Mineração de Textos

Tópicos de Inteligência Artificial

Prof. Dr. Dieval Guizelini

março/2023

22º Café Cultural - Inteligência Artificial

22º Café Cultural - Inteligência Artificial

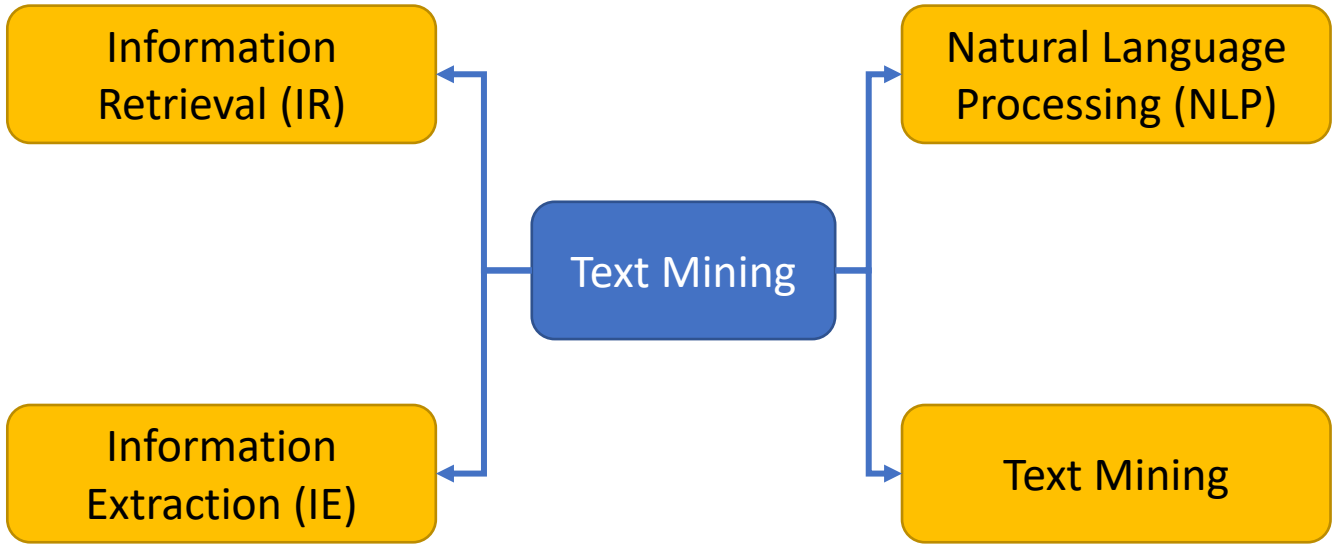
Sérgio Said Staut Junior
Diretor do
Setor de Ciências Jurídicas
da UFPR

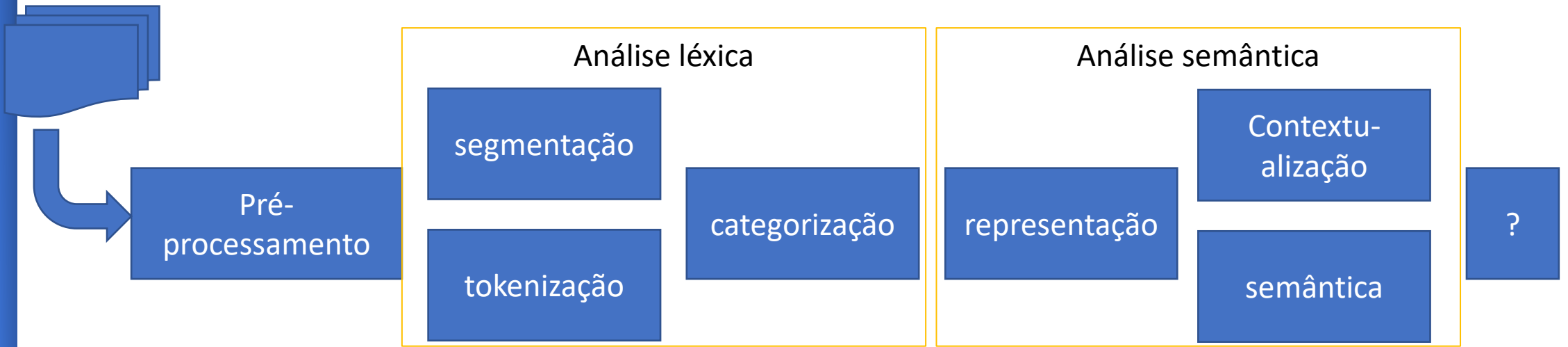
Link: https://www.youtube.com/watch?v=NHw33Cj5b80

