Wildcards and Morphological Inflections for the Google Books Ngram Corpus

Jason Mann, David Zhang, Lucille Yang, Slav Petrov and Dipanjan Das
Google Inc.
jcm2207@columbia.edu, dzhang21@gmail.com, ly77@cornell.edu
{slav,dipanjand}@google.com

Abstract

We present a new edition of the Google Books Ngram Viewer, which plots the frequency with which words and phrases were used over the last five centuries; it's data encompasses 23% of the world's published books. The new edition adds three tools for more powerful search: wildcards, morphological inflections, and case insensitivity. These additions allow for the discovery of previously unknown patterns in the Ngram Viewer data, and further facilitate the study of linguistic trends in printed text.

1 Introduction

The Google Ngram Viewer http://books.google.com/ngrams and its corresponding Google Books Corpus (Michel et al., 2011), since their release in 2010 have assisted in the analysis of cultural and linguistics trends through five centuries of data in eight languages. We present an updated version of the Viewer which introduces several new features.

First, users can replace one term or tag in their queries with a placeholder symbol (wildcard, henceforth), which will return the ten most frequent replacements in the Google Books corpus for the specific year range entered. Second, by adding a specific tag to any word in a query, morphological inflections (or variants) will be returned. Finally, the new interface has an additional option to allow for multiple capitalization styles. In addition, this demonstration presents an overhaul of the Viewer's user interface, with interactive features that allow for easier management of the increase in data points returned.

Mention related and prior work here.

In the following section, we provide an overview of our system's structure. We detail interesting use cases in section 3, which were difficult (or even impossible) to search in the previous versions of the NGram viewer that did not handle wildcards in the search queries. Additionally, we detail how the two other aforementioned features introduced in this demonstration paper result in interesting retrieval results. Beyond specific searches, we envision the new functionality of the tool uncovering trends and patterns not readily apparent in the data.

2 System Overview

In this section we present an overview of the Google NGram Viewer backend. Before going into the details, we first describe the corpora on which users of this tool can issue queries.

2.1 Ngram Corpus

The Google Books Ngram Corpus is available at http://storage.googleapis.com/books/ngrams/books/datasetsv2.html. The corpus provides ngram counts for eight different languages over more than 500 years; additionally, the english corpus is split further into British English and Fiction to aid domain restriction. This corpus is a subset of all the books digitized at Google, and represents more than 6% of all publicized texts in its newest edition. The differences between the first and second versions of the corpus are discussed at length in Lin et al. (2012) and the work in this demonstration is limited to the latest 2012 version. **Were there any improvements past this?**

2.2 Wildcards

We support the use of wildcards by utilizing an additional database that stores the most frequent replacements of queries to the ngram corpus. This wildcard database is created as a pre-compilation step when creating the Ng ram Corpus from the

Wildcard	Replacements			
a * man	a young man	a good man	a other man	a wild man
booked=>*_NOUN	booked=> flight_NOUN	booked=> passage_NOUN	booked=> room_NOUN	booked=>seat_NOUN
John said_INF	John says	John said	John say	John saying
'book_INF_NOUN	book	books		
'the cook' (case insensitive)	THE COOK	The Cook	The cook	the Cook

Table 1: A table showing examples of the possible precompiled wildcard, inflection, and case insensitive queries.

Google Books data. When a new ngram is created, one word or tag at a time is replaced with the wildcard symbol, '*', creating a wildcard query. The query becomes a key in a string indexed lookup table, and the original ngram is added to a list of ngrams which are its values. After collecting all the possible replacements for the corpus, the replacements are further pre-processed such that only the ngrams that appear in the top ten for any possible year range are kept in the database. On runtime, this collection of ngrams is processed for the specific year range and the top ten results are returned. For examples of expansions see Table 1. Although a generic wildcard '*' exists in stand alone form, we recommend the use of Part of Speech (POS) tags (i.e. '*_NOUN') for more specific results.

2.3 Morphological Inflections

Inflections of words in search queries are handled using a Google Search interface that can provide morphological variations of words for different syntactic categories (provide footnote about the ?define? keyword in Google Search). Unlike the wildcard substitutions, there is no need for pre-computation, while the results returned, even for languages such as Russian with vast morphological diversity, keep to a manageable number. While manageable, there can be more than 10 results returned per query, unlike the wildcard search; therefore we have updated the user interface to better deal with more data lines (see Section 2.5). Due to the time complexity of resulting queries, we do not allow the combination of morphological inflections with wildcards and/or case insensitive searches.

2.4 Case Insensitive

Case Insensitive searches are enabled by selecting a check box on the new interface. These queries, like the inflections are computed at runtime and include these variations: ALL CAPS, Camel Case, Independent camel Case (all possible variations), and all lower case. To further eliminate scanning errors (mostly in the case of all caps), we utilize a **threshold that is a certain percentage what percentage?** of the top results returned and ignore the results below.

2.5 User Interface

Should we talk about the interface?

3 Use Cases

We present multiple use cases that can be captured using the several features that we have presented in this paper. First, we show some examples of each of these individual features; next, we present some example queries that combine queries that use syntactic annotations (Lin et al, 2012) and the current additions to exhibit the type of results that the NGram Viewer can retrieve.

4 Conclusions

We have presented a new version of the NGram Viewer with some new functionality. With the introduction of these new features, users can perform more powerful searches that show trends which were not possible to extract from earlier versions of the tool.

We can cite examples from the media where this has been mentioned, and also show examples from several blog posts/entries from the internet: http://sciencerefinery.com/2013/10/28/google-ngram-viewer-now-more-powerful-than-ever/ http://www.devingriffiths.com/google-books/google-n-gram-studies/

http://languagelog.ldc.upenn.edu/nll/?p=8472
http://www.textualscholarship.nl/?p=14051

References

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. Association for Computational Linguistics.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*.