Team Number:

School Name: Albuquerque Academy

Area of Science: Behavioral and Social Sciences (linguistics)

Project Title: Linguistical Frequency Analysis

Proposal:

The Google Ngrams project has analyzed the occurrences of various words and phrases in over 5 million digitized books. Google has made this data open source for any use. The 1-grams consists of over 1.4 million rows. The sheer quantity of data makes this a computational challenge, not to mention that most 1-grams are non-words. This data can also be used for analysis on other 1-grams.

On top of this, I will be taping into the Twitter API, and possibly Facebook too, to get small sample of what is said on these social media. The Twitter API has support for random sampling of tweets, streaming 1% of tweets to me in real time, and I will perform my own 1-gram counts on this data. Through this, I will be able to get insights into the difference between more formal literature and day-to-day conversations. I can also analyze day-to-day trends between people.

This problem encompasses linguistics, history, and computational science. While this is a challenging problem, it is certainly feasible and provides plenty of room to draw conclusions. The limitation of this data is the source, as google has only processed this data for five million books, around 4% of all books ever written. This data is also limited to literature as opposed to every day conversation. This research relates to research being done by other linguists, and having applications in speech recognition, spell-check, translators, cryptography, language parsers, and even human-like AI.

Team Member(s):

- Jeremy Rifkin

Sponsoring teacher:

- Jim Mims

Project Mentors:

- Jim Mims
- Kevin Fowler