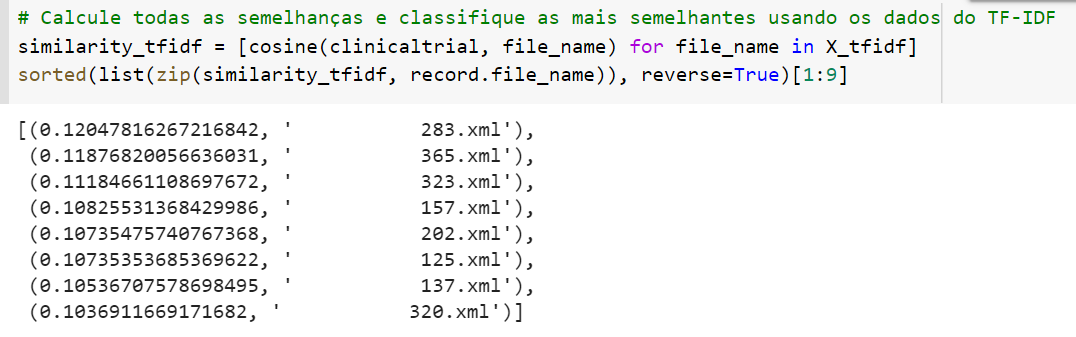


Fig 3.2.5 – Registros mais similares aplicando TF-IDF

**3.4.2. Cosine Similarity**

A similaridade de cosseno (Cosine Similarity) calcula a similaridade medindo o cosseno do ângulo entre dois vetores.

Matematicamente falando, a similaridade de cosseno é uma medida de similaridade entre dois vetores diferentes de zero de um espaço de produto interno 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, π] radianos.

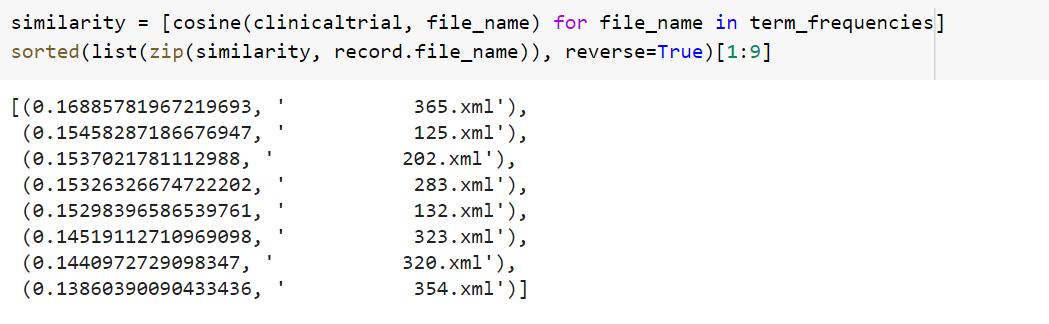


Fig 3.2.6 – Registros mais similares aplicando Similaridade do Cosseno

**3.4.3. 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.

O primeiro resultado ficou muito bastante pulverizado a nível de tópicos (Fig 3.2.7.1).

