Université Abdelmalek Essaadi
Faculté ses Sciences et techniques de Tanger
Département Génie Informatique
Master : AISD
NLP
Pr . ELAACHAk LOTFI

Lab 3

Objective : The main purpose behind this lab is to get familiar with NLP language models using Sklearn library.

Work to do :

Part 1: Language Modeling / Regression

Dataset:https://github.com/dbbrandt/short_answer_granding_capstone_project/blob/master/data/sag/answers.csv

1. Establish a preprocessing NLP pipeline (Tokenization stemming lemmatization, Stop words, Discretization, etc) of the collected Dataset.

2. Encode your Data vectors By using Word2vec (CBOW, Skip Gram), Bag Of words, TF-IDF.

3. Train your models by using SVR, Naive Bayes, Linear Regression , Decision Tree Algorithms (The embedding will be done by Word2Vec).

4. Evaluate the four languages models by using standards metrics (MSE , RMSE, etc), choose the best model then argument your choice.

5. Interpret the Obtained Results.

Part 1: Language Modeling / Classification

Dataset: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

1. Establish a preprocessing NLP pipeline (Tokenization stemming lemmatization, Stop words, Discretization, etc) of the collected Dataset.

2. Encode your Data vectors By using Word2vec (CBOW, Skip Gram), Bag Of words, TF-IDF.

3. Train your models by using SVM, Naive Bayes, Logistic Regression, Ada Boosting Algorithms (The embedding will be done by Word2Vec).

4. Evaluate the four languages models by using standards metrics (Accuracy, Loss, F1 Score, etc) and other metrics like blue score, choose the best model then argument your choice .

5. Interpret the Obtained Results.

Notes :

- **At the end each student must give a brief synthesis about what he has learn during the proposed lab.**
- **Push the work in the Github repository and write a brief report in Github readme file.**

Tools:

Google colab or Kaggle, gitlab/github, spacy , NLTK, Sklearn.