Uma definição... incompleta

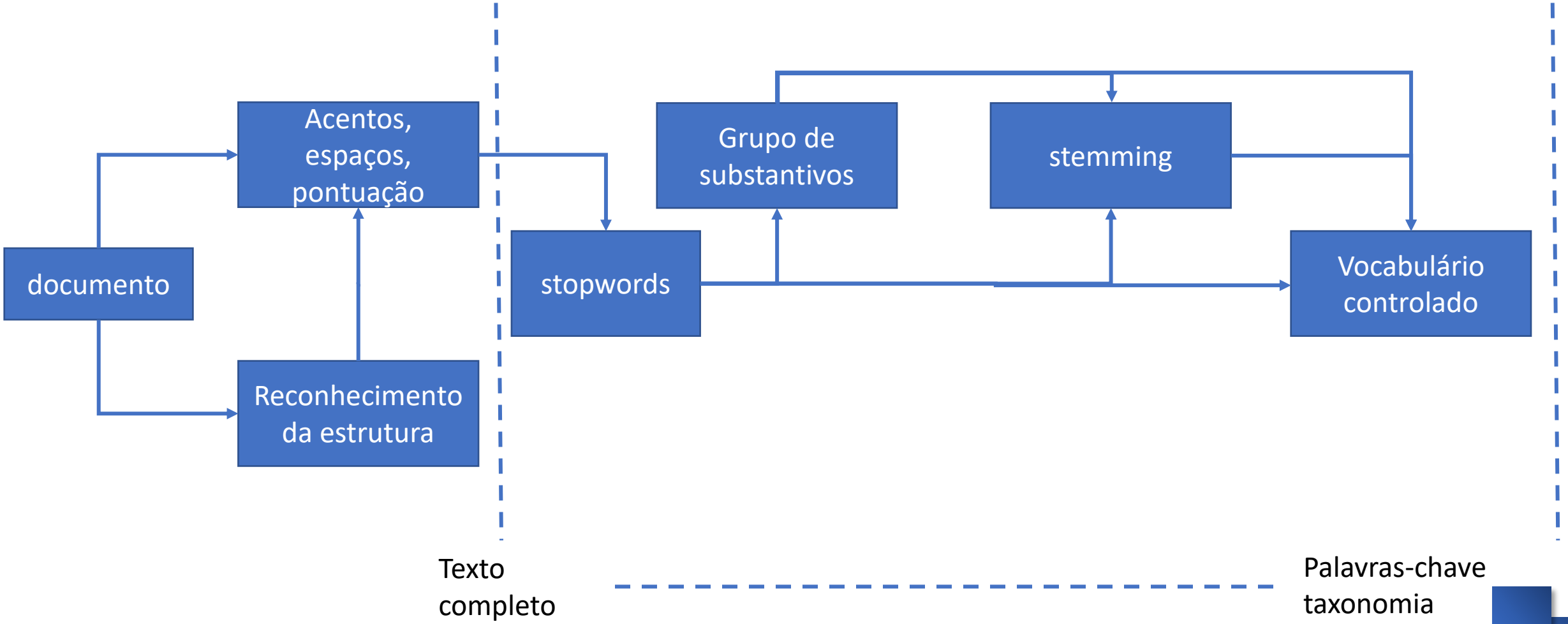
- A Mineração de Texto, também conhecida por Text Mining, Text Processing ou ainda Text Analytics, é um processo semiautomatizado para extração de conhecimento de fontes de dados não-estruturados.

