INTRODUÇÃO

O processamento de linguagem natural (PLN) tem se tornado cada vez mais presente em nossas tarefas diárias, com aplicações que vão desde assistentes virtuais e modelos de IA generativa a programas especialistas de áreas diversas, como direito, medicina etc. O volume e a velocidade de compartilhamento de informações cresce exponencialmente junto com a necessidade de extrair e interpretar os diversos tipos de dados disponíveis atualmente.

As dúvidas em relação a como essas tecnologias entendem a nossa escrita, fala e gestos vem à nossa mente quando alguma inteligência artificial nos responde parecendo compreender nossa mensagem como um humano. Diante disso, vemos a necessidade do estudo e da divulgação sobre o funcionamento dessas ferramentas que tanto usamos.

OBJETIVO

O objetivo deste estudo é utilizar os conceitos e técnicas básicas vistas na disciplina de *Computação Científica e Análise de Dados* para introduzir um dos métodos de PLN, técnica pela qual tentamos passar algum nível de interpretação de fontes de linguagem humana ao computador, permitindo assim determinadas análises sobre essas fontes de dados.

A ideia é extrair temas ou tópicos de alguma base de texto, isso é feito de acordo com quais palavras e com qual intensidade elas serão representadas pelo algoritmo. Para testar essas técnicas, escolhi a base de notícias e fake news da central de contas, além de dois livros famosos, a Bíblia e Harry e a pedra filosofal.

RASCUNHO INICIAL

A primeira abordagem foi analisar os autovetores das matrizes de frequência (notícias/fakenews x palavras) para identificar possíveis padrões visuais nos dados. Após a plotagem, percebi que interpretar a parte negativa do PCA poderia ser de maior complexidade, mas é possível notar distribuição simétrica interessante. Nele, os pontos vermelhos representam notícias projetadas em três componentes principais, enquanto os pontos azuis correspondem a fake news representadas da mesma forma.

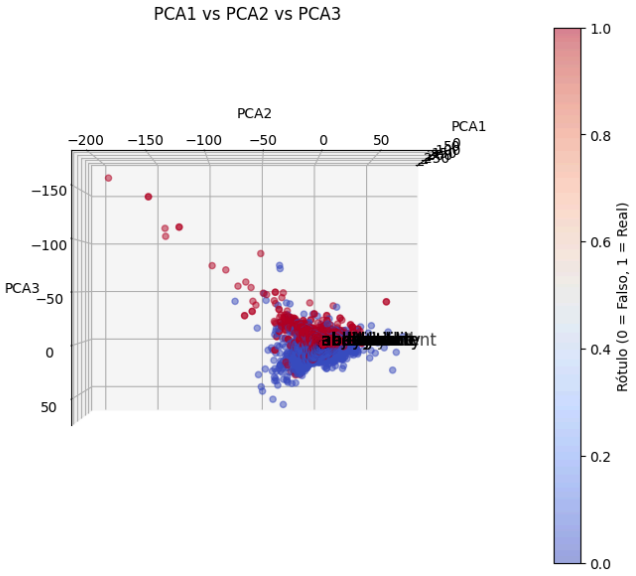


Figura 1: Plotagem de notícias e fake news nos PCA's

NORMALIZAÇÃO COM TF-IDF E LIMPEZA

O dataset escolhido foi da central de fatos que possuía inicialmente título, subtítulo e corpo do texto contendo as informações necessárias para o nosso objetivo. Nossa notícia é composta pela concatenação dos 3 campos, eliminando apenas linhas com corpo do texto nulas. Também foi utilizado o pacote stopwords, que descarta palavras que não acrescentam um sentido tão significativo ao texto, além de criarmos também uma lista customizável de palavras a serem descartadas da análise. EX: **a, e, o, ou, assim, como, que, etc**.

O dataset da Central de Fatos continha três campos principais da notícia: título, subtítulo e corpo do texto. Para padronizar as entradas e garantir que todas as informações relevantes estivessem presentes, concatenamos esses três campos em uma única estrutura de texto. Para aprimorar a qualidade da análise, utilizamos o pacote stopwords, removendo palavras de baixa relevância semântica, como **"a", "e", "o", "ou", "assim", "como"** e **"que"**, entre outras.

