Resumo das etapas básicas aplicadas em PLN

I - Limpeza e Pré-processamento

Basicamente, as atividades para essa primeira etapa são:

- remover caracteres especiais, números, pontuações, acentos e espaços em branco (direita e esquerda);
- remover *stopwords* (palavras de parada, em tradução livre): geralmente, são palavras irrelevantes que não adicionam significado ao processamento do texto. Podem ser artigos, conjunções, preposições ou palavras específicas dependentes do problema em questão;
- converter o texto para minúsculo;
- aplicar o processo de *Tokenization no texto* (ou tokenização, em tradução livre): quebra o texto em partes denominadas "*tokens*", os quais são utilizados como entrada para a etapa de transformação/ extração de características (feature extraction). Um texto pode ser separado em sentenças (s*entence tokenization*) palavras (*word tokenization*), caracteres ou subpalavras. Esse processo também pode remover a pontuação do texto; e
    - Exemplo>> input: `"I love my cat cat"`
      output: `{'my', 'I', 'love', 'cat'}`
- reduzir cada *token* à raiz por meio de *Stemming* ou *Lemmatizing*
    - *Stemming,* processo heurístico que simplesmente remove sufixos para gerar a forma raiz de uma palavra (*stem*)
        - Por exemplo: estudar, estudou, estudo = estud, que representa a base de todas as variações citadas.
        - É mais utilizado por mecanismos de buscas para indexar palavras . O uso desta técnica pode gerar dois tipos de problemas: (i) gerar um *stem* que não tem sentido, pois perdeu a informação durante a redução; e (ii) gerar dois *stem* iguais, porém com significados diferentes
    - *Lemmatization*, considera a morfologia da palavra e a reduz para a forma canônica/básica, conhecida como (lemma). Apesar de ser mais lenta, garante a geração de palavras gramaticalmente corretas e com maior precisão
        - Exemplo: andando, andar, andamos = anda.

II - *Feature extraction*

A etapa de "extração de características" significa extrair e produzir representações apropriadas para o tipo de tarefa de PLN realizada e o tipo de modelo de aprendizado de máquina que será utilizado.

As técnicas para essa extração podem ser classificadas em dois grupos, os quais são:

- *Frequency-based or Statistical based *text vectorization* ou "statistical word vectorization"* (primeira geração):
    - é a classe pré-era de "*word embedding*" e é composta por métodos que contam as co-ocorrências das palavras ou constroem matrizes de ponderação, não conseguem apropriadamente modelar todo o contexto ao redor da palavra;
    - *Vectorization* é o mapeamento de tokens do conjunto de dados para um vetor correspondente de números reais;
    - Exemplos: N-gram, Bag of Words (BoW), TF-IDF; e
    - Problemas: são caracterizadas pela alta dimensionalidade e por valores esparsos na representação matricial dos textos.

- *Prediction based Word Embedding* (segunda geração):
    - Métodos mais sofisticados são chamados de "*Word embedding*";
    - É uma representação vetorial de números reais que captura a relação sintática e semântica das palavras de um grande corpus* de texto; e
    - Exemplos: word2vec, GloVe, fastText, elmo, ulmfit, BERT.

**corpus são conjuntos de dados textuais legíveis e compilados com o propósito de servirem como fonte para diferentes tarefas de processamento de linguagem natural [Chiele et al. 2015]

III - Modelo

Construção de um modelo baseado em alguma técnica de aprendizado de máquina.

Observação: Algumas bibliotecas para o processamento de linguagem natural: spacY, Gensim, NLTK

Referências

[Chiele et al. 2015] Chiele, G. C., Fonseca, E., and Vieira, R. (2015). Geração de modelo para reconhecimento de entidades nomeadas no opennlp. Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Su (PUCRS).

Guia de NLP - conceitos e técnicas | Alura

https://medium.com/analytics-vidhya/traditional-text-vectorization-techniques-in-nlp-4e99218e7efe

https://medium.com/swlh/word-embedding-new-age-text-vectorization-in-nlp-3a2db1db2f5b