2nd Year Second Cycle Option: IASD
Module : Natural Language Processing

Practical work N° 02: Feature extraction and embeddings

Objective :

➔ Familiarize with vectorization methods.

A. Data Preparation

Import the dataset **spooky.csv** that you have already preprocessed (the final version after preprocessing).

B. Encoding of the Target Variable

Encode the labels using an encoding technique.

C. Construction of Training and Testing Sets

1. Split the dataset into two parts: training and testing, using train_test_split, with a test size of 30% and random_state=0.
2. The dataset is imbalanced, stratify the samples to achieve a similar distribution in each class of the dataset.

D. Vectorization Methods

1. Use the lexical frequency method and one-hot encoding to vectorize the training and testing datasets.
2. Train a TF-IDF vectorization model on the training part and vectorize it.
3. Using the same model, vectorize the testing part.

E. Training

1. Create three models of the MLPClassifier type. (You can change the learning algorithm: use other scikit-learn algorithms)
2. Train these three models on the three vector representations.
3. Predict the classes by applying the three models to the three training representations.
4. Display the classification report using performance measures (accuracy, precision, recall...).

F. Testing

1. Predict the classes by applying the three models to the three testing representations.
2. Display the classification report using performance measures (accuracy, precision, recall...).
3. Calculate the training time for each model.

G. Vectorizations based on word embeddings

1. Use techniques of vector representation based on word embeddings: a. Word2Vec (CBOW and Skip gram) b. Glove c. FastText.

H. Training / Testing

1. The same questions as sections E and F but with the new vector representations presented in section G.
2. Compare all models made in this practical work.