A normalização foi deita por meio do **TF-IDF**, que consiste em dar peso às frequências, de modo a diminuir a importância de palavras que aparecem muitas vezes na maioria das notícias e portanto não trazem grandes informações sobre as diferenças que queremos capturar do conjunto, temas/assuntos das notícias. O cálculo pode ser feito com mudanças sutis, mas normalmente tem o seguinte ajuste:

$$tf\text{-}idf(t, d) = tf(t, d) * log(N/N(t))$$

Onde:

tf: Frequência da palavra na notícia

log(N/N(t)): Peso da frequência da palavra, log do número de documentos/frequência da palavra nos documentos. Desse modo, quanto mais esse termo aparece em diferentes documentos, menor seu peso associado. Logo, de forma análoga, quanto mais a palavra aparece em poucos documentos, maior o seu peso.

Algumas semelhanças aparecem quando olhamos para o processo de extração de temas com NMF e para o processo de agrupamento com K-means. Nessa comparação os pontos teriam o papel das palavras, sua cor representaria o tema que essa palavra melhor descreve. Algumas palavras como "foto", “imagens”, “notícia”, “etc” podem aparecer na maioria das notícias, mesmo com o bloqueio de boa parte dessas palavras com o stopword, a normalização com o TF-IDF entra para diminuir a intensidade das palavras/pontos que estão mais a margem entre os grupos, de modo que as palavras/pontos mais centrais de cada grupo tenham mais relevância na descrição desses grupos.

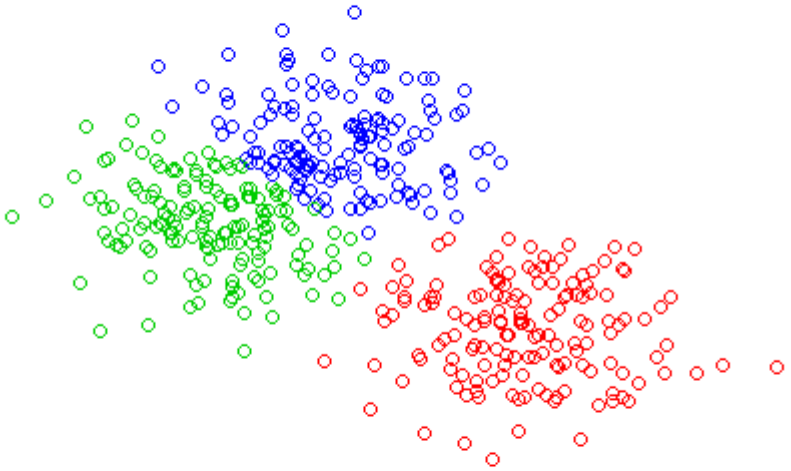


Figura 2: Agrupamento por K-means

NMF

Baseia-se na decomposição de uma matriz não negativa, fazendo parte do conjunto de técnicas de aprendizado não supervisionado, bastante usadas em modelagens como a nossa devido a sua natureza de não possuir elementos negativos. Essa matriz normalmente é preenchida com a frequência de determinadas características em uma amostra, ela é então decomposta em outras duas matrizes também positivas que são então analisadas a partir de um valor K passado.

A função recebe um K representando a quantidade de colunas da matriz W, seguindo as regras básicas de multiplicação de matrizes. Para nós o K representa a quantidade de grupos ou temas mais importantes que queremos extrair das notícias.

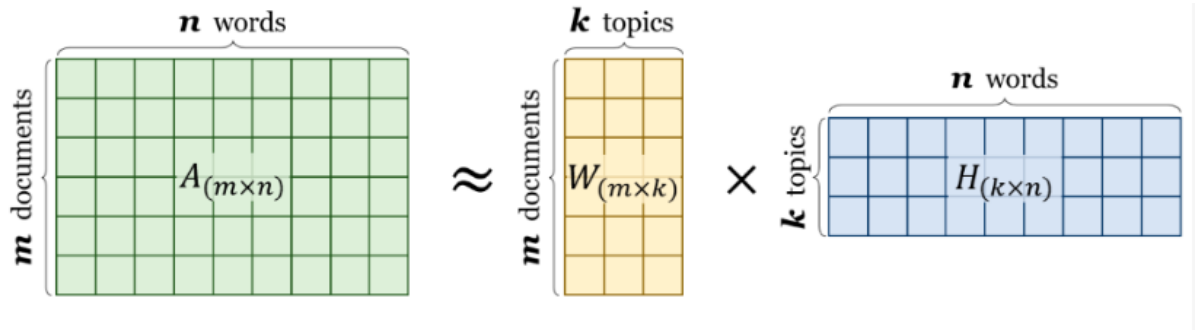


Figura 3: Decomposição da matriz base normalizada

Além da matriz de entrada, o método pode receber alguns argumentos como a semente de reprodução da aproximação e outras variáveis de estabilização.

```
modelo_nmf = NMF(n_components=num_topicos, random_state=42)
 matriz_documento_tema = modelo_nmf.fit_transform(dados["matriz_tfidf"])
 matriz_tema_por_palavra = modelo_nmf.components_
```

