Wildcards and Morphological Inflections in Google Books Ngram Viewer

Background

The Google Books Ngram Viewer plots n-grams derived from the Google Books corpus over time. The Books corpus captures more than 6% of the world's published books. This version adds complex searches to automate searches that previously relied on manual discovery.

New Features

Wildcards

Wildcard searches are created by including a '*' in the query in place of a word or POS tag. These searches are limited to the top 10 results that are ranked based occurrence in the time frame given.

- ‘*’s Theorem: Lagrange’s Theorem, Gauss’s Theorem, Euler’s Theorem, Pascal’s Theorem, ...
- ‘War=>*_NOUN’: War=World_NOUN, War=>Civil_NOUN, War=Second_NOUN, War=>Cold_NOUN ...
- ‘cook_*’: cook_NOUN, cooks_NOUN

Morphological Inflections

To search for inflections, attach the '_INF' tag to a specific word. All inflections with data are returned.

- ‘seine_INF’: ist, sind, war, sein, waren, sei, wäre, ...
- ‘book_INF’: book, books, booked, booking
- ‘book_INF_NOUN’: book_NOUN, books_NOUN

Case Insensitivity

For capitalization variations of a word, we have added a checkbox to the right of the query entry text box. We return all capitalization forms that meet a threshold in comparison to the other results.

User

‘the cook’(case insensitive): the cook, the Cook, The Cook, THE COOK

Part of Speech Tags

For further refinement, we allow the part of speech tags introduced in the previous version to be appended to any of the above searches. We also maintained the previous version's implementation of using a part of speech tag in place of a word.

The set of part of speech tags that can be appended (she_PRON) or stand alone (_PRON_) is:

-NOUN_ _VERB_ _ADJ_ _ADV_ _PRON_ _DET_ _ADP_ _NUM_ _CONJ_ _PRT_

These tags can only stand alone (_PRON_):

.START_ _END_ _ROOT_

Example Searches

- ‘*ADP Ukraine, *ADP the Ukraine’
- ‘University=>*_NOUN’
- ‘theater_INF: eng_us_2012’

System Architecture

Dotted lines indicate new architecture

Raw Ngrams

Ngram Viewer

User

Inflections

Case

Wildcards

System

Note: Due to computational complexity, only one of these new searches is allowed per query. 4- and 5-gram searches are also excluded from the new functionality.

Acknowledgments

We would like to thank John DeNero, Jon Orwant, Karl Moritz Hermann for many useful discussions.