Penn Data Science Group NLP Project

This document lays out a syllabus for a collaboration between Christopher Stewart and the Penn Data Science Group (PDSG). The goal of the collaboration is to work together on a project that will allow PDSG students to get hands-on experience with Natural Language Processing (NLP) methods and Machine Learning (ML) techniques similar to that required in industrial NLP applications.

Our proposal is to first work with the Kaggle Quora Question Pairs dataset to gain familiarity with linguistic structure and basic NLP techniques. This will cover 2 weeks in our curriculum, after which time we will use the expertise gained and code written to participate in a similar Kaggle competition. We anticipate dedicating a month of classes to that effort, making biweekly submissions and charting the progress in our model's accuracy. Note that this may have to be adjusted based on our progress and the availability of upcoming Kaggle competitions. Students should be prepared to invest **minimally** 5 hours a week outside of meetings.

At the end of the collaboration, students may choose to continue development of the model. Deliverables include experience with the world's foremost Data Science competition platform, a repository of code showcasing cutting-edge NLP skills and possibly some prize money!

Provisional Timeline

1. Week 1: Background 1
    a. Reading / Homework:
        i. Work through the Python, Machine Learning and SQL Kaggle Learn codelabs
        (AND/OR)
        ii. Read and work through the exercises in the "Language Processing and Python" chapter of the Natural Language Toolkit online book

    b. In class
        i. Getting started with NLP
            1. Find or write a Jupyter notebook going through the basics of NLP analysis with data from the Kaggle Quora Question Pairs competition, something like this

2. Week 2: Background 2
    a. Reading / Homework
        i. More advanced topics in NLP: NLP Best Practices

        ii. Read about the dataset, kernels and winning solution in the Kaggle Quora Question Pairs competition

    b. In class
        i. Work through a real submission (here) in preparation for our upcoming competition.

3. Week 3: Competition (Exploratory Data Analysis)
    a. Reading / Homework
        i. Fork and work through a notebook here and script here.

    b. In class
        i. Go through data in new competition together with an eye towards developing an intuition for the new dataset, thinking of next steps

4. Week 4: Competition (Feature Engineering)
    a. Reading / Homework
        i. Read up on a broad overview of feature engineering for NLP here, get some hands-on coding experience by forking this notebook, and get an idea of the offerings of Stanford NLP's CoreNLP tools

    b. In class
        i. Use intuitions from previous week to comb through dataset, looking to build features and transform data to make an optimal model

5. Week 5: Competition (Model Building)
    a. Reading / Homework
        i. Reading: "Deep Learning Applied to NLP" and "Deep Learning for NLP (without magic)"

    b. In class
        i. Roll up our sleeves and get to the modeling! Try to come to class with some good ideas for how to proceed and we'll work together on a joint submission

6. Week 5: Competition (Model Building II)
    a. Reading / Homework
        i. Fork the class' notebook from the previous session and work on improving our score using everything that you have learned up until now.
    b. In class
        i. Continue working to better our model, making submissions periodically and finishing up whatever needs to be completed.