**Bsc Progress Report: Properties of Text**

**Supervisor: Dr Tim Evans**

**Students:** Tahseen Hamid Omaar and Immanuel Adewumi

***Date: 13th November 2020***

**Current Status**

In our investigation of the Zipf Law phenomenon in natural language, according to the initial project schedule we have achieved:

* Collating large text corpora from 3 different languages: - English, Yoruba and Igbo for further analysis;
* Creating prototype code blocks for Zipfian analysis which graph the word frequencies of various words (tokens) in bodies of text against their rank.

***Corpora-types collated:***

BBC News articles:- Hausa, Yoruba, English and Igbo

The Bible:- Yoruba, Igbo and English (World English Bible - WEB).

NLTK - Spoken and Written Corpora; Chat (text) Corpora, US Inaugural address Corpora, and thousands of Novels from the Gutenberg Project.

**Milestones**

* Our Initial Plan
  + *Web Scraping Corpora:* In project week 4 we set out to gather text-corpora for 3 languages using web scraping methods. We expected the greatest challenge would be scraping text from bbc news articles in Yoruba and Igbo Nigerian languages.
  + *Scraping Tools discovered:* After doing some research on tools to use we zeroed in on webscraper.io a Google Chrome web extension. It took us more than 24 man hours to master.
  + *Difficulties:* We tried multiple approaches whilst attempting to use this to web-scrape both youversion.com, a multi-version bible website, and the BBC news website. Its efficiency in scraping youversion was less than satisfactory, as only a single configuration of the Web Scraper was functional, and it was temperamental. For the BBC website the web scraper was unable to parse through and thus could not achieve the set task.
* Resolutions and discoveries
  + In searching for better web-scraping alternatives we discovered the professional multipurpose tool Apify-Web Scraper. It recovered text from more than 1000 BBC articles in less than 30 minutes. This drastically reduced our initial estimated data acquisition time from about 36 hours to less than 5 hours. Thus, we were able to achieve our set objective of recovering Corpora for BBC *articles* segmented, according to the BBC site, by ***Sports*, *News*** and ***Most Popular*** for each of the 3 languages; *English, Yoruba and Igbo****.***
  + Despite the success in scraping bbc.com, we were however, with our current level of expertise in using this tool, unable to parse through an entire language-translation of the youversion bible site. Instead, we settled for scraping the site using the initial webscraper.io tool and using Python to clean up the extracted data.

**Analysis Tools Discovered**

* NLTK - We discovered NLTK an online NLP (natural language processing) book and python package for text and language evolution analysis, used by data science experts. We were able to read up and understand the pertinent code blocks. The NLTK functions most pertinent to this project include:
  + Searching for word-frequencies in large bodies of text using set language constraints (for example having a maximum of 2 identical vowels adjacent to one another).
  + Observing the rate of occurrence (frequency) of specific words in language and written text over long time-periods such as years and decades.
  + Providing tools and guidance for processing (cleaning up) raw text files.
  + Access to some of the world's largest banks for Corpora - Namely the Gutenberg Project.

**Current Objectives**

* Having achieved our initial objectives our immediate future objective is to use python and the NLTK package to do a final clean up of our data and begin exploring the presence of Zipf's law and Heap's law in various text types and languages.
* For Zipf law analysis we will do preliminary tests by passing different text bodies through the prototype Zipf analyser we developed. From results of preliminary tests we will fine-tune the code to analyse larger bodies of text.
* More in-depth and qualitative analysis of the general features of various text bodies (as well as the frequency of use of specific word-types across languages) will be done using the NLTK package.
* Artificial Language - Finally we will explore, using python, the possibility of creating an artificial language with a specific subset of Unicode text encoder and we will check for the presence of Zipf's Law phenomenon in this created language also.