NLP Project - Data Collection

Now that we have decided on how we are going to implement our system, we need to gather the data and format it so that it will be useful to our system. Formatting comes in many flavors, and not everyone's data is going to look the same once it has been gathered and embeddings have been added.

By the end of this week, you should have amassed the minimum amount of data you will need to implement your system. You can continue to gather more data as you work on the project, but you will need enough to develop a proof-of-concept prototype of your system by next week.

Each group will have a single submission:

A 1-2 page paper describing the data you will be using for your project. It will include:

- How the data was/is being captured
  - Where is it coming from?
  - Who are the authors of the data?
  - How did you collect the data?
  - What is the raw format of the data?
- How will the data be formatted to work with your system?
  - How will it be read in?
  - Any normalization or pruning?
  - What annotations will you be adding to the training data? (sentiment, POS tags, meaning, etc)
  - What will a trained model look like?
    - What information will it contain?
    - What features of the data will be used for decision-making?

In addition to the paper, you will also submit a sample of a formatted, annotated document from your data, along with the original raw document before it was processed and an explanation of what was done to this document during processing.