Visão geral

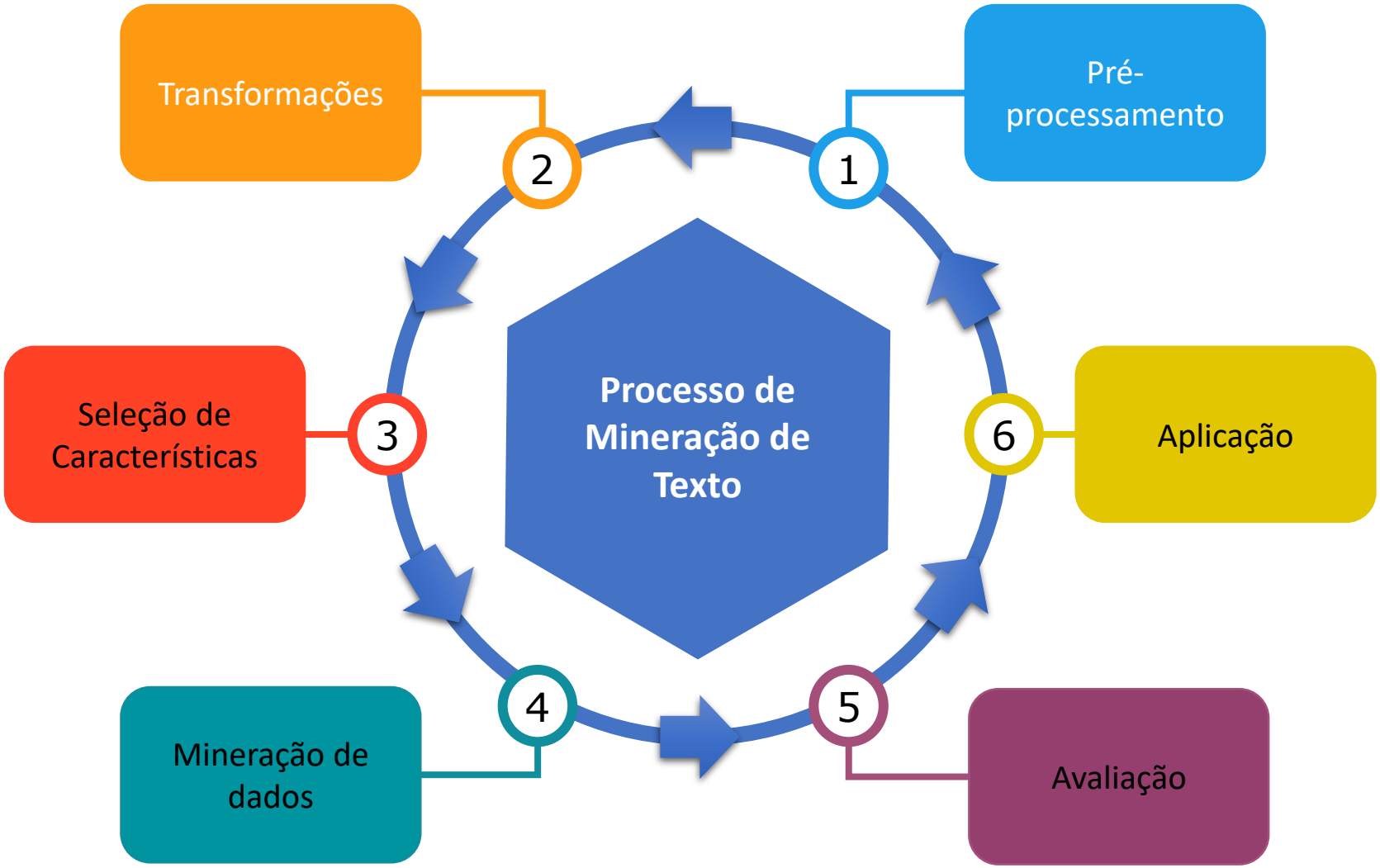

Information Retrieval (IR)
Natural Language Processing (NLP)
Text Mining
Information Extraction (IE)
Text Mining


