Tópico 1: O que é NLP?

1. Leitura inicial I: - O que é Processamento de Linguagem Natural? Medium.
2. Leitura inicial II: - What is Natural Language Processing. Oracle.
3. Leitura inicial III: - Abordagens clássicas de NLP, UFU.
4. Video ‘Code-First Intro to NLP’

Leitura inicial I: - O que é Processamento de Linguagem Natural? Medium.

Normalização

A normalização abrange tratativas como a tokenização, transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais e tags HTML(em caso de web scrapping).

- Tokenização

Remoção de *Stopwords*

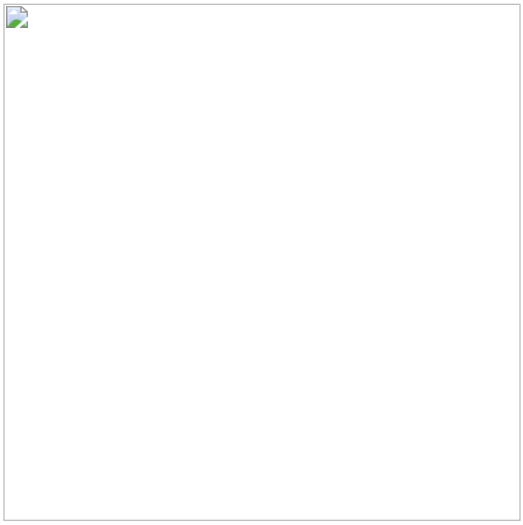
Uma das tarefas muito utilizadas no pré-processamento de textos é a remoção de *stopwords*. Esse método consiste em remover palavras muito frequentes, tais como “a”, “de”, “o”, “da”, “que”, “e”, “do”.

Remoção de numerais e símbolos que os acompanham como “R$”, “$”, “US$”, “kg”, “km”

Correção Ortográfica

Stemização e Lematização

Abordagens clássicas de NLP, UFU.



> 💡 Enquanto o nível **semântico-lexical** organiza e caracteriza semanticamente as classes verbais, o nível **sintático-lexical** prevê as possíveis configurações **sintáticas** e as alternâncias argumentais dessas classes. O que os dois níveis têm em comum e o que os relaciona é a raiz.

5.2.2 Análise léxica

Segundo [Hippisley, 2010], a análise léxica tem dois lados: o mapeamento da palavra até o seu lemma, chamado de parsing side e o outro lado é o mapeamento do lemma
para a palavra, chamado de geração morfológica (do inglês: morphological generation).
Na recuperação de informação (do inglês: information retrieval, IR), o parsing e o generation servem diferentes propósitos.

5.2.4 Análise semântica

Na linguística, a análise semântica refere-se à análise do significado das palavras,
expressões fixadas, sentenças inteiras e enunciados no contexto [Goddard and Schalley,
2010]. Na prática, isso significa traduzir as expressões originais em um tipo de metalinguagem

4. Video ‘Code-First Intro to NLP’

O estado da arte de processamento de linguagem digital é a utilização de aprendizado de máquina, especificamente redes neurais.

Historicamente, os problemas de NLP utilizavam métodos baseados em estatística dos comportamentos possíveis da linguística, como em Spell Corrector escrito por Peter Norvig.

Standart techniques:

- Stemming, Lemmatization, stop words removal
  - Não são úteis para as técnicas de redes neurais mas é interessante mantê-las. Entretanto, stop words removal por exemplo nem sempre traz benefícios no mundo real.

Norvig vs Chomsky:

Há debates entre a utilização de modelos de aprendizado de máquina para prever os outputs e a resolução de problemas utilizando mecanismos baseados em como os fenômenos se comportam.

- Chomsky: é contra a utilização de métodos de aprendizado de máquina usando apenas métodos estatísticos sem entender o sentido do comportamento
- Peter Novig: Discorda parcialmente de Chomsky, dizendo que a linguagem é um fenômeno que pode ser melhor representado por modelos estatísticos

> I believe language is such a phenomenon and therefore that probabilistic models are our best tool for representing facts about language, for algorithmically processing language, and for understanding how humans process language.
> Peter Norvig - https://norvig.com/chomsky.html

Yann Lecun vs Chris Manning

a favor de adicionar mais features linguísticas em arquiteturas neurais X não vê necessidade de extensivo feature engineering

Aplicações:

- Qualidade da informatividade/argumentação em textos
- Análise de sentimentos sobre políticos
- Utilização de meta-dados para aprimoramento de tarefas de NLP

Bias

- Tendenciosidade depende das particularidades de cada linguagem e contextos sociais, como em questões de gênero

Languagem Models

Modelos generativos como GPT-2

- Possui bias
- Possui capacidade de escrever textos que soam coerentes, mas com informações completamente imprecisas
    - Então, como filtrar informações falsas em gramática consisa escritas por modelos de linguagem?