НИС Культуромика

Борис Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

19 октября 2018

Содержание

Содержание

Google Books

Google решил создать искусственный интеллект и научить его на всех книгах, изданных человечеством.

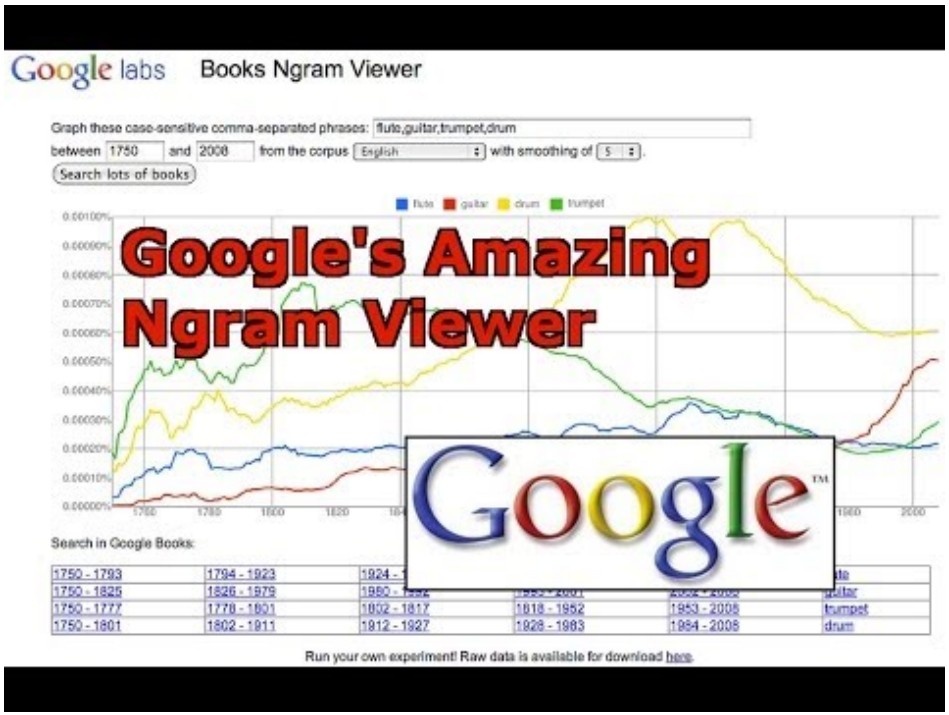
Но книги для этого нужно оцифровать:

Google Ngram Viewer

В 2011 году Google был корпорацией добра и выложил собранные из книг текстовые данные в открытый доступ.

Из-за ограничений, связанных с авторскими правами тексты пришлось нарезать на n-граммы.

Кроме того, к ним был приделан поиск, в результате получился веб-сервис:

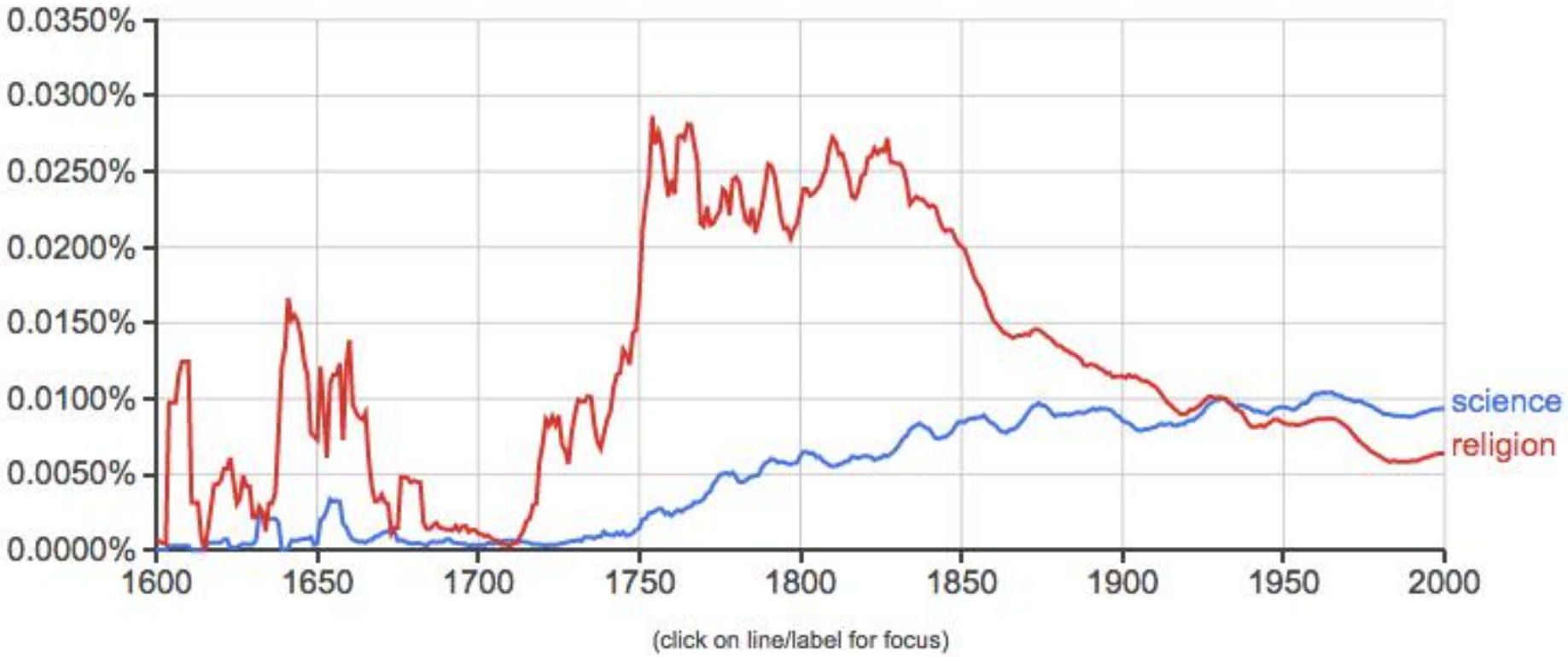


Культуромика

В 2011 году придумали «Культуромику».
J.-B. Michel et al. Quantitative analysis of culture using millions of digitized books (2011)

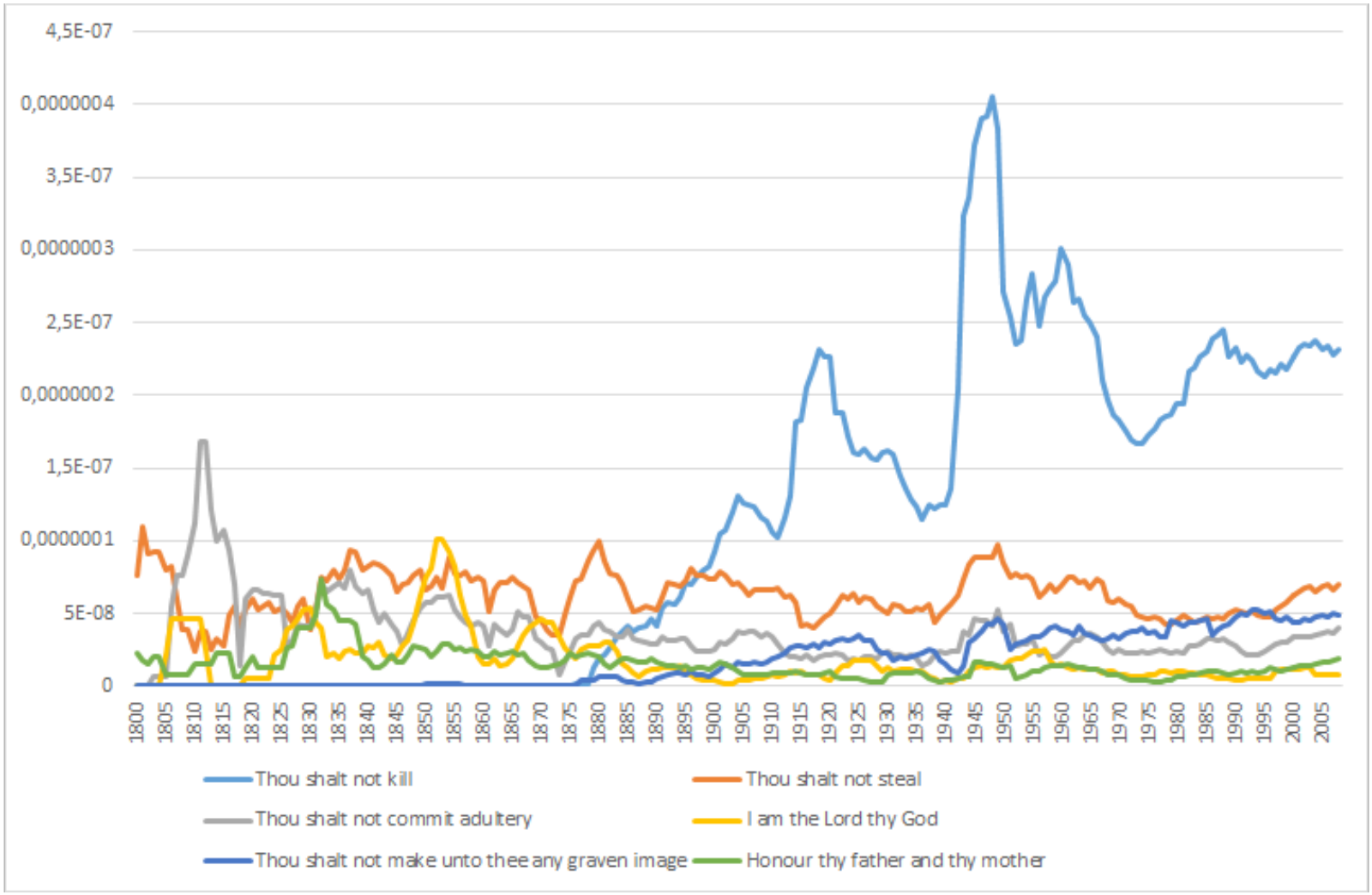
Слова и вещи

Основная гипотеза: частотность употребления слов в книгах отражает важные культурые тренды