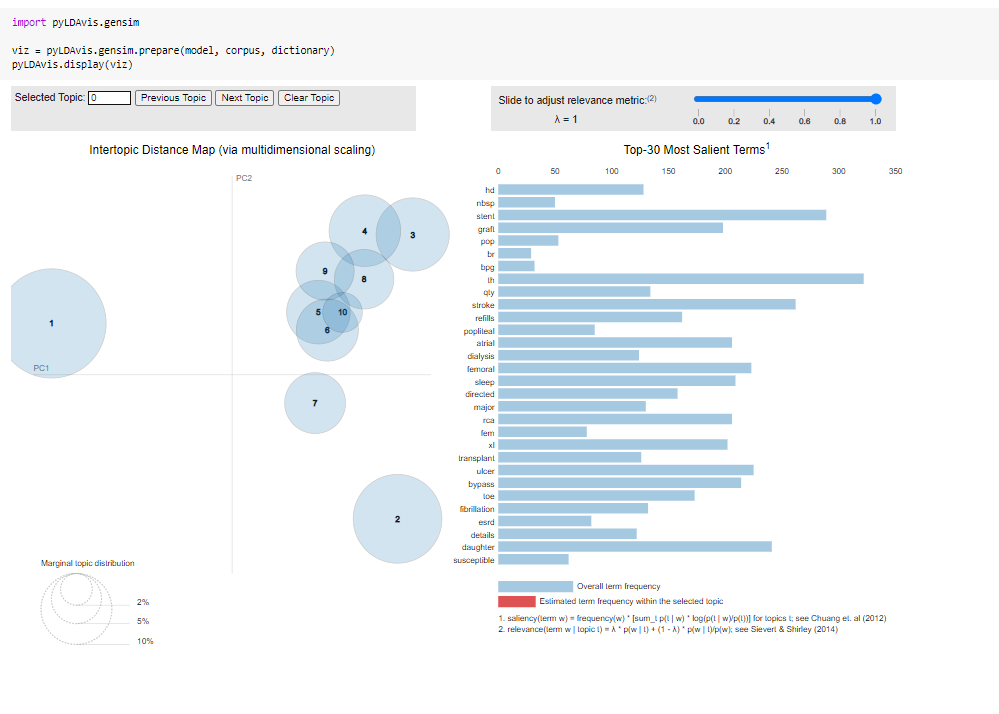


Fig 3.2.7.1 – Primeiro modelo com LDA

Foi usado o GridSearch para encontraro o melhor modelo LDA. O parâmetro de ajuste mais importante para o LDA é o numero 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 parãmetro que foi alterado foi a taxa de aprendizado (learning\_decay), essa taxa também não tem uma valor padrão.

Antes de aplicar o Grid Search foi feito ajustes no vetos de dados (CountVectorizer) alterando os parâmetro token\_patern para considerar como tokens apenas palavras alfanuméricas de maior que 5 caracteres (Fig 3.2.7.2).

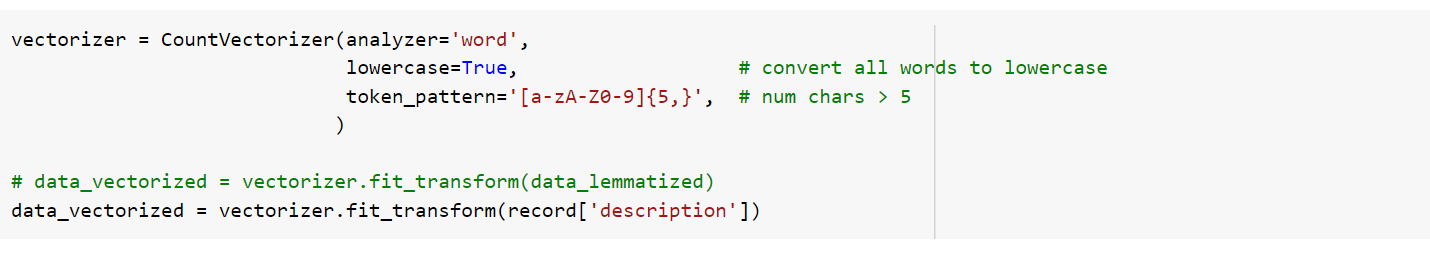


Fig 3.2.7.2 – CountVectorized ajustado

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 (Fig 3.2.7.3).

