UiT The Arctic University of Norway

Natural Language Processing

Week 5: Text Preprocessing

Shayan Dadman - Phd Candidate, UiT Narvik
Room: D3430
Email: Shayan.dadman@uit.no

Andreas Dyrøy Jansson - Phd Candidate, UiT Narvik
Room: C3190
Email: Andreas.d.jansson@uit.no

Department of Computer Science and Computational Engineering

Text Preprocessing

- Data preprocessing is as essential as model building.
- It is even more important for text data due to unstructured nature of it.
- Some common steps include:
  - Lower casing
  - Punctuations Removal
  - Stopwords Removal
  - Frequent words Removal
  - Rare words Removal
  - Emoji Removals
  - Emoticons Removal
  - Conversion of emoticons to words
  - Conversion of emojis to words
  - URLs Removal

Text Preprocessing

- Lower Casing:
  - It is the most common text preprocessing step.
  - The idea is to convert the input text intop same casing format.
  - This helps the model to treat all texts in the same way.
  - Nevertheless, it may not be helpful for tasks like:
    - Part of Speech tagging
    - Sentiment analysis
- Punctuations Removal:
  - This is part of the text normalization process.
  - For instance, it helps the model to treat *Hi* and *Hi!* The same.
  - Here are the punctuation symbols in python:
    - !"#$%&\'()*+,-./:;<=>?@[\\]^_{|}~`
  - We can add or remove based on the task

Text Preprocessing

- Stopwords Removal:
  - Stopwords are commonly occuring words like 'the', 'a', etc.
  - They can be removed, as they don't provide valuable information for analysis.
  - However, they are required for Part of Speech tagging task.
- Frequent Words Removal
  - It is the removing process of words that appeared frequently in text
  - It is common in domain specific corpus
- Rare Words Removal
  - Same as previous step, but for the rare words in the corpus

Text Preprocessing

- Emojis Removal:
  - Depends on the task, for textual analysis it is recommended to remove the emojis.
- Emoticons Removal:
  - Emoticons are different from emojis.
  - Emoticon represent a facial expression by putting characters together.
  - However, emoji is an actual image.
  - Again, based on the task, we may remove the emoticons.
- Conversion of Emoticons/Emojis to Words:
  - In case of sentiment analysis, emoticons/emojis provide valuable information.
  - In this case, instead of removing them we can convert them to words.
- URLs Removal
  - Simply removing the URLs from text for further analysis.

Text Preprocessing

- You can find the jupyter notebook of the text preprocessing steps in canvas.
  - We used Twitter dataset for this demonstration.