Kai Levy
Software Design Spring 2015
Project 3: Text Mining

Wikipedia: Comparing English and 'Simple' English

Wikipedia has a separate 'Simple English' language that I chose to investigate. In my project, I harvested text from both languages and analyzed complexity in terms of word length, paragraph length, overall length, as well as the similarity between two given articles of the same subject. I hoped to gain insight into how much simpler Simple English is, and how much similarity there is between the two features.

To achieve my goal of analyzing the articles, I start by downloading them from wikipedia and filtering out unwanted sections and symbols. For a given article, of which I have a list, I retrieve both the English and Simple English versions and store them as strings. I then analyze them, using a number of simple functions that compute the word count, average word length, and average paragraph length. I also compute the cosine similarity between the two languages, of each article. These statistics I pass back as a list. With that data, I can use matplotlib to graph the notable trends in my data.

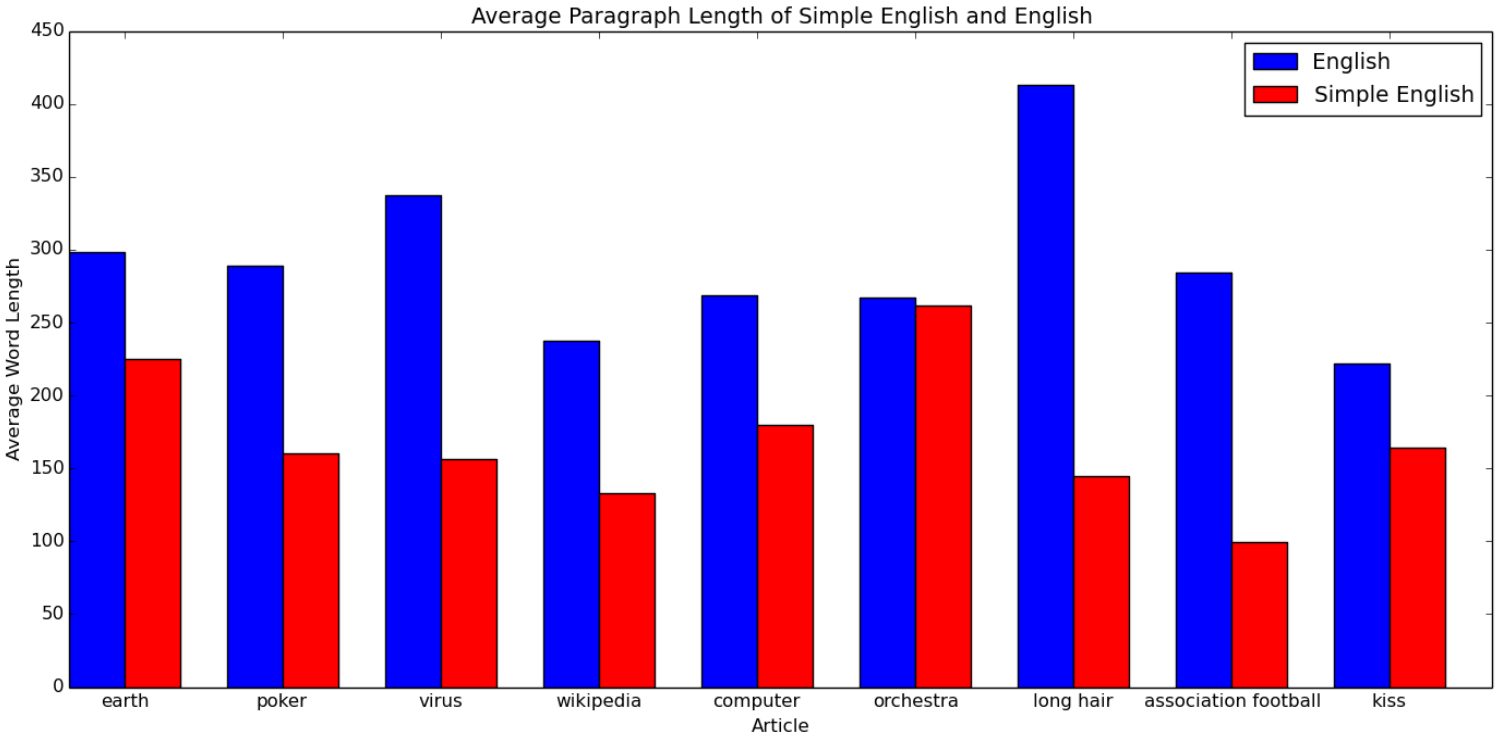
One design decision to note is my choice to ignore certain parts of each article. I consider the often ignored but very lengthy sections at the end of Wikipedia articles (such as Notes, References, etc.) to be less important and even statistically misleading, considering those sections are generally left off in Simple English. Additionally, I decided to ignore punctuation and digits, because they would skew the character count and be considered words by my functions, which would lead to less perceived similarity.

