Project Map

[Dan] Create Wikipedia Scraper (completed 28 November) (Data Acquisition)

[Dan] Finish downloading at least 1 million lines of Wikipedia. (Data Acquisition)

[Supratik] Separate the text into sentences using either Python or Notepad++ (Data Value)

[Supratik] Create a frequency list of all the worlds from our corpus (sentence list) (Data Value)

[Supratik and Dan] Assign the frequency of the least frequent word in each sentence to each sentence. (Data Value)

[] Order sentences based on frequency (descending order). This is a major success, with this we have created a 'path of least resistance' for language study. (Data Value)

[] Create a 'bigram', 'trigram', 'n-gram' frequency lists (Data Value)

Example of trigrams:

The sentence "the quick red fox jumps over the lazy brown dog" has the following word level trigrams:

the quick red

quick red fox

red fox jumps

fox jumps over

jumps over the

over the lazy

the lazy brown

lazy brown dog

[] Create any other data sets based on the corpus. (Data Value)

[] Store all our data in our database (Data Persistence)

[] Create a data API (Flask) or website (FLASK) to allow the data to be shared (Data Exposure)

[All] Create 5 – 9 minute presentation in a PowerPoint or demonstration. (Documentation/Presentation)

[Dan] Do a writeup of the project (Documentation/Presentation)