Pré-processamento
Análise léxica
segmentação
tokenização
categorização
Análise semântica
representação
Contextu-alização
semântica
?

Análise léxica



O processo ganha fases/etapas...



Qual o “peso” das palavras em uma sentença?

Parte das subtarefas comuns são:

- A **redução da dimensionalidade** é uma técnica importante para o pré-processamento de dados.
- A **recuperação de informações** ou **identificação de um corpus** é uma etapa preparatória
- O **reconhecimento de entidades nomeadas** é o uso de **dicionários geográficos** ou **técnicas estatísticas** para identificar recursos de texto nomeados/categorizados.
- A **desambiguação** - o uso de pistas contextuais.
- Reconhecimento de Entidades Identificadas por Padrão (regex)
- Clustering de documentos.
- Co-referência: identificação de sintagmas nominais e outros termos que se referem ao mesmo objeto.
- Extração de relacionamento, fato e evento.
- A análise de sentimento envolve discernimento subjetivo.
- A análise quantitativa de texto.

Cinco gerações

- Co-ocorrência (*a priori*)
- Baseadas em regras
- Baseadas em conhecimento e abordagens estatísticas
- Baseadas em aprendizado de máquina
- Sistema especialista

Grandes problemas que permanecem:

- Ambiguidade
- “gordura” (adjetivos)
- Prefixos e sufixos

Quais os objetivos da mineração de dados?

- Toda mineração de dados visa a descoberta de conhecimento (KDD, Fayad, 1996)
- “O objetivo do Text Mining é explorar as informações contidas em documentos textuais de várias maneiras, incluindo ... descoberta de padrões e tendências nos dados, associações entre entidades, regras preditivas, etc.” (Grobelnik et al., 2001).
- Recuperação de Informação (Ricardo Baeza-Yates e Berthier Ribeiro-Neto)
- Indexação
- Enriquecimento de dados
- Reconhecimento de entidades nomeadas (NER)
- Marketing: análise de sentimento (bag of words, Nayve Bayes)

Representação / codificação

- Redução da dimensionalidade
- Dicionário de termos controlados
  (MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed).
- Remoção de termos
- N-grams
- Dicionários dinâmicos
- Vetorial (palavras, transição de símbolos, codificação)

Tarefas relacionadas a mineração de textos

1) Classificação de Textos

Definição Básica:
Visa associar documentos de texto a classes temáticas pré-definidas. Os documentos são classificados a partir de características do texto como termos ou palavras presentes nos documentos.

Classificação de texto tem sido usada, por exemplo, para indexação de documentos, filtragem de documentos, e extração de informação. As técnicas aplicadas envolvem comumente o uso de Engenharia de Conhecimento (envolvendo sistemas de classificação com regras definidas por especialistas) e o uso de Algoritmos de Aprendizagem de Máquina Supervisionada (e.g. aprendizado bayesiano, kNN, redes MLP, Support Vector Machines,...), onde o classificador de texto é induzido a parir de um corpus de documentos previamente etiquetados.

Artigo: Sebastiani, F. Machine learning in automated text categorization. ACM Comput. Surv. 34(1), pp. 1-47, 2002

Tarefas relacionadas a mineração de textos

2) Agrupamento de Textos

Definição Básica:

Corresponde a identificar grupos de documentos similares entre si. Cada documento é similar aos documentos pertencentes ao mesmo grupo, e diferente dos documentos pertencentes a outros grupos. Ao contrário da classificação de texto, o objetivo do agrupamento é encontrar classes ou grupos de documentos não conhecidos a priori.

