1. This assignment is sort of the course in reverse: let’s use NLP tools others have built, so we can get a grasp on: a) what the linguistic distinctions are that we’re trying to create this semester, and b) how they can be useful. I strongly encourage you to work in groups of 2–4 for this assignment (and you need only turn in one report for the group). I also strongly encourage you to have fun.

   (a) Grab two novel-length English works of fiction (e.g., from http://www.gutenberg.org) and set one aside for development and one for testing. Report which texts you’re using and to what extent they are similar to or different from each other. Make sure the text contains dialogue in some fashion.

   (b) Download and install the Stanford CoreNLP suite of tools (http://nlp.stanford.edu/software/corenlp.shtml). You may use any of the CoreNLP tools except for the relation (RelationExtractorAnnotator) component. Make sure you can apply the tools to your development text and report any difficulties.

   (c) Comb over the development text, with the Stanford annotations added, in order to identify (manually, semi-automatically, and eventually fully automatically) when one character is speaking to another. Specifically, your goal is to extract relations of the form speak(X,Y), where X and Y are characters. (You are allowed to leave X and/or Y unspecified in some cases, noting that these are cases you would eventually like to make fully specific.)

      • Thus, you will be looking for patterns that help make this determination, probably across a range of verbs, styles of quoting, etc. (I expect that you’ll minimally find the lemma, ner, dcoref and depparse components to be useful.) Report both on your process in creating patterns and your specific list of patterns.

   (d) Run your pipeline over the test data. Report in a table the total frequencies with which, according to your system, the different characters are speaking to each other.

   (e) As a final part of the report, add a note on what you would do if you had more time to make the system better.