Project Proposal

Daniel Fagerlie
Sam Shinn
Nicolas Salas

1. What is the problem or task you propose to solve?

We aim to classify sides of debate in discussion posts which occur on social media or forums pertaining to arguments.

2. What is interesting about this problem from an NLP perspective?

Being able to classify a person's stance in a debate could have useful applications in summarization of debates. This could be very useful in social media contexts where hundreds or thousands of people are contributing to a discussion, but a participant cannot realistically read every contribution.

3. What technical method or approach will you use?

We're curious to try out the k-means clustering algorithm. The clustering will happen on feature vectors of words within a post and/or other features we can build from the data (see section 6 below).

4. On what data will you run your system?

One possible corpus might come from, http://mpqa.cs.pitt.edu/corpora/product_debates/ which was gleaned from convinceme.net. If, however, the data proves too sparse (see section 6 below), and if we have time, we might gather more data from convinceme.net because the posts might already be labeled (see section 5 below).

5. How will you evaluate the performance of your system?

Each posts' stance is identified as either A or B in a topic of A vs B. We believe this identification comes directly from convinceme.net when the users make a posting, they might be required to identify their stance. We will use that label only when testing the results.

To evaluate the system we will look at how accurately it clusters the data. This will be done by looking at how the algorithm clustered the documents compared to how the documents are actually grouped.

6. What NLP-related difficulties and challenges do you anticipate?

It's not entirely certain whether or not the k-means is appropriate for this type of classifier system. The feature vectors might not be deep enough. We may end up trying to build some abstraction on top of the sentences within each post. One possibility is to use an existing emotion classifier to identify a sentence as angry, happy, joyful, etc, and use that output for an element in the feature vector for the k-means. Other possible abstractions include sentiment and lexical analysis to determine almost explicit favorability of a topic. These methods are complex, however, and implementing and running them may not be possible within the scope of this project. Experimenting with these more complex feature vectors will depend on our progress within time considerations.

Additional difficulties might arise as a result of this added abstraction; How will we make comparisons of the feature vectors when they inherently represent conflicting types? We will pursue possible solutions. If none are immediately apparent, then we will stick with the basics, learning about the k-means along the way. Failure is oftentimes a better teacher than success.

Unfortunately, the already compiled data doesn't amount to very much. This data sparsity doesn't leave much to prove or deny in our tuning of the clustering technique. As stated above, there is the possibility of gathering more data, but this is highly dependent on what progress we make within the time constraints.

References:

- Swapna Somasundaran, Janyce Wiebe. 2009. Recognizing Stances in Online Debates.
- Sarvesh Ranade, Rajeev Sangal, Radhika Mamidi. 2013. Stance Classification in Online Debates by Recognizing Users' Intentions.