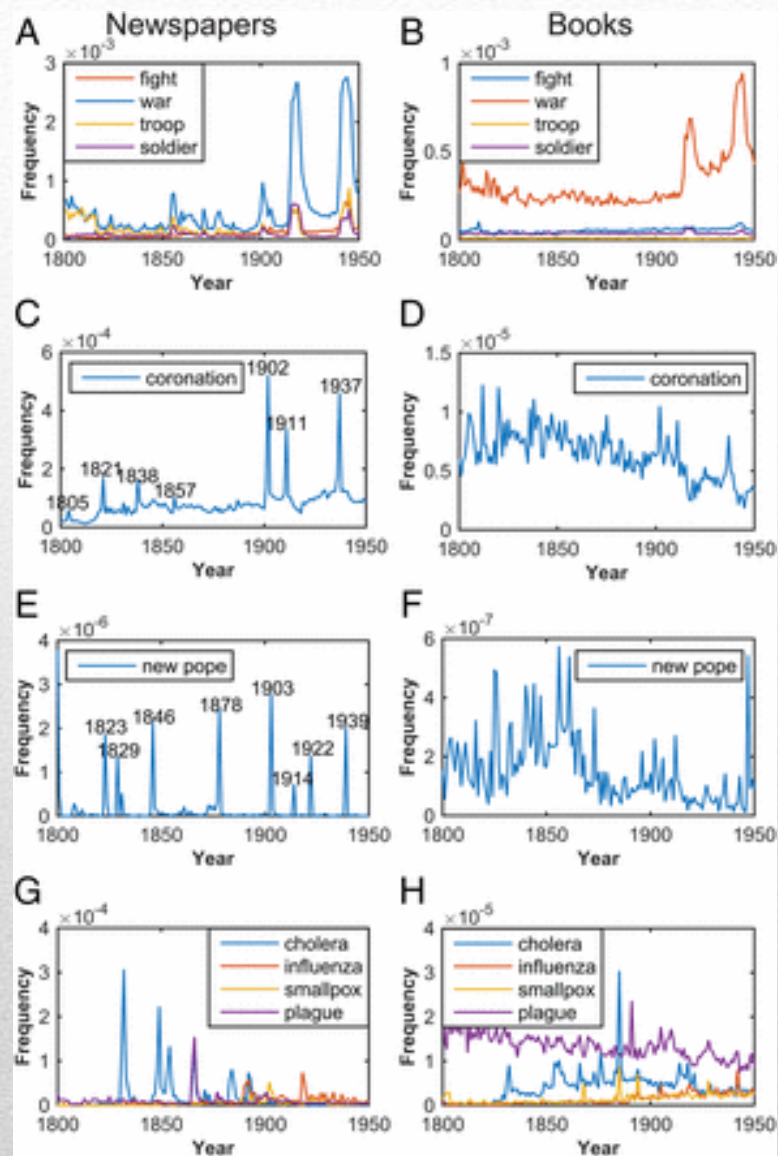


Content analysis of 150 years of British periodicals

Thomas Lansdall-Welfare, Saatviga
Sudhahar, James Thompson, Justin
Lewis, FindMyPast Newspaper Team, and Nello
Cristianini

Плюсы	Минусы
Охватывает 200 лет	Подсчёт слов вместо контент-анализа
5 миллионов книг	Отсутствие географических данных
	Отсутствие точных временных данных (или обращение к далёкому прошлому)
	Игнорирование семантики и контекста
	Ориентация на внутренние переживания

Статистический анализ содержания исторических книг ('Culturomics')



- Большой корпус исторических британских газет (www.britishnewspaperarchive.co.uk)
- Большое количество инструментов искусственного интеллекта

Исследование британской периодики

- Охватывает 150 лет (1800-1950)
- 835 изданий
- 35,9 млн статей
- 28,6 млрд слов
- 120 региональных или местных новостных агентств (14% всех британских региональных изданий)

- Разнообразие тем в издании
- Возраст издания
- Регион издания (Восток, Восточный Мидленд, Северная Ирландия, Лондон, северо-запад, северо-восток, Шотландия, Юго-Восток, Юго-Запад, Уэльс, Уэст-Мидлендс и Йоркшир)

Особенность выборки

- История
- Культура
- Гендерные стереотипы
- География
- Технологии
- Политика
- Экономика
- Общественные ценности и убеждения

**Поиск тенденций в
следующих сферах:**

- **N-грамма** — последовательность из n элементов.
- С семантической точки зрения, это может быть последовательность звуков, слогов, слов или букв. На практике чаще встречается N-грамма как ряд слов, устойчивые словосочетания называют коллокацией.