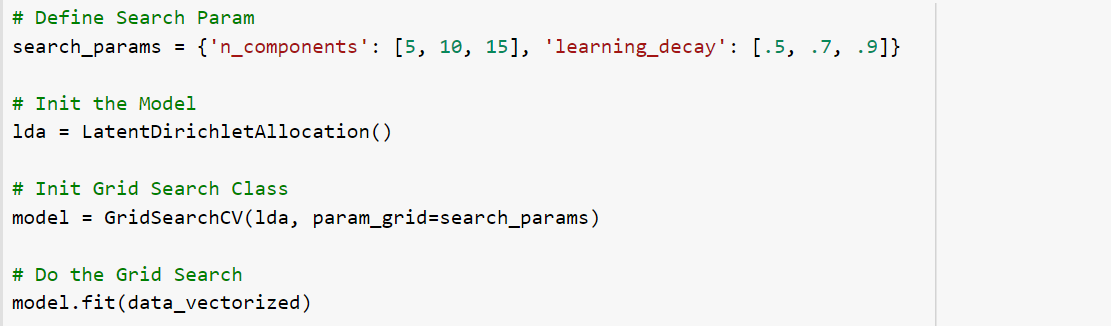


Fig 3.2.7.3 – Parâmetros do Grid Search

O resultado do Grid Search aponta os melhores parâmetos para o modelo LDA (Fig 3.2.7.4).

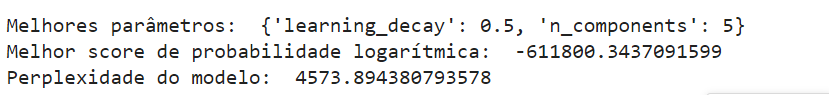
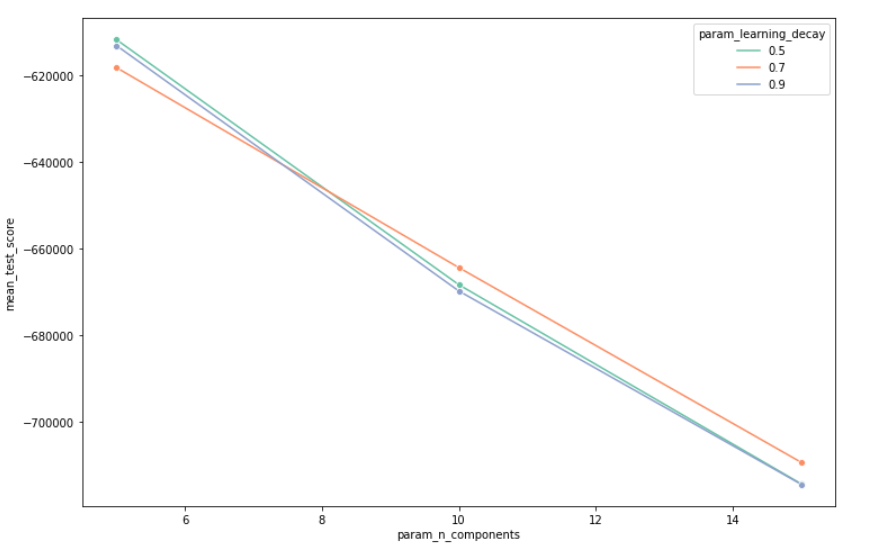


Fig 3.2.7.4 – Primeiro modelo com LDA

Também foi feita uma comparação entre os scores de performance dos modelos (Fig 3.2.7.5) .

Fig 3.2.7.5 – Comparação entre scores dos modelos

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)( Fig 3.2.7.7).

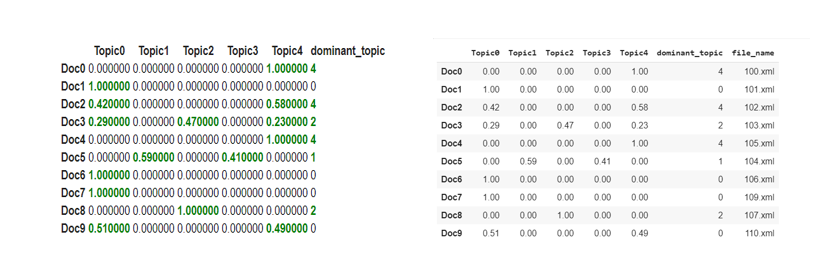


Fig 3.2.7.7 – Análise de tópicos dominantes (objeto style x dataframe), correlacionado os documentos aos registros de paciente

