[INSERT TITLE, PROJECT INFO, NAMES]

The data that we will use is Wikipedia dumps. We have already downloaded the latest complete english Wikipedia dump that totals (when extracted) to 37.7GB. Figure/Table 1 shows the word frequency for the top twenty words and Zipf's curve of the entire Wikipedia file.

Because Wikipedia contains so many articles, our initial goal of performing hierarchical clustering on the entire Wikipedia might be unrealistic. Also because of the size we would have to store it on disk making the entire process of analyzing it extremly slow. Because of this we used Wikipedia’s Special::Export tool to download a 13MB XML file that contains around 1500 articles. The articles downloaded are based on science, engineering, computer, mathematics and other science related categories. We will use this small file for testing purposes. Figure/Table 2 shows the word frequency for the top twenty words and Zipf's curve of this small file. By using only a small subset of Wikipedia we should be able to keep everything in RAM, speeding up the entire program, and only requiring the tf-idf vectors to be saved on disk.

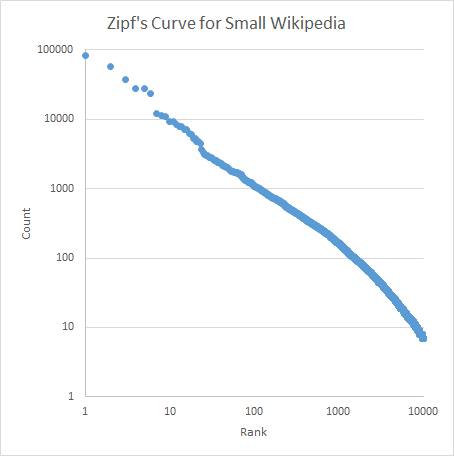
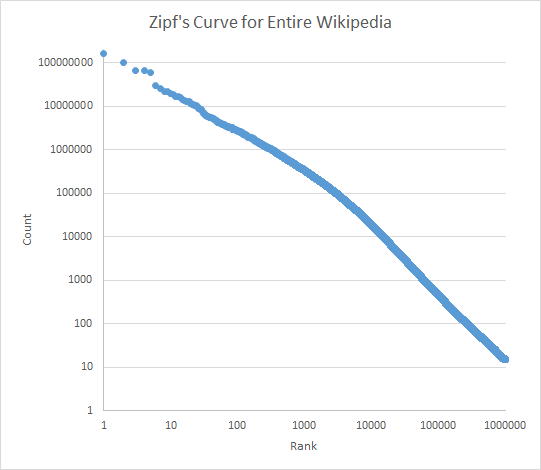
We are making a modification to our Project by only using a smaller subset of Wikipedia. Instead of allowing users to browse through the entire Wikipedia we are going to focus more Wikipedia articles that are relevant to college students.

Table 1 Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| the | 160910158 |  | the | 82678 |
| of | 102109943 |  | of | 57403 |
| and | 67048607 |  | and | 37979 |
| in | 66568469 |  | to | 28094 |
| to | 59590186 |  | in | 27742 |
| is | 29245795 |  | is | 23908 |
| for | 25508367 |  | for | 12271 |
| on | 21745345 |  | that | 11434 |
| was | 21133251 |  | as | 10965 |
| as | 19790864 |  | by | 9397 |
| by | 18485800 |  | are | 9164 |
| that | 16723028 |  | be | 8373 |
| with | 16702919 |  | with | 7827 |
| it | 16119454 |  | or | 7752 |
| from | 14528676 |  | on | 7088 |
| at | 13255091 |  | an | 7033 |
| this | 12873397 |  | this | 6281 |
| be | 11240966 |  | it | 6021 |
| he | 10847206 |  | from | 5225 |
| an | 10842524 |  | can | 5223 |