CS 124 Problem Set 4

Ted An

May 9, 2011

1. PREPROCESSING CODE:

   - Initial Positive and Negative Seeds contains the, well, initial positive and negative seeds mentioned in the section "Meaning Functions Through MapReduce" in our writeup.
   - The following describes the process by which you can run my preprocessing code given that you download WordNet from http://wordnet.princeton.edu/ and change the calls to wn.exe appropriately in the code. (In both python files it is the only os.system call) Wordnet is over 60 megabytes so for the sake of having a filesize of under 100 megabytes I will not include it.

     In "Code Related to WordNet Output Processing" we have the code we used to generate a list of words alongside their synonyms. The starting list of words is contained in "words.txt." The results are in Intermediate Files, with "Intermediate Files/finalOutput.txt" being our final processed result. **This is the input to the hadoop code in the no-heuristic approach in Meaning Functions Through MapReduce**.

     You can generate finalOutput.txt by running ./graphGen.py.
   - In "Code Related to Generating Informative Priors" we generate the "informative priors" mentioned in the Meaning Functions Through MapReduce section of the writeup. **Note that graph.txt in this file is the input to the hadoop code in the heuristic 1 approach.** The modified file with the informative priors is contained in "Intermediate Files/finalOutput.txt". **This is the input to the hadoop code in the heuristic 1 + heuristic 2 approach.** We can generate this finalOutput.txt by running ./graph2Gen.py.

     negwords.txt and poswords.txt contain the list of positive and negative words found online as referenced in the writeup.

2. HADOOP CODE:

   There are three directories, each corresponding to one of the runs we took. (See Meaning Functions Through MapReduce for more information) For each directory, simply do ./run_pagerank.sh graph.txt. (Note: Each graph.txt is different and the heuristic 1 + 2 code is different)

3. Results is pretty self-explanatory. We only graphed/spread-sheeted our best results, from the heuristics 1 + 2 approach. In Code Output you can find results from our runs of all the code. (Results contained in part-00000)