Faculty of Computers and Information
Cairo University
NLP-2019

NLP
Assignment-3
Performing text classification using Embedding Models

Objectives

- Practice how to perform text classification using a machine learning classification model and combinations of word embeddings or sentence embeddings as a feature vector

Dataset

Movie reviews data set V2.0 contains 2000 text samples divided into 1000 positive reviews and 1000 negative reviews. Reference and download:
http://www.cs.cornell.edu/people/pabo/movie-review-data/

Task:

- Load review samples (both positive and negative) and generate **sentence embedding vector** for the samples. You are required to use **TWO** of the embedding methods below:
    a. Sentence embedding based on the sum of word embeddings
    b. Sentence embedding based on the average of word embeddings
    c. Sentence embedding using doc2vec model
- Generate labels vector for the dataset.
- Randomly divide data to training and testing sets. Note that each set should contain samples of the two types
- Train a classification model to predict the label of the review

Output

- Comparison report for the results of model (1) from assignment 2 and models (2) and (3) from assignment 3.

Teams: Form teems of 3 for this assignment

Needed References:

- Glove dictionary of word embeddings
  https://nlp.stanford.edu/projects/glove/
- ML library
  https://scikit-learn.org
- Doc2vec
  https://radimrehurek.com/gensim/