Agrupamento de documentos textuais tem sido usado para a navegação de uma coleção de documentos (i.e. gerar uma taxonomia de documentos semelhante, por exemplo, aos diretórios do Yahoo), e para organizar os resultados de uma consulta resolvida por um engenho de busca (e.g. Vivisimo).

Envolve em geral o uso de técnicas de Aprendizagem de Máquina Não-Supervisionada (e.g. k-means, clustering hierárquico, redes SOM,...).

Steinbach, M.; Karypis, G.; Kumar, V. A comparison of document clustering techniques. KDD Workshop on Text Mining, 200

Tarefas relacionadas a mineração de textos

3) Extração de Informação

Definição Básica: Identificar dentro de um documento textual, trechos que correspondem a dados relevantes para um usuário (ex.: extrair nome e preço de produtos a partir de anúncios em páginas web).

Os dados extraídos são armazenados em um banco de dados que podem ser acessados diretamente pelo usuário, ou que podem servir como entrada para processos posteriores de mineração de dados.

Sistemas de Extração de Informação envolvem comumente o uso de Engenharia de Conhecimento, Processamento de Linguagem Natural e Aprendizado de Máquina.

Silva. E. Extração de Informação. In.: Um sistema para extração de informação em referências bibliográficas baseado em aprendizagem de máquina. Dissertação de Mestrado, Centro de Informática, UFPE, Cap.2, pp. 5-29, 2004

Tarefas relacionadas a mineração de textos

3) Análise de Sentimentos

Definição Básica: Identificar a opinião expressa sobre um determinado objeto (produtos, pessoas, empresas, etc...) a partir da análise de documentos de texto contendo como reviews, comentários, opiniões, dentre outros.

Em uma forma mais simples consiste em determinar graus de polaridade (positiva ou negativa) que um texto expressa sobre determinado assunto.

Análise de sentimentos é associada ao tema de mineração de opiniões e tem sido usada em aplicações práticas para monitoramento de mídias sociais, a fim de se identificar de forma automática o que se fala sobre determinado objeto ou pessoa e se fala de forma positiva ou negativa. Envolve técnicas de Recuperação de Informação, Processamento de Linguagem Natural e Aprendizagem de Máquina.

Pang, B; Lee, L. E. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, Vol. 2, pp. 1-135, 2008

São temas estudados e relacionados a text-mining

1. Medidas descritivas para texto.
    1. Noções de linguística.
    2. Índices de diversidade, abundância e riqueza.
2. Abordagens para a mineração de texto.
    1. Bag of words (BOW) e NLP.
    2. Manipulação de cadeias de caracteres.
    3. Pré-processamento para BOW.
3. Visualização em mineração de texto.
    1. Termos frequentes, associações e redes de relacionamento.
    2. Dendogramas, núvem de palavras.
4. Análise de sentimentos.
    1. Aplicação de acervos léxicos.
    2. Métodos alternativos.
5. Análise de agrupamento.
    1. Medidas de distância e similaridade.
    2. K-médias e variações.
6. Modelagem de tópicos.
    1. Latent Dirichlet allocation (LDA em linguagem natural).
    2. Abordagens text2vec.
7. Modelagem preditiva apoiada em texto.
    1. Classificação e predição.
    2. Engenharia de características.
8. Introdução ao processamento natural da linguagem.
    1. Rotulação de partes do discurso (POS tagging).
    2. Extração de entidades.

