HW03 – ISIS 4221

Natural Language Processing 2021-I

Due date: 10-05-2021

Groups are allowed up to a maximum of 3 students or 4 only if they are the same project group. Individual work is also allowed.

Coding rules: Use jupyter notebooks and be sure that the notebook is executed and contain the results before submitting. All classes, methods, functions and free-code MUST contains docstrings with a detail 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: Assignments are submitted via Brightspace. Do not email us your assignments. Please upload all files and documents.

Datasets

- **Simpsons Dialogues: https://www.kaggle.com/pierremegret/dialogue-lines-of-the-simpsons**
- **Friends Dialogues https://www.kaggle.com/blessondensil294/friends-tv-series-screenplay-script.**

PLEASE READ DATASET DESCRIPTIONS

You can download all datasets from:

https://www.dropbox.com/sh/trzd0mv7orvi0xi/AACql2T1-LA89suzIlwCMPLa?dl=0

[50p] Simpsons Dialogues

I. Using the Simpsons dataset build words embeddings using GENSIM and word2vec.
   - Prepare the dataset using only the dialogs.
   - Using appropriate text preprocessing steps.
   - Try different embeddings dimensionalities (at least 3) and save them to disk using appropriate GENSIM methods:
     - Simpsons_<size_1>_<group_code>
     - Simpsons_<size_2>_<group_code>
     - Simpsons_<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.
   - Find interesting relationships using analogous reasoning.

III. You are going to build a classifier to identify the most likely character for an input line of dialogue. It is a multinomial classification task. Only use the main characters.

- 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 neural network architectures in Keras that make use of the previously built embeddings.
  - Explain the dimensions of each layer of the architecture.
- Describe the results of combining the 3 architectures with the 3 types of embeddings in terms of accuracy, precision and recall in training, validation, and testing sets.
  - Explain how you implement precision and recall in Keras.

IV. Repeat III but instead of using GENSIM embeddings, train the embeddings in the neural network architectures.

- Use three different vectorize_layer output_modes so you will have three different embeddings.

V. Compare the results obtained in III and IV.

[50p] Friends Dialogues

I. Using the Friends dataset build words embeddings using GENSIM and word2vec.

- Prepare the dataset using only the dialogs.
- Using appropriate text preprocessing steps.
- Try different embeddings dimensionalities (at least 3) and save them to disk using appropriate GENSIM methods:
  - Friends_<size_1>_<group_code>
  - Friends_<size_2>_<group_code>
  - Friends_<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.

III. You are going to build a classifier to identify the most likely character for an input line of dialogue. It is a multiclass classification task. Only use the main characters.

- 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 neural network architectures in Keras that make use of the previously built embeddings.
  - Explain the dimensions of each layer of the architecture.
- Describe the results of combining the 3 architectures with the 3 types of embeddings in terms of accuracy, precision and recall in training, validation, and testing sets.
  - Explain how you implement precision and recall in Keras.

IV. Repeat III but instead of using GENSIM embeddings, train the embeddings in the neural network architectures.

- Use three different vectorize_layer output_modes so you will have three different embeddings.

V. Compare in details the results obtained in III and IV.

NOTE: please experiment with the HPC.