Outra análise que pode ser feita é a quantidade de documentos por tópico (Fig 3.2.7.8)

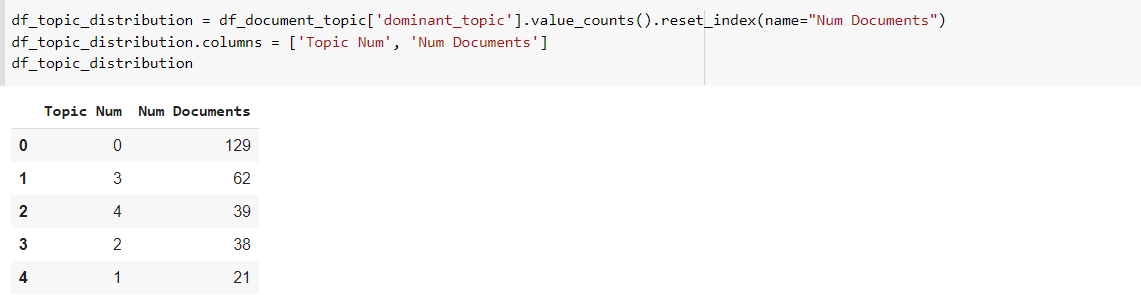


Fig 3.2.7.8 – Análise de tópicos dominantes

Com esses ajustes vejamos como ficou o modelo ajustado usando o pyLDA (Fig 3.2.7.9).

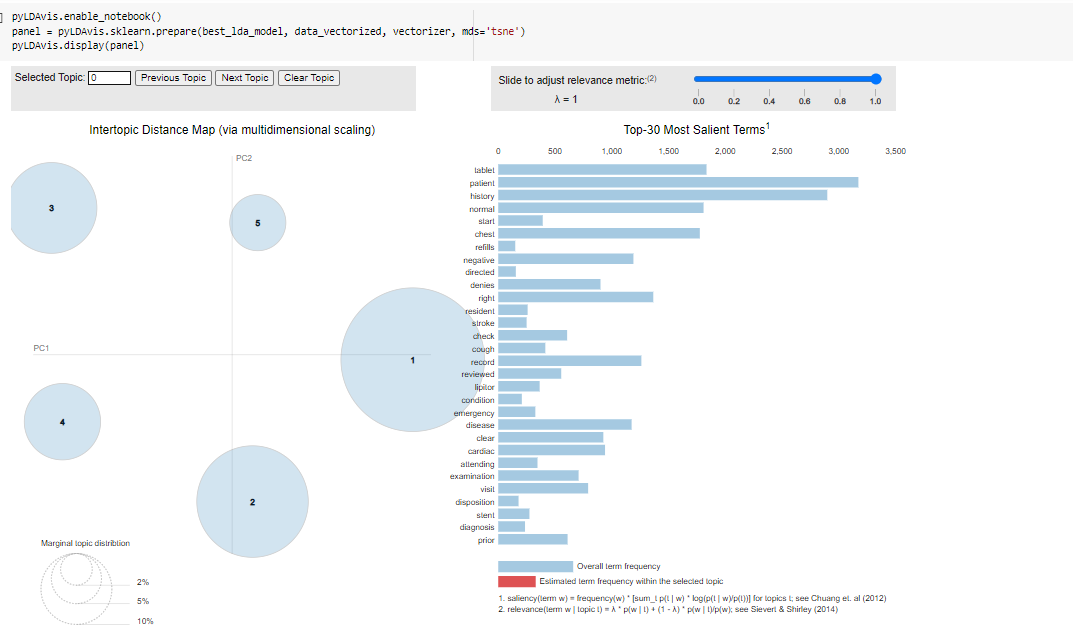


Fig 3.2.7.9 – Modelo ajustado

Verificando que o ensaio clinico foi classificado no tópico 2, buscamos a lista dos 15 mais do tópico 2 (Fig 3.2.7.0)

