| Tópico 2

Aula 3 — Normalização de texto

> https://www.youtube.com/watch?v=z4ZJsjHnEEE&list=PLLrlHSmC0Mw73a1t73DEjgGMPyu8QssWT&index=14

Aula 3.1: Normalização de Texto

Ferramentas de linguística computacional para estudo morfológico.

- Processamento e análise das unidades linguísticas da língua.

Normalização — conjunto de tarefas computacionais de **pré-processamento** de texto.

- Busca uniformizar os textos para uma forma mais conveniente de serem processados.
- Em geral, envolve 4 tarefas:
  - Tokenização.
  - Lemarização.
  - Radicalização.
  - Tokenização de sentenças.

Aula 3.2: Introdução à Tokenização

Tokenização é a tarefa **mais importante** da fase de normalização do texto.

- Busca **identificar** as diferentes **unidades linguísticas** de um texto.
- Unidades linguísticas → tokens.
- Forma mais comum: tokenização em palavras.

- Segmentação de uma *string* (sequência de caracteres) e em uma lista de *tokens*.

Tokenizador 1 — segmenta com base no delimitador espaço entre as palavras.

Tokenizador 2 — segmenta com base em delimitadores como espaço, pontuação e início/fim da sequência..

Aula 3.3: Tokenização em Sub-Palavras

- *Token vs. type*
  - *Token*: conjunto de unidades linguísticas dentro de um texto.
  - *Type*: conjunto de unidades linguísticas **únicas** dentro de um texto.
- Pode-se analisar a variedade linguística de um córpus comparando-se a contagem de tokens e types (razão entre contagem de types e tokens).
- Limitação: alto esforço manual para definição das regras.
- Tokenização em sub-palavras — segmentação de palavras raras em sub-palavras frequentes.
  - Baseado em dados (*data-driven*), i.e., são treinados.
  - Vocabulário mais enxuto.

Aula 3.3.1: Tokenização: Byte-Pair Encoding (BPE)

Treinamento influenciado em sistemas de compressão de arquivos.

Treinamento:

1. Segmentação das palavras em cada texto do córpus.
   - Uso de um pré-tokenizador, geralmente, baseado em regras.
2. Criação de uma lista de tokens únicos e suas respectivas frequências.
3. Criação de vocabulário base com todos os símbolos que compõem os tokens.
4. Aprendizado de regras de junção entre 2 símbolos até que um vocabulário do tamanho desejado seja obtido.
   - Tamanho é um valor estático definido anteriormente.
   - O par de símbolos unidos são aqueles que apresentam as maiores frequências.

Aula 3.3.2: Tokenização: Byte-Level BPE

- Limitação do BPE: possibilidade de ausência de alguns símbolos no vocabulário.
  - Solução: inclusão de todos os caracteres unicode no vocabulário, sendo esses representados como bytes.

- Vocabulário enxuto sem nunca obter alguma unidade linguística que não está contida no vocabulário.
    - Correção das deficiências presentes nos tokenizadores baseados em regras.

Aula 3.3.3: Tokenização: WordPiece

- Semelhante ao BPE.
- Difere na forma que realiza a junção de símbolos:
    - A escolha do par de símbolos (x1 e x2) que será combinada é feita buscando maximizar a seguinte probabilidade no conjunto de treinamento:

$$\frac{P(x_2|x_1)}{P(x_2)}$$

Aula 3.3.4: Tokenização: Unigram e SentencePiece

Aula 3.4: Capitalização e Tokenização de Sentenças

Aula 3.5: Lexicalização e Radicalização

- Lexema — unidade (abstrata) de significado que corresponde a um conjunto de formas relacionadas.
    - Exemplo: conjunto de conjugações de um verbo.
- Lema — forma canônica (dicionarizada), escolhida por convenção para representar um lexema.
    - Exemplo: verbo no infinitivo.
- Raiz — morfema básico, sem afixos derivativos ou flexionais.
    - Ao se considerar apenas a raiz, o vocabulário pode ser reduzido.
    - Redução da esparcidade do vocabulário.
- Processos automatizados:
    - Lematização — dado o conjunto de lexema, obtêm-se o lema correspondente.
    - Radicalização (stemming) — dado o conjunto de lexema, obtêm-se a raiz correspondente.

2.4 — Text Normalization

> https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

2.4.2 Word Tokenization

Tokenizadores mais sofisticados devem ser capazes de tratar numerais e pontuação em vez de apenas removê-los, como é feito por algoritmos mais simples.

- Esses símbolos são uma fonte de informação útil que dá significado a certas partes do texto.

Expressões com múltiplas palavras devem ser tratadas como tokens únicos em casos como nomes por exemplo.

- Por isso, o reconhecimento de entidades nomeadas é importante.

Exemplo de tokenizador padrão muito utilizado:

- Penn Treebank Tokenization

Algumas línguas realizam a separação de suas palavras de forma diferente, enquanto língua latinas e anglo-saxônicas utilizam o espaço como separador, outras como mandarim, japonês e tailandês definem palavras como conjuntos de caracteres básicos, sem separação por espaços.

- Assim, há várias formas de realizar a tokenização mantendo o significado do texto.
- Deve-se considerar se é melhor considerar caracteres ou palavras como entrada, já que em algumas dessas línguas já tem caracteres com um nível considerável de semântica.