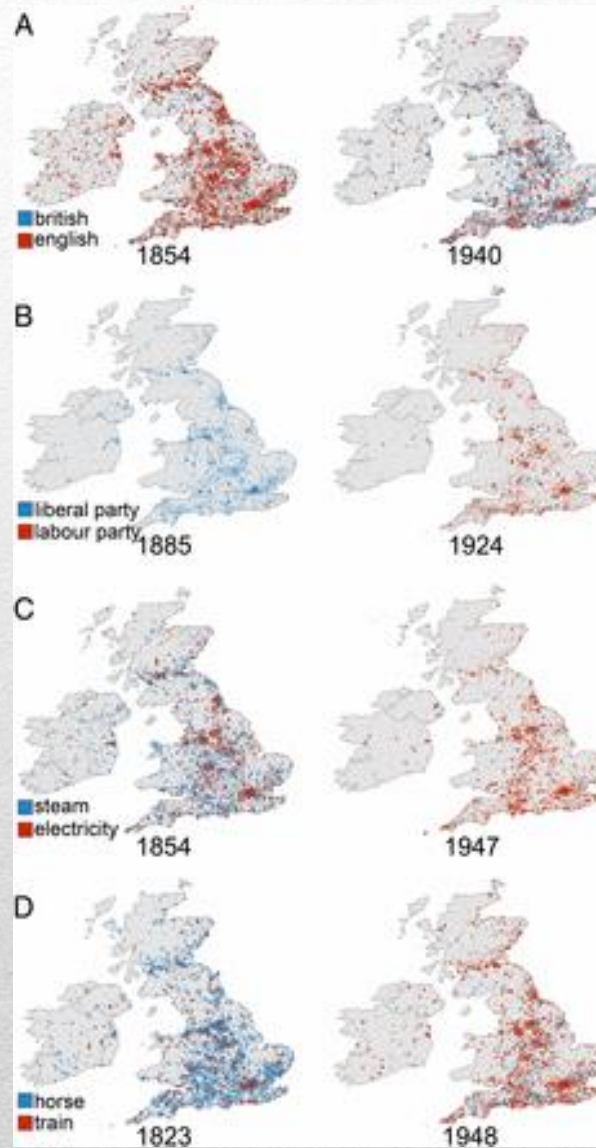
N-грамма

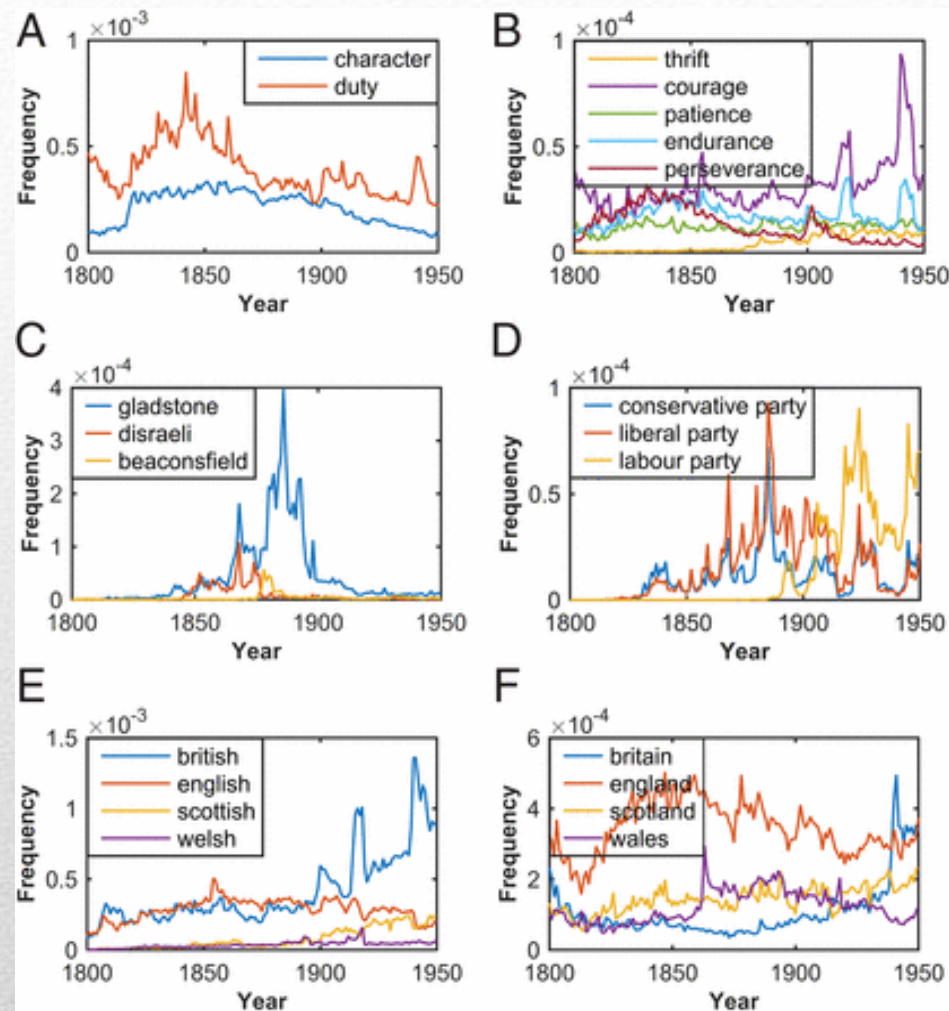
- 1-грамма - строка символов без пробелов: слова, цифры, опечатки ('wrods')
- N-грамма - последовательность 1-грамм: 'United Kingdom' (2-грамма), 'in the past' (3-грамма)
- Частота использования n-грамма = $\frac{\text{число экземпляра n-грамма в данном году}}{\text{общее количество слов в этом году}}$