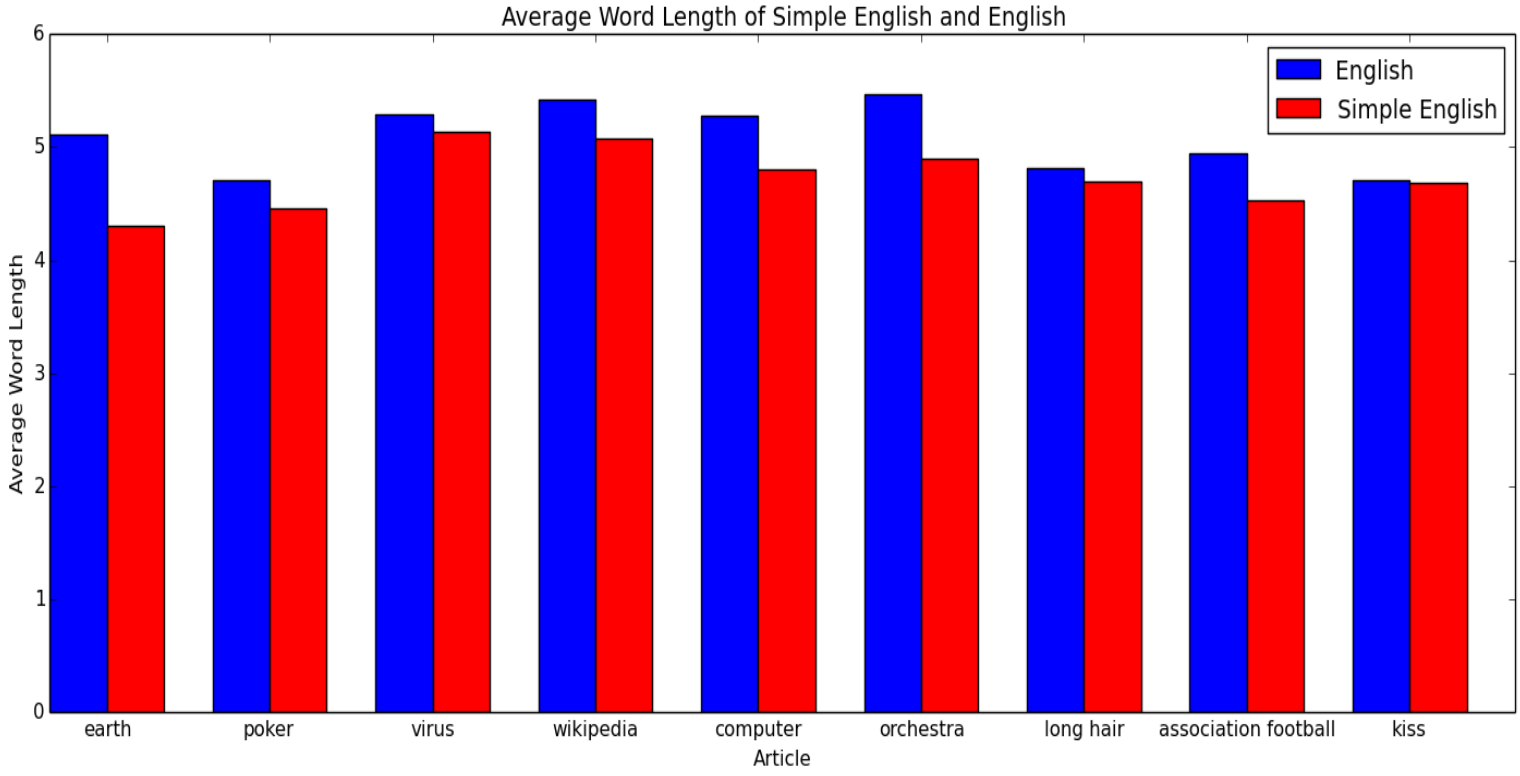
The bottom line that I discovered was: Simple English IS simpler than English. As apparent in the Graph of Average Paragraph Lengths, English had much longer paragraphs on average, with only 'orchestra' coming close. The difference in word length, while less extreme (because word lengths very much less than paragraph lengths) is still apparent and consistent.

Additionally, most of the articles were very similar by my analysis. One notable outlier was the article about kissing. Upon further investigation, that discrepancy is the result of a very, very simple Simple English article matched up against a very extensive English article. The English article covers history, culture, roles in religion, legality, and as well as it's role in evolution, while the Simple English article simply describes what it is, and how it can spread disease. My data is below.

Average similarity: 0.884

Title	Similarity
Poker	0.954
Association Football	0.894
Wikipedia	0.864
Orchestra	0.918
Virus	0.938
Computer	0.927
Kiss	0.679
Earth	0.941
Long Hair	0.844


Average Paragraph Length of Simple English and English
Average Word Length
Article
English
Simple English
0
50
100
150
200
250
300
350
400
450
earth
poker
virus
wikipedia
computer
orchestra
long hair
association football
kiss


Average Word Length of Simple English and English
Average Word Length
Article
English
Simple English
0
1
2
3
4
5
6
earth
poker
virus
wikipedia
computer
orchestra
long hair
association football
kiss

Looking back, I thought this project went pretty and I was able to discern a clear result with my program. However, I chose a topic which had a very clear answer already, as Simple English's purpose is to make Wikipedia more understandable. Additionally, my results were based on several selected articles that had both English and Simple English articles. To be more thorough, I should have designed a method for randomly selecting eligible articles. My unit testing worked fairly well throughout the project, but I wish I had a firmer grasp on what was happening in the Cosine Similarity, because I mainly relied on NINJA help and articles I found online to do that part.