GOOGLE N GRAMS

Julia Kuznetsova

GOOGLE BOOKS: GOOGLE COLLECTION OF ELECTRONIC BOOKS

- Books are scanned using a camera at a rate of 1,000 pages per hour
- Automatically corrected for the curvature of pages in a book
- By constructing a 3D model of each page and then "de-warping" it, Google is able to present flat-looking pages without having to really make the pages flat

N-GRAM

- A combination of N words directly following each other in a text (bigrams, trigrams etc.)

- *Two households, both alike in dignity,*
  In fair Verona, where we lay our scene…

- 1 grams: two, households, both, alike, in, dignity…
- 2 grams: two households, households both, both alike, alike in, in dignity, dignity in…
- 3 grams: two households both, households both alike, both alike in, alike in dignity, in dignity in, dignity in fair…
- 4 grams: two households both alike, households both alike in, both alike in dignity, alike in dignity in, in dignity in fair, dignity in fair Verona…
- 5 grams: two households both alike in, households both alike in dignity, both alike in dignity in, alike in dignity in fair, in dignity in fair Verona, dignity in fair Verona where…

GOOGLE NGRAMS

- The corpora used for the search are composed of total_counts, 1-grams, 2-grams, 3-grams, 4-grams, and 5-grams files for each language. The file format of each of the files is tab-separated data. Each line has the following format:
- total_counts fileyear TAB match_count TAB page_count TAB volume_count NEWLINE
- Version 1 ngram file (generated in July 2009): ngram TAB year TAB match_count TAB page_count TAB volume_count NEWLINE
- Version 2 ngram file (generated in July 2012): ngram TAB year TAB match_count TAB volume_count NEWLINE
- http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

GOOGLE NGRAM VIEWER

- The Google Ngram Viewer uses match_count divided by total volume_count for that year to plot the graph
- Tea and coffee
  - Tea and coffee in British English
  - Tea and coffee in German
  - Tea and coffee in American English
  - Tea and coffee in Russian
  - Tea and coffee in Hebrew
- Freedom and equality
  - Freedom and equality in Russian

FINDING TRENDING TOPICS

- Hitchcock 2010:
  - usage ratio (number of word occurrences divided by size of subcorpus) for each decade from 1890
  - explore changes by decade
  - filter out any words with a ratio less than or equal to 0.001%

Hitchcock, Andrew. 2010. Finding trending topics using Google Books n-grams data and Apache Hive on Elastic MapReduce. Electronic resource: https://aws.amazon.com/articles/Elastic-MapReduce/5249664154115844.

TOP 30 RESULTS PER DECADE

- **1900**
  radium, ionization, automobiles, petrol, archivo, automobile, electrons, mukden, anopheles, marconi, botha, ladysmith, lhasa, boxers, suprema, aboord, rotor, turkes, wireless, conveyor, manchurian, erythrocytes, shoare, thirtie, kop, tuskegee, thorium, audiencia, bvo, arteriosclerosis
- **1910**
  cowperwood, britling, boches, montessori, venizelos, bolsheviki, salvarsan, photoplay, pacifists, joffre, petrograd, pacifist, bolshevism, airmen, kerensky, foch, boche, serbia, serbian, hindenburg, madero, serbians, bombing, ameen, anaphylaxis, aviators, syndicalism, aviator, biplane, taxi
- **1920**
  bacteriophage, fascist, mussolini, fascism, sablin, latvia, insulin, peyrol, volstead, czechoslovakia, iraq, vitamin, kenya, curricular, swaraj, reparations, broadcasting, slovakia, vitamins, gandhi, automotive, kemal, zoning, jazz, isotopes, isoelectric, airscrew, shivaji, czechoslovak, stabilization

HOW WORDS THAT PEAK CHARACTERIZE A YEAR

- Russian 1-grams that are found in more than 0.001% of books
- Word peaks in a year: *гитлеровской* 'Hiltler's' 1942