CS 410 Final Project Proposal

Group: LastMile

Requirements

In your proposal, please answer the following questions:

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

I will be working on the project by myself.

- Winston Zhu (hezhiz2@illinois.edu)

2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

Free topic: Sentiment or Hate Speech classification on Twitter data

Description: The main objective is to classify the twitter text dataset regarding a realistic label. The label is multi-level and the input is pure text input, similar to what we worked on in MP2.2 or MP2.3. For example, given the Twitter text, we will identify the sentiment associated as being positive or negative. Positive and negative can be determined either via cross-comparing the positive or negative words by searching from an existing word dictionary (bad of words) or training and developing a neural network. The majority of the knowledge will come from the lectures in Week 12. I am planning to apply 1 or 2 models and compare the results for sentiment classification and perform a basic data visualization to the result.

Task: I define the tasks into 3 parts, full stack framework, model construction and visualization.

- Because I want to build a full stack, local product for users to interact with. I will start working on a full stack framework. Ideally this is done by flask or django, written in python.
- The most important part comes to the model. I am planning to use one we learnt in class, such as Naive Bayes or LSTM(not learnt but mentioned in reading

material) and another one using neural networks, such as CNN (convoluted neural network).

- The last part I am planning to demonstrate the result using some of the evaluation methodologies we learnt in class, such as F1 score, average precision. Then I will create visualizations to compare different models.

Why is it important or interesting: Nowadays people usually publish what they want on Twitter, given the freedom of speech. This is an advantage of information society, however, sometimes people offend other people or groups of people. This is where the fight begins and everyone has seen it while surfing the internet. I want my model to classify the sentiment or hatred of a text so that people at least get some alarm before publishing. Specifically as an Asian, I have seen the anti-Asian twits on the Internet, especially last year.

What is your planned approach: First collect public labeled data, then do data cleaning and curation. The formalized data will then be fed into the model and we will use the trained model to test on some new data. We do this in an iterative manner and we document the results for different models.

What tools, systems or datasets are involved:

- Flask (web framework)
- NLTK and MetaPy (Stemming and Tokenization)
- Python autocorrect (Data Cleaning)
- Panda (data structure)
- Tensorflow (optional)
- Dataset: https://www.kaggle.com/vkrahul/twitter-hate-speech
- Dataset2: https://www.kaggle.com/kazanova/sentiment140

3. Which programming language do you plan to use?

- Python
- HTML in frontend
- Some bash code for pre-defined script

4. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

- Report writing (4 hours in total)
- Data collection (2-3 hours)
- Data cleaning and curation (5 hours)

- Model implementation, include training, testing and model comparison (5 hours)
- Data visualization and presentation (5 hours)