**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**

**Programa de Pós-graduação em Ciência de Dados e Big Data**

Elisangela Souza

Priscila Lara

# ANALISANDO A SILIMARIDADE ENTRE RESULTADOS DE EXAMES

Belo Horizonte

2020

Elisangela Souza

Priscila Lara

# ANALISANDO A SIMILARIDADE ENTRE RESULTADOS DE EXAMES

Trabalho apresentado ao Instituto de Educação Continuada (IEC) da Pós-Graduação em Ciência de Dados e Big Data da PUC Minas, como requisito parcial para obtenção de créditos na disciplina Processamento de Linguagem Natural.

Professor: Bárbara Silveira

Belo Horizonte

2020

## INTRODUÇÃO

NLP (natural language processing) ou processamento de linguagem natural (PLN em português) é o ramo da Ciência da Computação focado no desenvolvimento de sistemas que permitem que os computadores se comuniquem com pessoas que usam a linguagem cotidiana (falada/escrita) de maneira a responder suas demandas.

Com a NLP podemos extrair informações relevantes do texto, classificar documentos, fazer análise de sentimentos, encontrar textos relevantes no documento, modelagem de tópicos, etc.

## OBJETIVO

Desenvolver um projeto prático que utilize conceitos e tarefas de NLP aprendidas. Utilizaremos uma situação da vida real na área da saúde para aplicar os conhecimentos adquiridos durante as aulas.

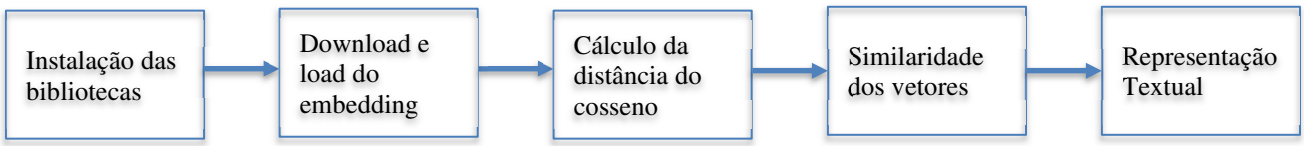
## DESENVOLVIMENTO

Nossa base de dados é um arquivo Excel. Os dados foram extraídos de uma base onde são imputados resultados de exames laboratoriais e contém as seguintes informações:

- Id: número sequencial que identifica unicamente um registro.
- Parametro: tipo do exame realizado.
- Divulgado: resultado informado pelos técnicos responsável pela análise do exame. É de preenchimento livre e cada responsável técnico preenche conforme lhe é mais adequado, sem uma padronização. Observa-se nesse campo palavras/sentenças escritas de formas diversas.
- Resultado: classificação da coluna divulgado.

Utilizamos os dados da coluna Divulgado para fazer as relações (análise de similaridade) entre algumas informações contidas nesta coluna. Como estamos analisando a similaridade entre palavras, utilizamos um embedding de 100 dimensões em português.

### Etapas do desenvolvimento



**ANÁLISE**

Usando o embedding temos um valor para todos os pares de resultado, ganhamos mais informação e maior poder de análise. Temos dimensões de análise e valores latentes que ajudam identificar a semelhança entre as palavras e consequentemente entre os resultados. Quanto menor a distância, mais próximo. Dessa forma, os resultados 1 e 2 estão mais próximos, fazendo sentido pois os resultados "nao detectado" e "nao detectável" estão no mesmo contexto.

Um outro passo do nosso projeto foi calcular a similaridade de vetores, onde inicialmente exibimos as palavras similares com as palavras "detectado" e "influenza", através da distância do cosseno. Para a palavra "detectado", temos uma similaridade de 0,8518 com a palavra "detectada", seguida de "observado" e "detectados". Já a palavra "influenza" tem 0,8674 de similaridade com a palavra "h1n1", seguida das palavras "gripe" e "h3n2".

Utilizamos também operações com vetores e verificamos que a segunda palavra está fazendo mais sentido do que a primeira:

- detectado, detectável, indeterminado: detectada (0,7428) | indetectável (0,6545)
- detectado, detectável, inconclusivo: detectada (0,7187) | indetectável (0,6927)

Além disso, calculamos a similaridade entre as seguintes palavras:

- "detectado", "detectável"= 0.76558065
- "indeterminado", "inconclusivo" = 0.5674166
- "detectado", "reagente" = 0.46118718

Por último, realizamos o treino com o embedding. Para isso, executamos um pré-processamento mínimo para tokenizar a coluna Resultado, criando uma coluna no dataframe para armazenar o dado processado. Os dados desta coluna foram tratados previamente, diretamente no arquivo xlsx. Identificamos a necessidade de transformar o X_train_embedding para duas dimensões, pois tinha um pouco mais de mil linhas, mas não tinha colunas, retornando apenas uma dimensão. Quando transformamos para embedding temos 100 colunas, pois está em fatores latentes. O modelo treinado consegue predizer 88% das vezes e tem boa precisão e revocação tanto para positivo quanto para negativo. Como o modelo está balanceado, podemos considerar apenas para exemplificação, onde conseguimos aplicar todas as etapas descritas acima e conhecimentos adquiridos durante as aulas.