Word2vec

- O word2vec utiliza dois modelos de rede neurais em sua arquitetura

- Modelo 1: CBOW - Este modelo é utilizado para descobrir a palavra central de uma sentença, baseado nas palavras que a cercam.
  Ex: O cachorro correu atrás do gato
  (cachorro, [O, correu])
  (correu, [cachorro, atrás])
  (atrás, [correu, do])
  (do, [atrás, gato])

- Modelo 2: Skip-Gram - da palavra central, tentaremos descobrir as palavras de contexto.
  Ex:
  ([O, correu], cachorro)
  ([cachorro, atrás], correu)
  ([correu, do], atrás)
  ([atrás, gato], do)

Gensim word2Vec

- O Word2Vec parte da premissa:
  se você tiver duas palavras que têm vizinhos muito semelhantes (onde o contexto que são usadas é quase o mesmo), então provavelmente essas palavras possuem significados semelhantes ou pelo menos estão relacionadas.

- Por exemplo: as palavras chocado, horrorizado e espantado são geralmente usadas em um contexto semelhante.

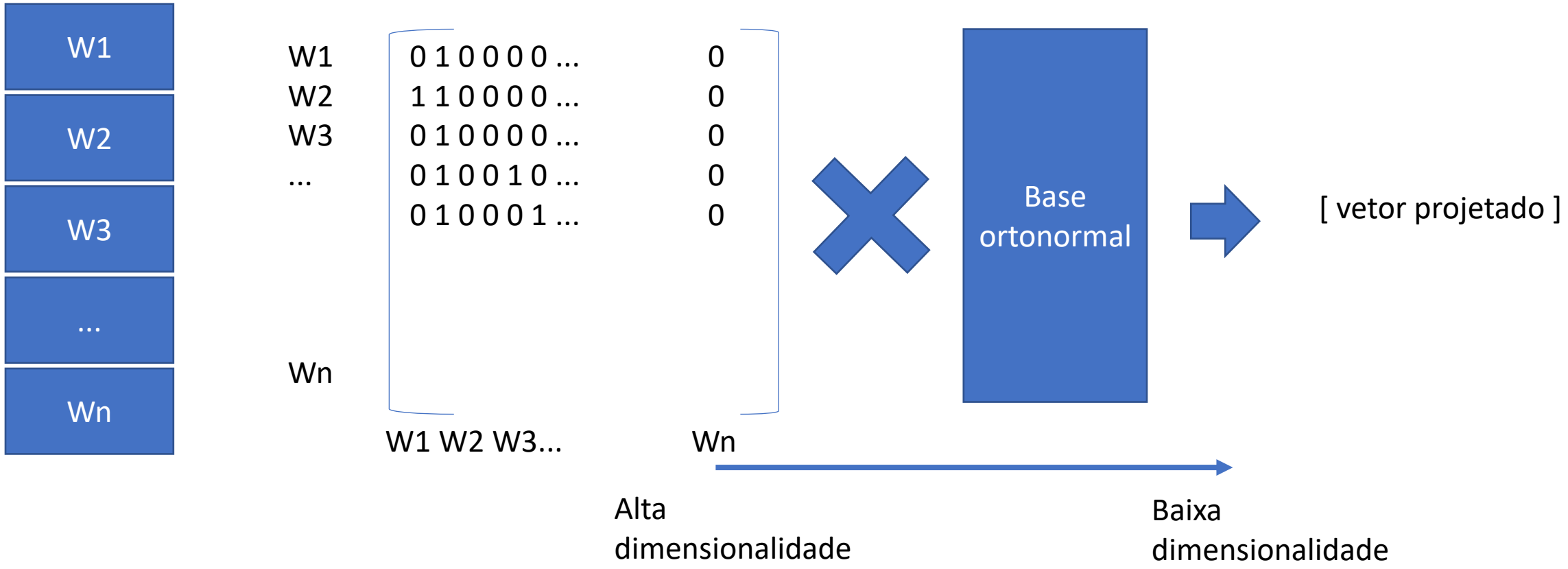
- Usando essa suposição subjacente, você pode usar o Word2Vec para:

1. Identificar superficialmente conceitos semelhantes
2. Encontrar conceitos não relacionados
3. Calcular a semelhança entre palavras.

