Assignment 4

L615

Due Wednesday, March 27

1. Fun practical experience:

   (a) Download the Enron corpus (http://www.cs.cmu.edu/~enron/). This may take some time.

   (b) Unpack the corpus (which also may take time). In unix, this can be done in a terminal by calling: tar xvfz enron_mail_20110402.tgz

   (c) Select a directory of one person’s emails. (You can do more directories, or go through the entire corpus, depending upon how much time, experience, and interest you have.)

   (d) Use the Ngrams Statistics Package (NSP) to find collocations in this directory. Do you find any interesting patterns? Does one particular collocation method work best? Do you find anything illegal? You likely will want to remove some meta-data in this process. e.g., you could use something like cat * | egrep -v '^X-' | ... to get the content of all files in a directory, excluding those starting with X-.

   (e) Just for fun, POS tag the data using TreeTagger, and then see what the significant collocations of POS tags are. What are you doing when you do this? (i.e., what does it mean?) Do you find anything interesting?

2. Final project stuff:

   (a) Acquire the corpus that you need for your project, and/or describe to me at what stage you’re at in the corpus acquisition process.

   (b) Provide a one-page (or so) outline of what your final paper will consist of. That is, what is the overall thesis? What will be your main arguments? Where does your corpus data fit in? etc.