n-граммы можно скачать и посчитать самим

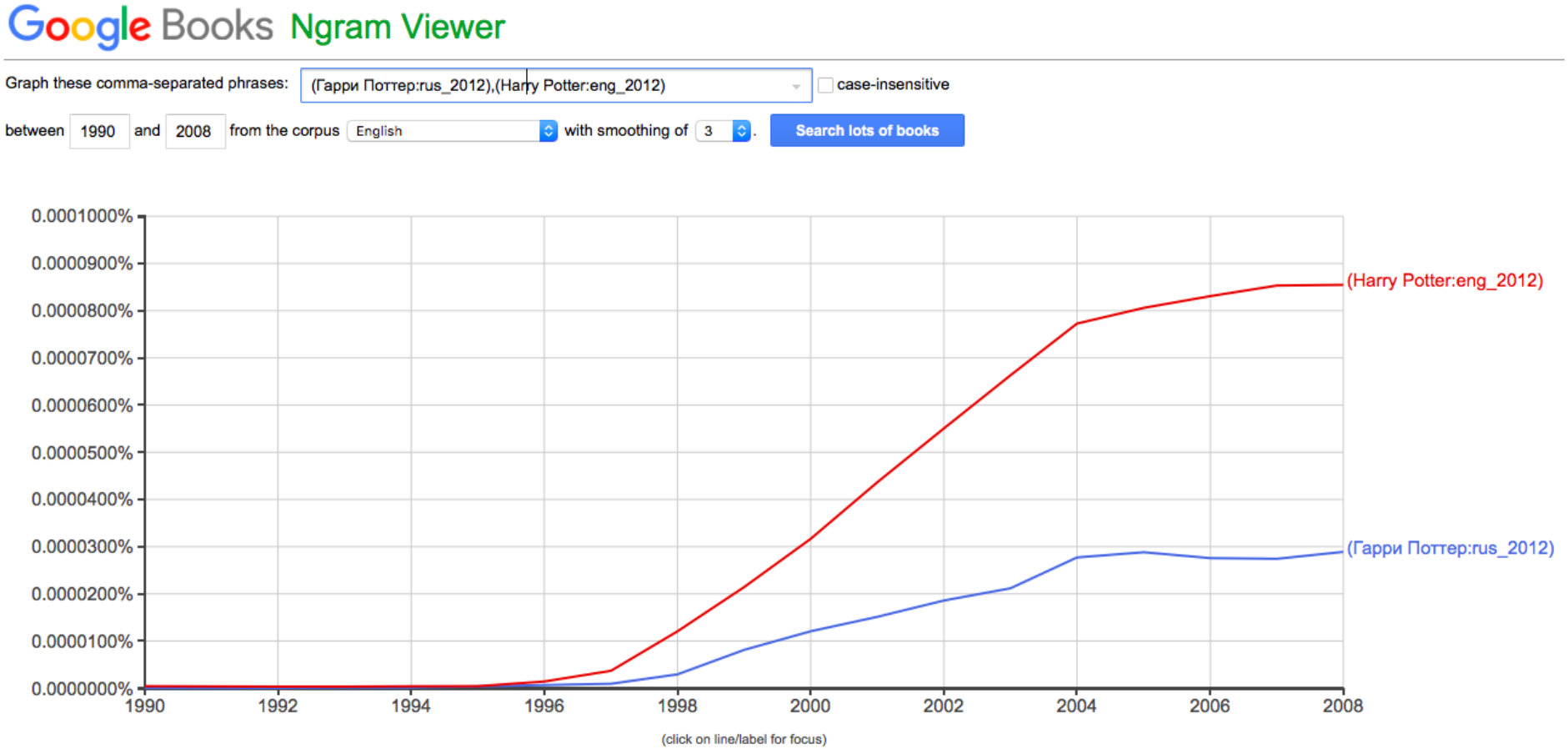
Частотность упоминания библейских заповедей



Содержание

Книги на русском в данных есть

Гарри Поттер у нас и у них:



Проблемы

Не распознана вся старая орфография

Качество распознавания вообще страдает

Содержание

Задачи: spell-checking, data analysis, linked open data

- Исправить орфографию в данных по дореволюционным книгам.
- Вывесить исправленные наборы ngrams в открытый доступ.
- Искать статистические выбросы и тренды автоматически:
  - Где пики и спады?
  - У каких слов и n-граммов отрицательная корреляция?
  - О чём не подумал исследователь, но что найдёт бездушная машина?
- Связывать слова и n-граммы с Wikidata и так автоматом искать связи.

Ngrams + Wikidata

Объяснение пика в связанных данных

WIKIDATA

Main page

Item Discussion

1980 Summer Olympics (Q8450)

Games of the XXII Olympiad, held in Moscow in 1980

Moscow 1980 | Games of the XXII Olympiad