https://radimrehurek.com/gensim/

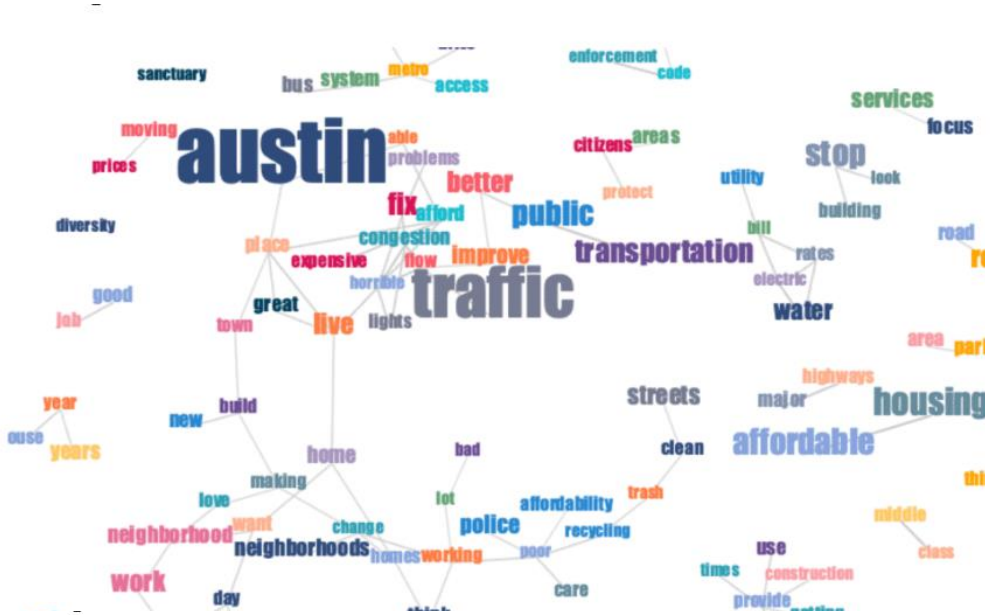
Spaced Words Projection (SWeeP)

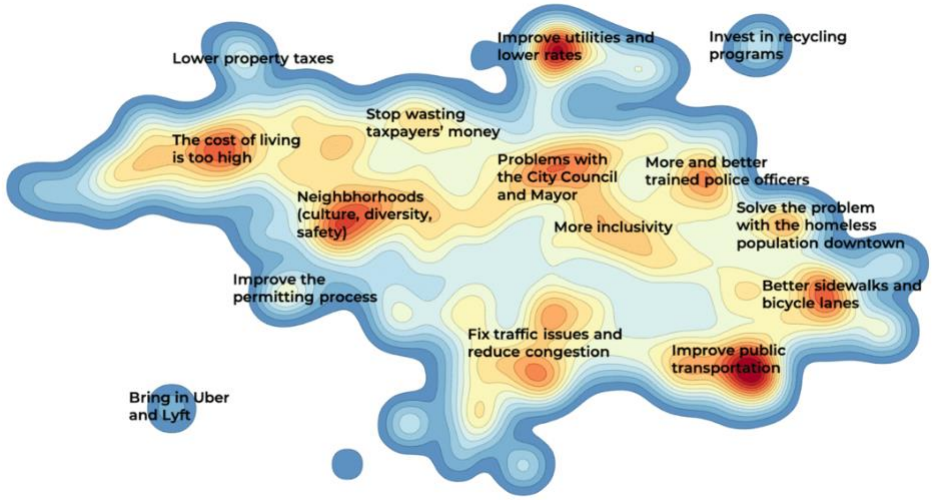
Sentenças, frases, parágrafos, textos ou documentos

