Unyleya
EDUCACIONAL

Fundamentos de mineração de texto e processamento de linguagem natural

Mineração De Texto e Processamento De Linguagem Natural

Mineração de texto == Mineração de dados textuais

Processo de obtenção de informações importantes de um texto

Reconhecimento de padrões

Linguagem natural vs. linguagem de programação

Linguagem natural não se destina a ser traduzida em um conjunto finito de operações matemáticas e por isso é mais complexa.

O processamento de linguagem natural (PNL) é uma área de pesquisa voltada para o processamento de linguagens naturais tão distintas como o inglês e o mandarim.

O PNL envolve sua tradução em dados (números) que um computador pode usar para aprender sobre o mundo.

Aplicações práticas

- Notícias
- Atribuição de autoria e veracidade
- Análise de sentimento
- Predição de comportamento
- Escrita criativa

Aplicações práticas

Busca	Internet	Documentos	Autocompletar
Edição	Ortografia	Gramática	Estilo
Dialogo	Chatbot	Assistentes	Agendamento
Escrita	Índices	Concordância	Tabela de conteúdos
E-mail	Filtro de Spam	Classificação	Priorização
Documentação	Resumos	Extração de conhecimento	Diagnósticos médicos
Justiça	Inferência legal	Busca de precedente	Classificação de intimação
Notícias	Detecção de eventos	Verificação de fatos	Composição de manchete
Atribuição	Detecção de plágio	Literatura forense	Treinador de estilo
Análise de sentimento	Monitoramento de moral	Análise de críticas	Tratamento a clientes
Predição de comportamento	Finanças	Previsão de eleições	Marketing
Escrita criativa	Roteiro de filmes	Poesia	Letra de música

Tradução livre: LANE, Hobson; HOWARD, Cole; HAPKE, Max Hannes. Natural Language Processing in Action. [S.l: s.n.], 2019.

2- A representação Bag of Words para análise de frequência

Bag of Words

- Um texto é representado como um pacote (bag) de suas palavras, desconsiderando a gramática e até mesmo a ordem das palavras, mas mantendo a multiplicidade.

- Exemplo: João gosta de assistir filmes. Maria também gosta de filmes.

João	gosta	de	assistir	filmes	Maria	também
1	2	2	1	2	1	1

Comparando com Bag of Words

- João gosta de assistir filmes.

João	gosta	de	assistir	filmes	Maria	também
1	1	1	1	1	0	0

- Maria também gosta de filmes.

João	gosta	de	assistir	filmes	Maria	também
0	1	1	0	1	1	1

3 - Visualizando textos com nuvem de palavras

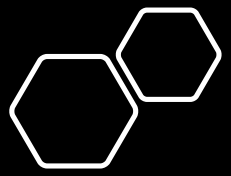
Nuvem de palavras (Word Cloud)

- É uma lista hierarquizada visualmente, uma forma de apresentar os itens de um conteúdo.
- Que música está nessa nuvem ao lado?

Nuvem de palavras (Word Cloud)

- Útil para comparar textos pelas palavras mais frequentes.
- E agora qual música está nessa nuvem ao lado?

4 - Segmentando textos em tokens



Segmentando textos em tokens

- O processo de segmentação de texto em que se divide uma string de caracteres em uma lista de tokens.
- Tokens podem ser palavras em uma frase e frases e um parágrafo.
- Palavras também pode ser separadas em tokens menores como silabas.
- A tokenização pode ocorrer por separação por espaços ou pontuações ou qualquer regra para separar significados.

Segmentando textos em tokens

Exemplo: Esta é uma sentença, e esta é outra sentença.

[“Esta é uma sentença,”, “e esta é outra sentença.”]

Segmentando textos em tokens

Outro exemplo: Esta é uma sentença.

Considerando só espaços.

[“Esta”, “é”, ”uma”, “sentença.”]

Considerando espaços e pontuação.

[“Esta”, “é”, ”uma”, “sentença”, “.”]

