SML 310 Mini-Mini-Project #3 -- NLP and Neural Networks Spring 2022

The following mini-project is focused on sharpening your Data Science skills in the context of Natural Language Processing and Neural Networks! Your writeup should be submitted as a well-documented Python / Jupyter Notebook by e-mail to sulinl@princeton.edu and jonhanke@princeton.edu by 10pm (EST) on Wed April 20th, 2022.

Collaboration: All class projects are required to be solely your work, with no collaboration on modelling or design choices, and you will be required to certify this when you submit your assignment. If you want to ask others about or share general references, please feel free to post them on Slack where all comments are clearly visible.

References: Please feel free to use online resources (with clear inline citations in your code) to better understand the process of using your basic data science tools (Python / Jupyter / Pandas / Scikit-Learn) to perform a specific task (i.e. make a scatter plot, fill in missing values, etc.), but ***you may not consult*** online references specific to your given dataset, or solicit/receive comments outside of course discussions about the choices needed to perform EDA or modelling specific to your dataset. When in doubt about a reference, please feel free to ask on our Slack channel, or during Class / Precept / Office Hours.

For these projects you may find some of our Precept discussions and walkthroughts useful!

Mini-Project Questions:

1. NLP and Word2Vec (30%)

    a. Please describe what is meant by a “vector embedding” of words in Word2Vec.

    b. Please use the pre-trained gensim `'glove-wiki-gigaword-50'` Word2Vec model to determine reasonable synonyms for the following words:
        i. Tiger
        ii. Awesome
        iii. Song
        iv. Data

    c. Please use the pre-trained gensim `'glove-wiki-gigaword-50'` Word2Vec models to determine reasonable answers for the following analogies:
        i. puppy : kitten :: dog : ?
        ii. freshman : sophomore :: junior : ?
        iii. brother : sister :: grandson : ?

2. NLP and Topic Modelling (30%)

    a. Please prepare the built-in "`fake-news`" corpus of text using the commands:

    ```
    import gensim.downloader as api
    corpus_data = api.load("fake-news")
    docs = [x['text']  for x in corpus_data]
    ```

    From the Gensim github website at:
    "`https://github.com/RaRe-Technologies/gensim-data`"

    b. Please use Gensim to preprocess these documents by tokenizing and lemmatizing them and removing other small text/strings that you decide are not meaningful for NLP, as well as rare and scarce words.

    You may find the following Gensim tutorial useful:
    https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html

    c. Please use the Gensim `Dictionary` and `dictionary.doc2bow` to create a dictionary and a bag of words representation of your tokenized corpus.

    d. Please use the Gensim `LDAModel` to perform topic modelling of the corpus into 3 topics.

    e. For each topic show the main words and use these to give a rough name to each topic.

3. Data Classification (40%)

    a. Please take the MNIST training dataset and split off the last 10,000 images as a validation data set. Then assign the other training images to a new training set.

    b. Please create a Neural Network with three dense hidden layers each having 32 nodes, and train it to classify the MNIST Data set over five epochs. For this please use your new training set and validation set respectively for model training and model performance reporting.

    c. What is the performance of the model on the training and validation sets over the 5 epochs? Which of these do we expect is characteristic of model performance in new data?
    ss

    d. Check your model performance on the test set, and compare with your expectations in part c.