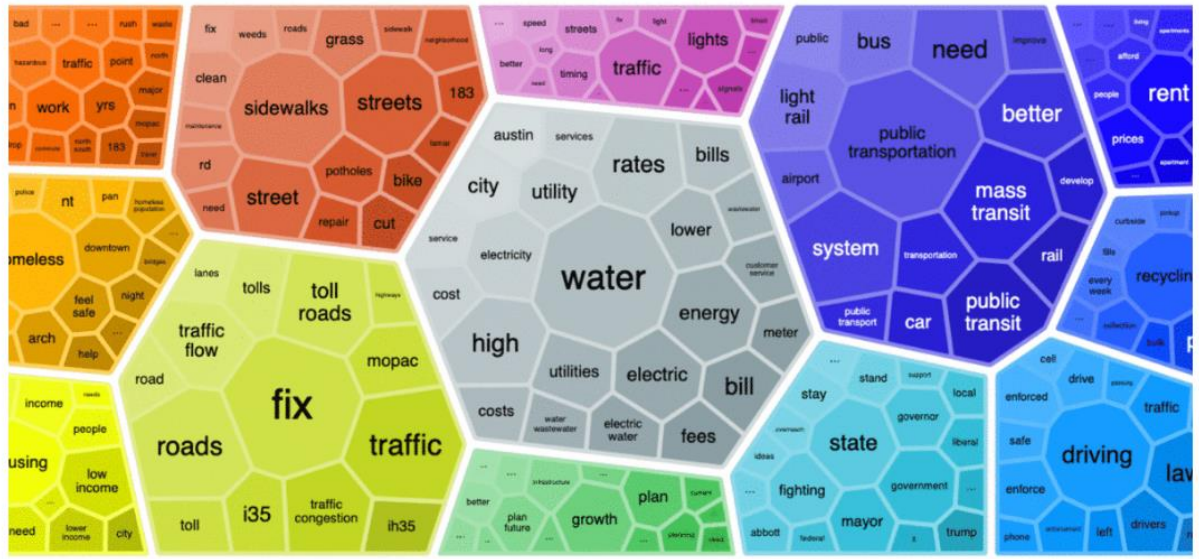


Melhor visualização com PCA...

Representações gráficas







Atividade

- Utilize um dos livros disponíveis no moodle ou qualquer outra obra com pelo menos umas 80 páginas para essa atividade.
- Faça o processo de pré-processamento, quebrando o texto em parágrafos ou sentenças e removendo stopwords, símbolos e pontuações (cuidado com nomes, personagens, lugares etc).
- O que a matriz de co-ocorrência pode nos revelar? Para duas, três e quatro palavras? (considere os 10 resultados mais frequentes).
- O que o modelo de regras pode nos revelar?
- O que o modelo vetorial pode nos revelar?
  - O que a PCA nos revela do modelo vetorial?
  - Qual a diferença de resultado de uma clusterização de k-mens para uma clusterização db-scam em cima dos resultados do modelo vetorial?

Referencias e conjunto de dados

- PLOS COMPUTER & INFORMATION SCIENCES. Text Mining:Curated Collections, 2016
- Mikolov et al, Efficient Estimation of Word Representations in Vector Space, 2013
- Gensim word2vec
  https://radimrehurek.com/gensim/
- Dois de cinco conjuntos disponíveis com o software SENNA (2011)
  http://ronan.collobert.com/senna/
- Tripadvisor and Edmunds
  https://github.com/kavgan/OpinRank/tree/master
- Alocação latente de Dirichlet
  https://pt.wikipedia.org/wiki/Aloca%C3%A7%C3%A3o_latente_de_Dirichlet

Referencias

- Anna Huang. Similarity Measures for Text Document Clustering http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.4480&rep=rep1&type=pdf
- Cláudia Freita. Sobre a construção de um léxico da afetividade para o processamento computacional do português. https://www.scielo.br/j/rbla/a/jxSZLGKJQVZgxRDVkpR9Dxn/?format=pdf&lang=pt
- Evaluating CETEMPúblico, a free resource for Portuguese https://dl.acm.org/doi/10.3115/1073012.1073070