5 - O problema das palavras vazias

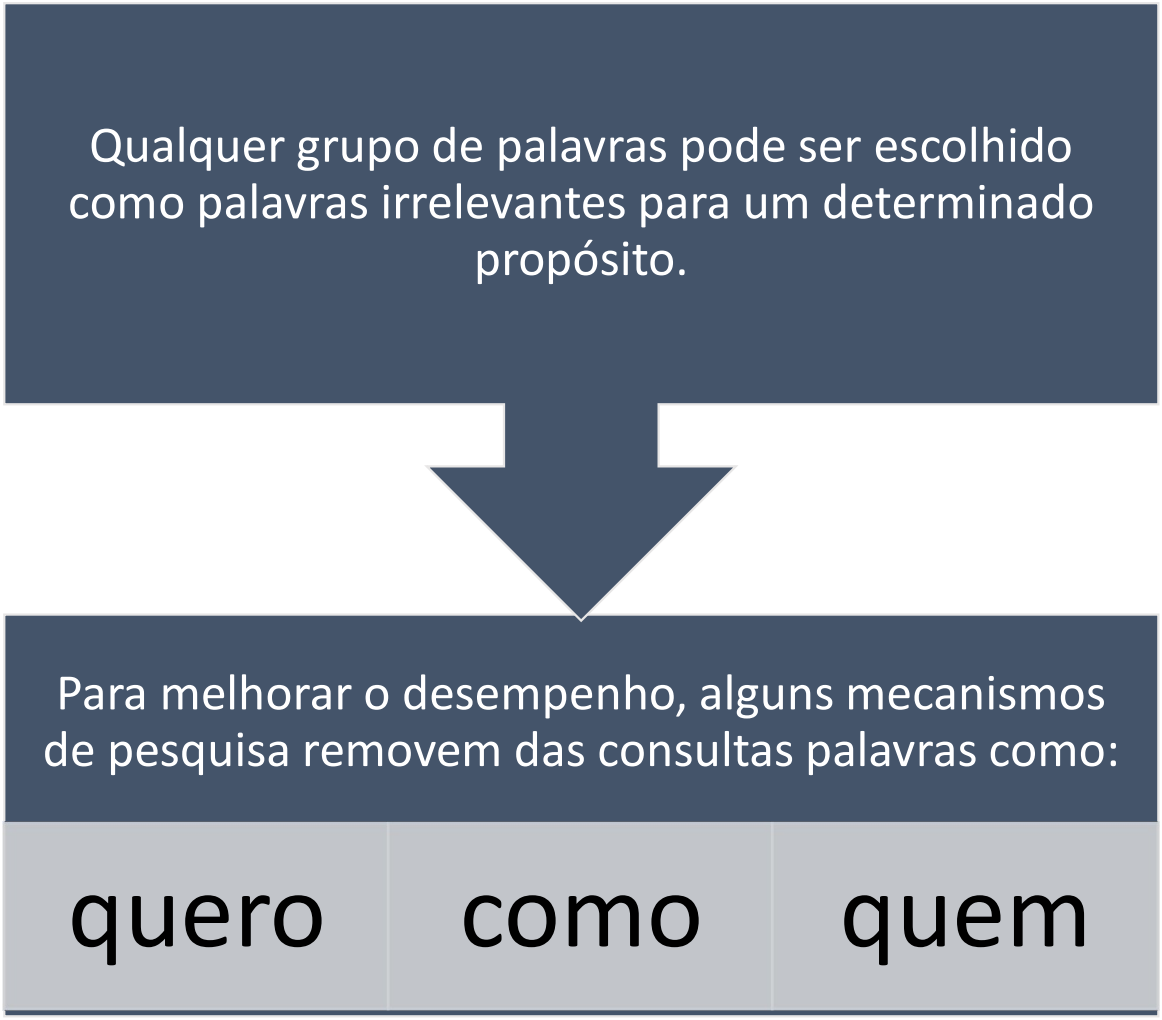
O problema das palavras vazias

Palavras vazias (stop words) são palavras comuns em qualquer idioma que ocorrem com alta frequência, mas contêm informações muito menos substantivas sobre o significado de uma frase.

Exemplos de algumas palavras vazias comuns incluem:

o, a	e, ou	em, na

O problema das palavras vazias



Unyleya EDUCACIONAL

Obrigado!