Write Up

Brandon Parrish, Conner Taylor, Zheng Li

May 2, 2016

Our data is the (approximately) 1,000,000 most frequent 5-grams from the Corpus of Contemporary American English (COCA). The source for this from `http://www.ngrams.info`. To download the data we had to give our email and name and agree not "distribute this list to others, and to not develop any other frequency lists that are based on this data, which will be sold commercially."

The data we use contains eleven attributes all together, but we only use six attriutes for our word correction algorithm. Our data consist of a frequency attribute and a 5-gram that is split into 5 separate string attributes per row. The frequency represents the relative number of times the 5-gram is used in comparison to the rest of the 5-grams. The 5 unused attributes are the corresponding parts of speech for each word in the 5-gram.

Choosing the data set for our spell checker was a challenge; many of the English language corpuses we looked at were either missing frequencies or ended up being too large to load into the database. We decided to use the COCA 5-gram data set because it was small enough to load in the database while containing all the data we needed. We didn't end up using the part of speech data in our program, but it could be useful if we decided to improve the program in the future. For example, if we could identify the parts of speech in our input, we could improve the spell checker's accuracy by only choosing corrected words which have the same part of speech as the original word.

Once we had selected our data set, uploading it into the database was relatively straightforward. We used a scripting language to convert the tab-separated data into .csv format and a SQL script to create the ngrams table with the appropriate attributes. Once the table was created, we were able to use the COPY command to load the data directly into the table.

1 Additional Links

http://norvig.com/ngrams/ (list of dataset files)

http://norvig.com/spell-correct.html (how to implement a basic spellchecker)

https://datahub.io/dataset/beautiful-data-natural-language-corpus-and-code (download for files)

ota.ox.ac.uk/headers/0643.xml (spelling error corpus)

aspell.net/test/ (possibly better spelling error corpus)

www.ngrams.info/download_coca.asp (corpus of contemporary english, 2-, 3-, 4-, and 5-grams)

www.postgresql.org/docs/9.5/interactive/textsearch.html (full text search if we want to use stemming or something)