**Processamento de linguagem natural com Python [25E2\_2]**

**Aluna:** Rachel Reuters

**Implementar técnicas de lematização**

1. Qual o endereço do seu notebook (colab) executado? Use o botão de compartilhamento do colab para obter uma url.

🡺<https://github.com/rachelreuters/posGraduacaoIA/blob/main/NLP/Projeto_de_Disciplina_de_Text_Mining.ipynb>

1. Em qual célula está o código que realiza o download dos pacotes necessários para tokenização e stemming usando nltk?

Texto

O conteúdo gerado por IA pode estar incorreto.

1. Em qual célula está o código que atualiza o spacy e instala o pacote pt\_core\_news\_lg?

Texto

O conteúdo gerado por IA pode estar incorreto.

1. Em qual célula está o download dos dados diretamente do kaggle?

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

1. Em qual célula está a criação do dataframe news\_2016 (com examente 7943 notícias)?

Interface gráfica do usuário, Texto, Site

O conteúdo gerado por IA pode estar incorreto.

1. Em qual célula está a função que tokeniza e realiza o stemming dos textos usando funções do nltk?

Texto

O conteúdo gerado por IA pode estar incorreto.

1. Em qual célula está a função que realiza a lematização usando o spacy?

Texto

O conteúdo gerado por IA pode estar incorreto.

1. Baseado nos resultados qual a diferença entre stemming e lematização, qual a diferença entre os dois procedimentos? Escolha quatro palavras para exemplificar.
   1. O **stemming** é um processo mais heurístico e mais simples que remove sufixos (e às vezes prefixos) das palavras para obter o seu radical (ou stem). Ele não se preocupa se o radical resultante é uma palavra válida no dicionário.

Já a **lematização** é um processo mais sofisticado que visa reduzir uma palavra à sua forma canônica ou de dicionário. Este processo leva em consideração a morfologia da palavra e o seu contexto na frase.

* 1. Alguns exemplos:

|  |  |  |  |
| --- | --- | --- | --- |
| Stemming | Lematização | Palavra original | Onde encontrar |
| mor | morador | moradores | 10 documento |
| veloc | velocidade | velocidades | 10 documento |
| condi | condição | condições | 10 documento |
| telefn | telefônico | telefônica | 10 documento |

**Construir um modelo de reconhecimento de entidades (NER) usando Spacy**

1. Em qual célula o modelo pt\_core\_news\_lg está sendo carregado? Todos os textos do dataframe precisam ser analisados usando os modelos carregados. Em qual célula isso foi feito?

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

1. Indique a célula onde as entidades dos textos foram extraídas. Estamos interessados apenas nas organizações.

Texto

O conteúdo gerado por IA pode estar incorreto.

1. Cole a figura gerada que mostra a nuvem de entidades para cada tópico obtido (no final do notebook)  
   Calendário

   O conteúdo gerado por IA pode estar incorreto.

**Criar modelos utilizando vetorização de textos baseado em Bag of Words**

1. Quando adotamos uma estratégia frequentista para converter textos em vetores, podemos fazê-lo de diferentes maneiras. Mostramos em aula as codificações One-Hot, TF e TF-IDF. Explique a principal motivação em adotar TF-IDF frente as duas outras opções.

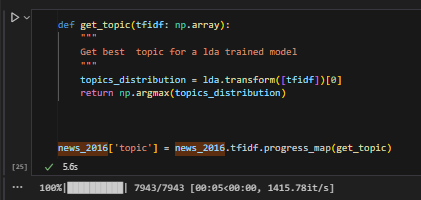
🡺 O TF-IDF põe mais peso às palavras que são importantes dentro de 1 documento, mas não muito frequentes nos outros, ele também reduz a influência de palavras comuns, aumentando o impacto de termos específicos . Esse método também reduz o peso dos termos que são muito comuns em toda a coleção.

1. Indique a célula onde está a função que cria o vetor de TF-IDF para cada texto.

Texto

O conteúdo gerado por IA pode estar incorreto.

1. Indique a célula onde estão sendo extraídos os tópicos usando o algoritmo de LDA.



1. ~~Indique a célula onde a visualização LDAVis está criada.~~
2. Cole a figura com a nuvem de palavras para cada um dos 9 tópicos criados.

Calendário

O conteúdo gerado por IA pode estar incorreto.

1. Escreva brevemente uma descrição para cada tópico extraído. Indique se você considera o tópico extraído semanticamente consistente ou não.   
   Tópico 0: Setor financeiro, talvez mercado de ações e câmbio

Tópico 1: Eventos Acadêmicos e debates (mais voltados para o lado Político-Sociais)

Tópico 2: Atividade Empresarial e Comércio Nacional

Tópico 3: Setor de petróleo e geopolítica energética

Tópico 4: Declaração de Imposto de renda, como declarar, dicas e afins.

Tópico 5: Política monetária, inflação

Tópico 6: Setor elétrico, tarifa, conta de luz e transporte aéreo

Tópico 7: Setor de óleo e gás do mercado brasileiro, mais focado na Petrobras que é a principal estatal do país nesse ramo.

Tópico 8: Cenário político do Brasil, 3 poderes, decisões políticas.

Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

**Criar modelos baseados em Word Embedding**

1. Neste projeto, usamos TF-IDF para gerar os vetores que servem de entrada para o algoritmo de LDA. Quais seriam os passos para gerar vetores baseados na técnica de Doc2Vec?
   * 1. Coletar um conjunto de documentos de texto
     2. Tokenizar os textos
     3. Remover stopwords
     4. Criação e Treinamento do Modelo Doc2Vec
     5. Transformar os documentos em vetores
     6. Aplicação dos Vetores: Os vetores gerados podem ser usados para clusterização com K-Means, DBSCAN, classificação de textos (SVM, Random Forest) ou treinamento de modelos de deep learning.
2. Em uma versão alternativa desse projeto, optamos por utilizar o algoritmo de K-Médias para gerar os clusters (tópicos). Qual das abordagens (TF-IDF ou Doc2Vec) seria mais adequada como processo de vetorização? Justifique com comentários sobre dimensionalidade e relação semântica entre documentos.

🡺 O processo de vetorização mais adequado no caso de utilizar um algoritmo de clusterização tipo o Kmeans seria o Doc2Vec pois reduz a dimensionalidade e torna a abordagem mais escalável. Além disso essa abordagem leva em consideração a relação semântica entre as palavras pois captura o contexto em que a palavra está inserida.

🡺Ao final do notebook, eu apliquei o algoritmo de Kmeans nos vetores gerados e também com redução de dimensionalidade para 2 dimensões. Os resultados foram bem próximos ao TF-IDF.

1. Leia o artigo "Introducing our Hybrid lda2vec Algorithm" (https://multithreaded.stitchfix.com/blog/2016/05/27/lda2vec/#topic=38&lambda=1&term=) .  
   O algoritmo lda2vec pretende combinar o poder do word2vec com a interpretabilidade do algoritmo LDA. Em qual cenário o autor sugere que há benefícios para utilização deste novo algoritmo?

🡺 Essa abordagem requer um esforço computacional bem grande, portanto mais indicada quando a quantidade de dados é muito elevada. Cenários onde os tópicos precisam ter relações mais profundas com documentos ou usuários. Uma utilização indicada para esse modelo seria personalizar um sistema de recomendação de artigos baseado nos interesses do usuário ou análise de tendencia de um blog ao longo dos anos.