Week 2 Research Stay

Context

This week's focus was on how text should be pre-processed so that the model can give accurate results.
This pre-processing includes how to divide the text into tokens and how they should be handled, as well as the considerations that must be taken in order to generate good pre-processing of the data.

Search Metodology

I reviewed the references in the readings, read the abstracts of the ones that caught my attention, and chose the one that was most closely related to the three readings.

Week 2 Research Stay

Preliminary Terms

- *Token*: How to "break" a text into words or phrases to make it processable for the model.
- *Bag-of-Words (BoW)*: Text representation model that counts the frequency of occurrence of words.
- *Corpus*: A collection of text documents representing a specific variety of language

Week 2 Research Stay

Comparison [1]

In the readings provided to us, the common topic is how a text should be pre-processed, what practices should be followed and when to use them, and that many of these techniques depend closely on the context and nature of the text in order for the model to produce accurate results.

Some of these preprocessing techniques include removing punctuation or infrequent terms, lowercasing, stemming and lemmatization, among others.

The use of these techniques depends on the context of the corpus, since the way in which these data are initially treated can profoundly change the results of the model.

Week 2 Research Stay

Comparison [2]

There is no universal preprocessing technique, but there are tools that help in this process.

An example of these tools is preText, which is mentioned in Reading 3 (“Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It”).
Evaluate the sensitivity of analysis results to different preprocessing techniques. It provides a more structured and systematic assessment of how preprocessing affects results, helping to identify optimal configurations and avoid decisions that may distort data.

The article I found mentions another technique (NPFST) to improve pre-processing, this seeks to be an improved version of BoW, which only counts the frequency of words in their simplest form without counting the continuity of certain words. “NPFST uses a part-
of-speech tagger and a finite state transducer to extract multiword phrases to be added to a unigram BOW.”

References

- What is a word, What is a sentence? Problems of Tokenization
- Text preprocessing for text mining in organizational research
- Text preprocessing for unsupervised learning

- Bag of What? Simple Noun Phrase Extraction for Text Analysis
- Implementation of NPFST: http://slanglab.cs.umass.edu/phrasemachine/