Auto Complete Design Document

This algorithm takes in a set of words and produces a set of the most commonly used words after the input. It does this using Google Ngram data. This data set tracks how many times a word or phrase was used in the last 500 years. Google started at one grams and have realised all the way up to five grams. Utilizing this data, the algorithm will produce speedy results for in time use.

The idea behind this project is to create an algorithm that can be utilized in different applications. This algorithm can be applied numerous different fields and solve different problems, so I use this algorithm to produce products. I would be selling products that I create with the algorithm as the core. People could come to me with a product in mind that would require my algorithm, and I would create it for them. Depending on the application, this algorithm can make learning in school easier and learning languages better. This is only one use case.

Major Components:

Google Ngram data

- 803.5 GB of compressed data
- Formatted: gram, year, word frequency, book frequency

Data Processing

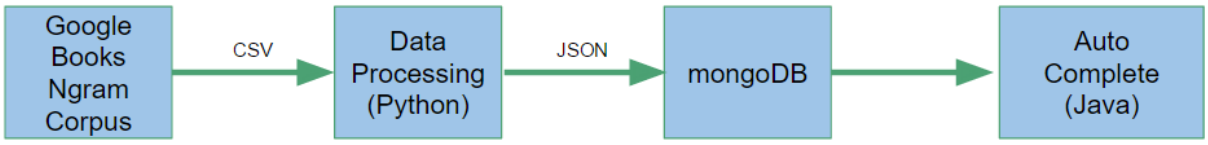
- Process of cutting down from the original data set
- Uses Python's Json library to convert files to json for easy importing into MongoDB

MongoDB (Data storage)

- limits each database to no more than 16000 data files.
- database has a maximum size of 32TB.

Auto Complete

- Uses Mongo Java Drivers
    - Allows to open connection and query data
    - Uses Mongo database object when data comes out of MongoDB
- Requires an in time speed so results do not seem slow to users



Control Flow Graph