DECOMPOSIÇÃO

A técnica principal empregada nesse tipo de decomposição é a da atualização multiplicativa, que consiste em minimizar alternadamente W e H fazendo com que o módulo ao quadrado de M - WH seja o menor possível. O processo de aproximação é descrito no artigo "*Algorithms for Non-negative Matrix Factorization"*.

Usando derivação parcial nas matrizes é possível buscar uma aproximação para cada um dos valores a serem preenchidos.

$$W_{i,j} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}}$$

$$H_{i,j} \leftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T WH)_{ij}}$$

Figura 4: Representação de aproximação das matrizes

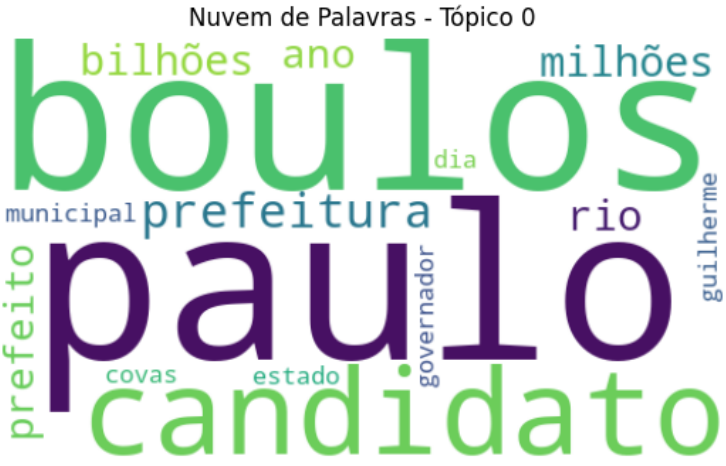
A principal vantagem da abordagem por NMF vem justamente dessa decomposição, que permite uma análise tanto na matriz W quanto na H, trazendo e aumentando a interpretabilidade dos dados e das reduções.