N-грамма

- Максимальное количество элементов в n -грамме - 3
- 3-грамма должна появиться более 10 раз

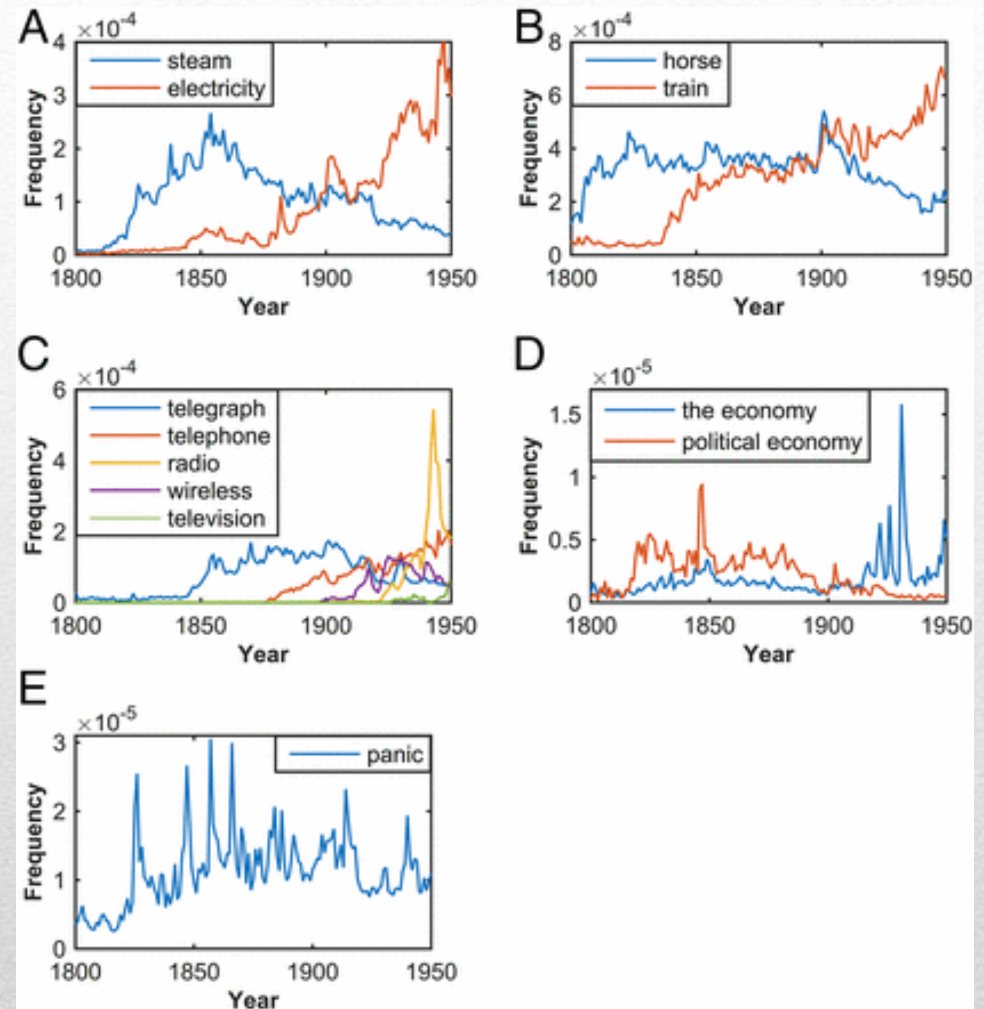
