LING 165 Lab 6: Term Frequency – Inverse Document Frequency

Synopsis

Write a program that calculates term frequency – inverse document frequency (tf-idf) of words in a document.

Task

(1) Download `/home/ling165/lab6/wsj.zip` from the gray server and unzip it.

(2) You should now have a directory named `wsj` that contains 840 files. Assume that (i) each file in the directory is a *document* and (ii) the set of 840 files is a *collection* of documents.

(3) Identify top ten words with high tf-idf in `WSJ_2325`.

(4) Email me the top ten words and where I can find your code.

Data

Each file in the `wsj` directory has just one line consisting of words that are separated from each other by white space. For example, `WSJ_2325` looks like this:

```
oil industry middling profits persist rest year major oil companies next few days ...
```

I've already pre-processed the text for you. Do not process it any further.

FYI, the original file from the Wall Street Journal corpus looked like this:

```
.START

The oil industry's middling profits could persist through the rest of the year.

Major oil companies in the next few days are expected to report much less robust
 earnings than they did for the third quarter a year ago, largely reflecting det
eriorating chemical prices and gasoline profitability.
...
```

Basically, I extracted nouns, verbs, adjectives, and adverbs from the original, lower-cased them, and dumped them into a single line. If you're curious, see the originals under `/data/TREEBANK/RAW/WSJ/` on the gray server.