Final assestment NLP 2

Prof. Emmanuel Páez

October 9, 2023

1 Instructions

This is the practical evaluation assessment of the course, please read each section carefully, and remember this must be delivered through Canvas and it is required to be completed by class teams. You must upload one notebook file per team (".ipynb"). You must include a header for each section and you must all the code used to complete the assessment.

2 Tasks (total 100 pts)

2.1 Corpus creation (25 pts)

Select any source and theme you like to create a corpus, and provide information about the method you used (web scrapping, data sets, files, etc).

Also, provide some context about the text, then tokenize your corpus and clean the tokens to create some descriptive statistics (frequency analysis, collocations, word cloud, etc).

Describe your findings and explain in detail every text-processing task in your data pipeline.

2.2 Embeddings (25 pts)

By using any model you prefer to create a vector embedding representation of your vocabulary.

Include information about the model you use for embeddings (word2vec, glove, bert, etc), include plots and figures about the semantic inferences you find in your corpus vocabulary after using embeddings.

2.3 Clustering (25 pts)

By using any unsupervised method you prefer (nearest neighbor for example), analyze your embedding vocabulary in order to find some clusters of relevant topics.

Include information about the model you choose and your findings, and include any plot you like to explain.

2.4 Zero shot model (25 pts)

Research fine-tuned "zero-shot models", choose a pre-trained "zero-shot model" (you can use a hugging face) and include some information about it: Who trained? How it was trained? and include some instructions on how to use it.

Propose some classes (topics) for your corpus and process each document with the "zero-shot model" you choose, include your findings and comments for example: Why do you choose those classes? what is the relation between zero-shot predictions and the clusters you found in the 2.3 task?, Include any plot or image to explain your findings.