- Overview

Similar to competitions held by Netflix to refine their recommendation system or Zillow to better their appraisal system, OpenClass is sponsoring a competition to better their question tagging system. We will be provided 200 open linguistic questions to be tagged as 1 or more of several topics. Scoring will be based off of $F_2$ score. The full project description can be found [here](here).

- Motivation

The motivation for this assignment comes from two avenues. The first comes from the promise that I see for a platform like OpenClass and its potential. Mastery is not often sought from many classroom settings and so anything that can integrate usefully into a classroom setting and focus on mastery is worth pursuing. Being a part of that development is just as noble. The other avenue is the potential $100 cash prize that would be useful to a graduate student like myself.

- Anticipated Challenges

The first and largest challenge of this project will be the small volume of training data at our disposal. Since we are only provided 200 questions for tagging, we won't have the opportunity to train a particularly expressive model. This problem is further exacerbated by the need to tag multiple tags per question rather than just one. Having a more expressive model is useful in that endeavor. One potential solution would be to use feature selection to determine which features contribute the most to each tag and perhaps weigh those more heavily. Looking into auxiliary data sets might also prove useful.

A second challenge is one inferred from the project description where it gives that each question contains a name, an instructor solution, a follow-up multiple-choice question name, between 2-5 options for the follow-up multiple-choice question. This seems like quite a lot of information to parse through to determine tags. The solutions proposed to the problem above would also be a good fit for trying to solve this problem.

- Planned Approach

I will be implementing this question tagging system using an artificial neural network constructed using the [Keras](Keras) library. I will also use the [GloVe](GloVe) embedding of words to represent the tokens of each question as vectors. This was very useful in other NLP multi-tag tagging application that I have worked on in the past.

Using a library like [SpaCy](SpaCy) may also be useful in determining other aspects of the text that aren't visible just in their normal form such as part-of-speech tags.

Finally, using a stemmer or lemmatizer such as [snowball stemmer](snowball stemmer) or SpaCy's implementation would be useful in reducing the variation in the corpus of words that

we will generate (participate and participated would be considered the same word) thus reducing the level of expressiveness that may be needed in the model.

- Delegation of Responsibilities

I will be working on this project by myself.

- Timeline

| Date | Task |
|---|---|
| 2/23/20 | Data Collection Completion |
| 3/5/20 | Parsing, Tokenization, Embedding of Text Completion |
| 3/23/20 | Baseline System Completed |
| 3/23/20-4/30/20 | Tweaking and Refinements |
| 3/23/20-4/10/20 | Testing with Parts of Speech included |
| 4/10-20-4/20/20 | Testing with Auxiliary Dataset |
| 4/30/20 | Final Code Submission to Completion |
| 5/5/20 | Poster Completion |