NLP &Text Mining
Project

The goal of this work is to process a text dataset using word text embedding and data analytics methods in order to extract knowledge from it. Prepare a report for this work and deposit it on moodle.

In this work you will use 20 Newsgroup dataset, but you a free to use any text data (UCI datasets repository, kaggle, data.gouv.fr, …) informing the Professor. It is better if you use another dataset.

The work should contains at least the following 6 parts:

1. **Analysis of the text dataset**
2. **Text processing and Transformation**
3. **Apply different embedding techniques**
4. **Clustering and/or classification on the embedded data**
5. **Results analysis and visualisation**
6. **Theoretical formalism**

1. Analyse the dataset : the context, size, difficulties, detect the objectives.

2. Text Processing and Transformation
For this part, you should use scikit-learn and you can follow the tutorial:
https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html#tutorial-setup

A. Extracting features from text files

In order to perform machine learning on text documents, we first need to turn the text content into numerical feature vectors using:

- Bags of words:

The most intuitive way to transform the text into a vector is to use a bags of words representation:

1. Assign a fixed integer id to each word occurring in any document of the training set (for instance by building a dictionary from words to integer indices).

2. For each document #i, count the number of occurrences of each word w and store it in X[i, j] as the value of feature #j where j is the index of word w in the dictionary.

B. Tokenizing text with scikit-learn
Text preprocessing, tokenizing and filtering of stopwords are all included inCountVectorizer, which builds a dictionary of features and transforms documents to feature vectors.

C. Preprocessing with Lematization, Stematization, and others techniques.

3. Apply different embedding techniques

Occurrence count is a good start but there is an issue: longer documents will have higher average count values than shorter documents, even though they might talk about the same topics.

To avoid these potential discrepancies it suffices to divide the number of occurrences of each word in a document by the total number of words in the document: these new features are called TF for Term Frequencies.
Another refinement on top of TF is to downscale weights for words that occur in many documents in the corpus and are therefore less informative than those that occur only in a smaller portion of the corpus.
This downscaling is called TF_IDF for "Term Frequency times Inverse Document Frequency".
Both TF and TF-IDF can be computed as follows using TfidfTransformer
You should test different embedding approaches as:

- word2vec,
- document2vec,
- Glove,
- BERT,
- …

4. Clustering and/or classification on the embedded data
Use a machine learning method (clustering and/or classification) in order to predict the class of a new set of objects. You can use the methods as K-Nearest Neighbours (K-NN), Support Vector Machine (SVM), Decision trees, Neural Networks ... The obtained results should be validated using some external indexes as Prediction Error or others. The obtained results should be analysed in the report and provide a solution to ameliorate the results. If the class information is not available, use a clustering method as K-means or/and hierarchical Clustering.

5. Results analysis and visualisation
Analyse the obtained results i.e. validation indexes and compare between different embedding methods. You can visualise the knowledge extracted from the classification/clustering in order to present the results i.e. scatter plots using predicted colors,…

6. Theoretical details:
Give the algorithmically (mathematical) formalism of the embedding method which give the best results. Explain all the parameters of the used method and their impact on the results.
Some comparison should me made to conclude the project.