Frequency Analysis of Ngrams

Interim Report

Team Number: ABQAcad-1

School Name: Albuquerque Academy

Area of Science: Behavioral and Social Sciences

Project Title: Frequency Analysis of Ngrams

Problem Definition:

Linguistics is a field of research that is rarely thought of as a supercomputing problem. However, it most certainly is. Computers are a powerful tool for linguistics as it can provide a means of analyzing massive amounts of data. My goal is to perform statistical analysis on word frequencies in the English language and to compare the frequencies of words in English Literature as opposed to online social media.

This project is important because of its relevance to many areas of computer science, such as speech recognition, spell-check, translators, cryptography, language parsers, and even human-like AI. The research I am doing relates to work being done by other computer scientists, and I will be able to extend some work done by others.

Problem Solution:

I am dividing this project into two main parts. I will be using n-gram (a 1-gram is a phrase unseparated by spaces, one word, a 2-gram is a phrase separated by one space, two words, etc.) data provided by Google as my initial starting point. This data will provide a solid basis for my project because Google used a dataset of over 5 million books. With more than 1.4 million 1-grams in the dataset I plan on using. I will perform analysis on this data to better understand word frequencies in Literature.

The second part is to use randomly sampled tweets gathered from the Twitter API to assemble a new n-gram data set. The tweets will have to be collected over a period of a couple of days as tweets from the API trickle in at a measly 39 tweets/sec, and of those only 12 usable tweets due to length or language factors. I have a goal sample size of 10 million tweets to use to construct the n-gram database.

Progress to Date:

I have performed my analysis on the Google Ngrams, and have worked out a system of working with the data so that if I need to change anything at a later date, it doesn’t take 3 hours to re-run. I have also written the code for gathering tweets, as well as done some preliminary analysis and visualization of the data.

I have yet to do a multi-day long twitter sampling to achieve 10 million tweets, and I also need to tackle the problem of counting and splitting n-grams. The final step will be the comparisons between literature and social conversation.

Expected Results:

As expected with literature n-grams, a couple of words are extremely common; however, everything else is pretty even concerning frequency. This frequency distribution makes sense because in literature there is a significant effort to increase vocabulary. When comparing to social media, however, I expect to see a bunch of words which are extremely common, with the majority less frequent.

Sources:

Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010)

Damashek, Marc. Gauging similarity with n-grams: Language-independent categorization of text (http://search.proquest.com/openview/27eb4f15f42e26562a57cf8224a9ccd1/1?pq-origsite=gscholar)

Igor Santos, Yoseba K. Penya, Jaime Devesa, and Pablo G. Bringas. N-GRAMS-BASED FILE SIGNATURES FOR MALWARE DETECTION

Irene Langkilde, and Kevin Knight. THE PRACTICAL VALUE OF N-GRAMS IN GENERATION

Mitja Trampus. Evaluating language identification performance (https://blog.twitter.com/2015/evaluating-language-identification-performance)