Characterizing the Google Books Corpus

Eitan Pechenick
Dr. Peter Dodds

Department of Mathematics and Statistics &
The Computational Story Lab
University of Vermont

April 16, 2014

Abstract:

A wealth of data is available for the study of the volume and frequencies of words and phrases over time. In particular, the Google Books corpus represents a compilation of billions of word and phrase occurrences in written works spanning five centuries and several languages. It is tempting to treat the entries for a given year in a Google Books dataset as an indicator of the popularity of various words and phrases in that language during that year. Doing so allows one to examine trends in various classes of phrases and draw novel conclusions about the evolution of public perception regarding a given topic. However, sampling published works and doing so by availability and ease of digitization leads to several significant effects. One is the ability of a sufficiently prolific author to effectively insert new phrases into the observed language at noticeable levels. A greater effect arises from professional (especially technical) texts which have become numerous in the last several decades and are heavily sampled in the corpus. This results in a surge of phrases that are typical in these texts, but not common in the language at large, as well as an increase in references to time through citations. This work aims to highlight these artifacts and to filter trends due to the nature of the datasets. One available tool is the English Fiction dataset from the same corpus, since it is not as heavily affected by professional texts.