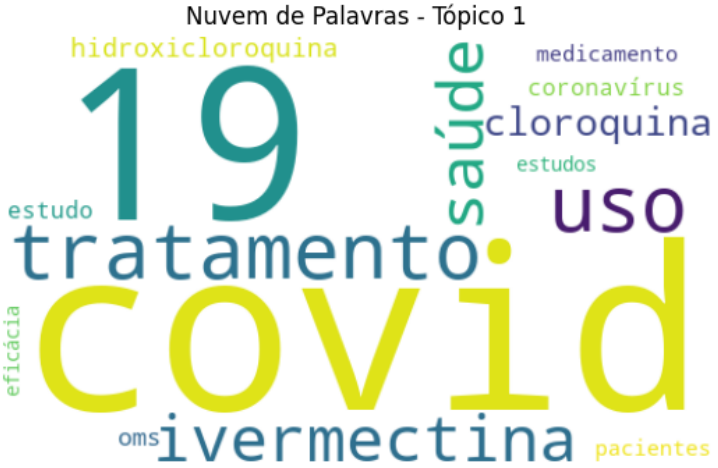
RESULTADOS

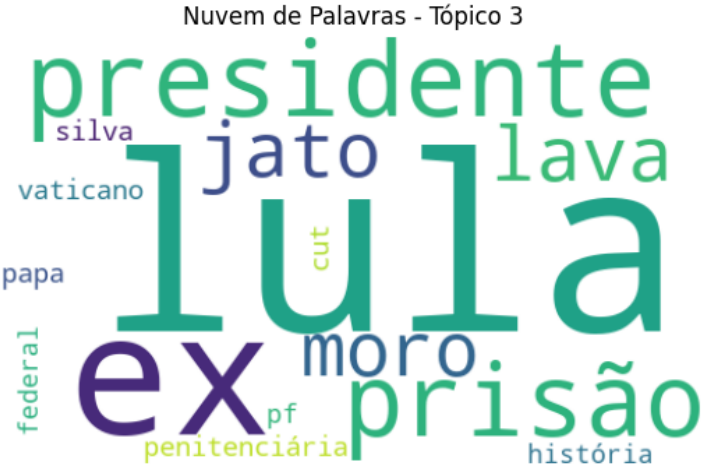
Para melhor visualização dos dados o wordcloud foi utilizado, esse estilo de visualização nos ajuda a entender não só as palavras principais, mas também sua respectiva intensidade. A função ***extrair_tópicos*** recebe 4 parâmetros, onde escolhemos a tabela a ser analisada, a coluna onde os textos se encontram, o número de temas que deseja extrair e o número de palavras a ser mostrada na nuvem. Exemplo de uso:

```
extrair_topicos(fake_df, coluna_texto= "texto", n_topicos=10, max_words=15)
```

Queremos da tabela de fake news, pegar as notícias que se encontram na coluna “texto“, analisar os 10 principais temas e mostrar suas 15 palavras mais relevantes. Veja os 4 primeiros resultados desta chamada.



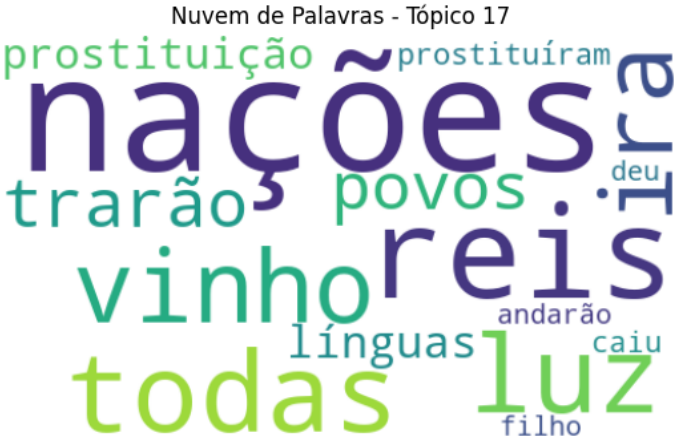




Podemos imaginar que o principal assunto das notícias são fakenews direcionadas a política de são paulo, citando o candidato Boulos. O segundo maior assunto das fake news do nosso conjunto de dados parece ser sobre tratamentos ineficazes contra Covid 19. Promoções/correntes parece ser o terceiro e fake news sobre o presidente lula o quarto. Alguns outros podem ser vistos conforme aumentamos o número de tópicos a serem extraídos.

OUTROS TEXTOS

A motivo de curiosidade, testei o algoritmos algumas vezes na Bíblia, obtendo algumas imagens curiosas, veja alguns exemplos.



Na *nuvem de palavras - tópico 24*, podemos imaginar que o algoritmo caracterizou **números por extenso** como o vigésimo quarto principal “tema” da bíblia.

Já na *nuvem de palavras - tópico 5*, o tema parece ser **família**.

A *nuvem de palavras - tópico 16* foi retirada do livro de Apocalipse.

Já no livro de ***Harry Potter e a pedra filosofal*** os resultados foram piores, mas consegui obter essa nuvem, que juntou nela a maioria das palavras relacionadas a casa onde Harry morava. O oitavo tema principal do livro poderia ser "**parentes chatos/casa".**

Os resultados obtidos parecem satisfatórios uma vez que boa parte dos temas extraídos é visualmente identificável.

Uma das coisas interessantes sobre esse processo está na possibilidade de alterar a quantidade de temas que queremos extrair, podendo obter temas relevantes bem diferentes ou mais específicos. Supondo que queremos extrair 3 temas, obtendo saúde, política e entretenimento como as áreas mais relevantes do conjunto. Se aumentarmos o número de temas para 6, a área de entretenimento pode não aparecer mais, uma vez que os 6 temas mais relevantes podem estar dentro das áreas de saúde e política. De modo análogo, se diminuímos é possível que outras áreas apareçam ou sumam.

Os resultados obtidos parecem satisfatórios uma vez que boa parte dos temas extraídos é visualmente identificável, mas apesar dos resultados tentarem nos “resumir” os principais temas, a interpretação ainda é feita por nós, após visualizarmos as palavras e seus tamanhos, ainda é preciso associar essas representações a conhecimentos e impressões pessoais.