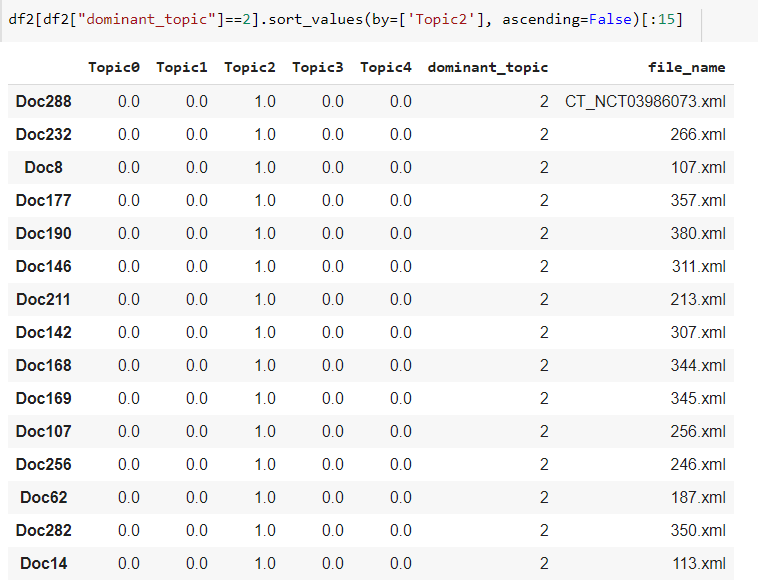


Fig 3.2.7.0 – 15 mais do tópico 2

**3.4.4. 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 [10].

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 (Fig 3.2.8).

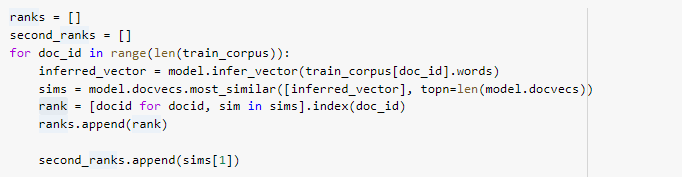


Fig 3.2.8 – Ranking de documentos

O resultado foi mostrado abaixo com os 5 documentos mais similares pelo modelo Doc2vec(Fig 3.2.8.1) .

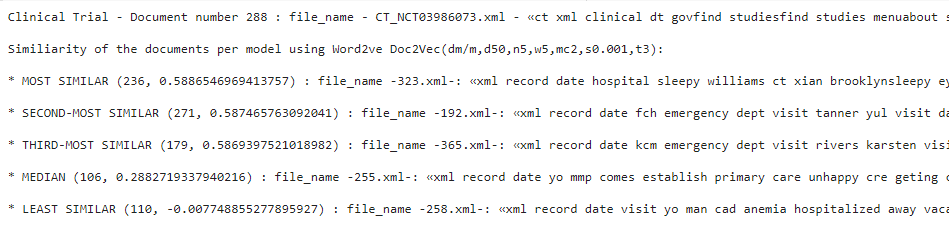


Fig 3.2.8 – resultado do racking

1. **Arquitetura do sistema proposto**
2. **Resultados**
3. **Conclusão e trabalhos futuros**

**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] Do site https://medium.com/swlh/latent-dirichlet-allocation-lda-eff969bda284

[7] Powerpoint da aula2 de BI

[8] https://www.machinelearningplus.com/nlp/gensim-tutorial/#2whatisadictionaryandcorpus

[9] Do site http://www.tfidf.com/

[10] NLP: Word Embedding Techniques Demystified. Rabeh Ayari, PhD