HW04 – ISIS 4221

Natural Language Processing 2023

Due date: 24-10-2023

Coding rules: Use jupyter notebooks and be sure that the notebook is executed and contains the results before submitting. All classes, methods, functions and free-code MUST contain docstrings with a detailed explanation. Build a notebook for each point.

Report: Together with the notebooks, you must submit a written report (please use pdf format) with the answers to the questions and a short summary of the implementation.

Submission: individually or in a group of two. Make a zip that includes all the files.

Datasets

- **Books from Gutenberg project https://www.gutenberg.org/**

I. Choose three authors whose works are available in the Gutenberg project. For each of these authors, it is imperative to carefully select and download a minimum of three books, totaling at least nine books in all. Utilize these selected literary works to train word embeddings using the GENSIM library and the word2vec model.
   - Prepare the training dataset using appropriate text preprocessing steps.
   - Try different embeddings dimensionalities (at least 3) and save them to disk using appropriate GENSIM methods:
     - Books_<size_1>_<group_code>
     - Books_<size_2>_<group_code>
     - Books_<size_3>_<group_code>

II. Investigate and explain a strategy for plotting embeddings in two dimensions. Plot the most similar words to the main characters names in each book.
   - Find interesting relationships using analogous reasoning.

III. You are going to build a classifier to identify the most likely author for a set of input lines of text (I suggest utilizing text segments comprising 150 to 250 words). It is a multinomial classification task (3 classes).
   - Describe how you prepare the dataset. Create the training, validation, and testing sets. Make a summary table with the dimensions (number of samples) by class for each one of the previous data sets.
   - Define three feed-forward (dense) neural network architectures in Keras that make use of the previously built embeddings.
     - Explain the dimensions of each layer of each architecture (model summary).
   - Describe the results of combining the 3 architectures with the 3 types of embeddings in terms of accuracy, precision and recall in tests set.

IV. Repeat III but instead of using your embeddings, use the Google-Word2Vec or Glove pre-trained embeddings with different dimensionalities (at least 3). You can download these embeddings from different sources like Gensim data repository (https://radimrehurek.com/gensim/models/word2vec.html#pretrained-models), Stanford Web Page (https://nlp.stanford.edu/projects/glove/), or TensorFlow Hub.

V. Compare the results, and answer the following questions:
   - Is it better to use pretrained embeddings? or custom embeddings on the corpus?
   - According to the results, what influence does the dimensionality of the embedding have? Analyze the two cases: when you use pre-trained embeddings and when you use customized embeddings.