Выборка n -грамм



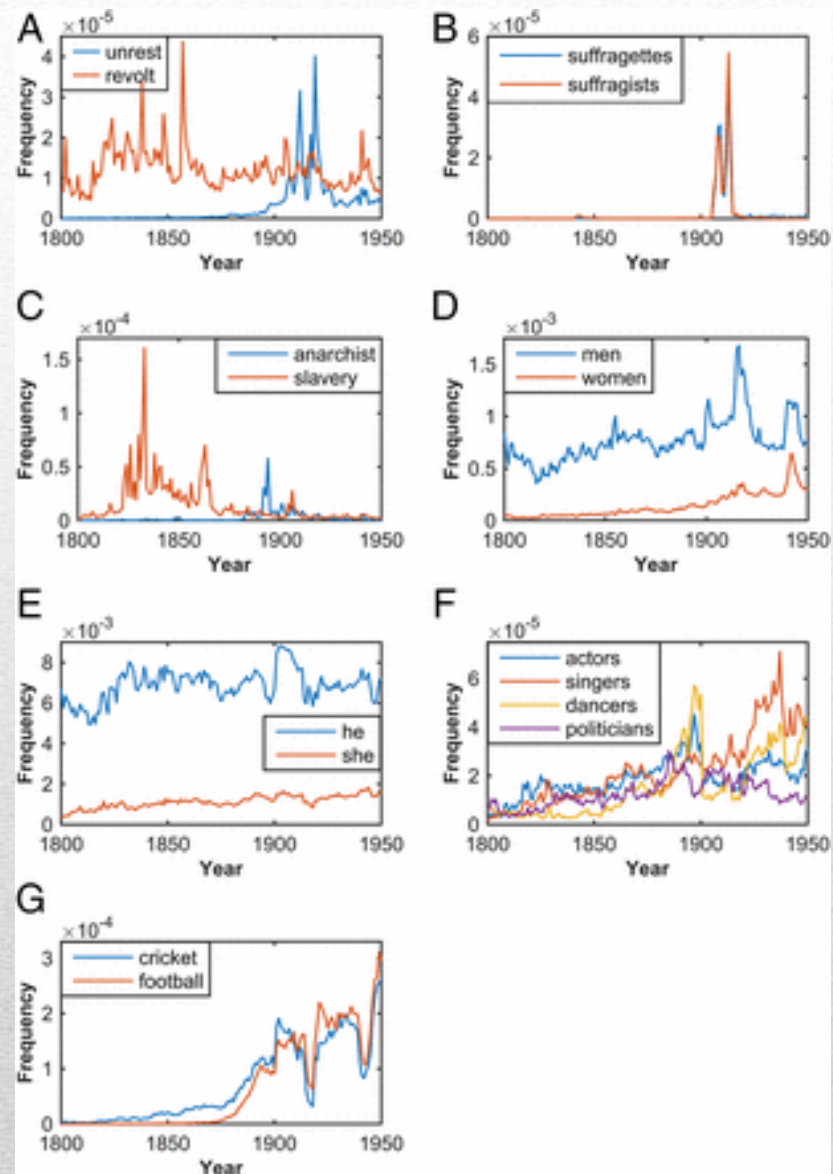


Гипотеза Гиббса и Коэна

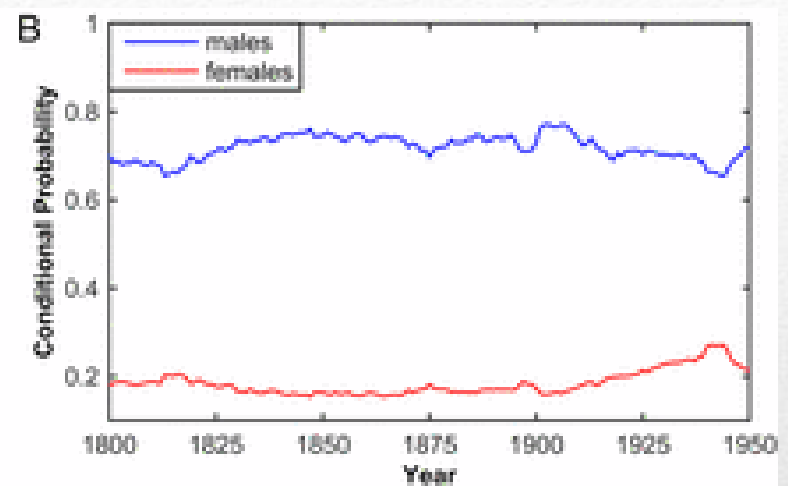
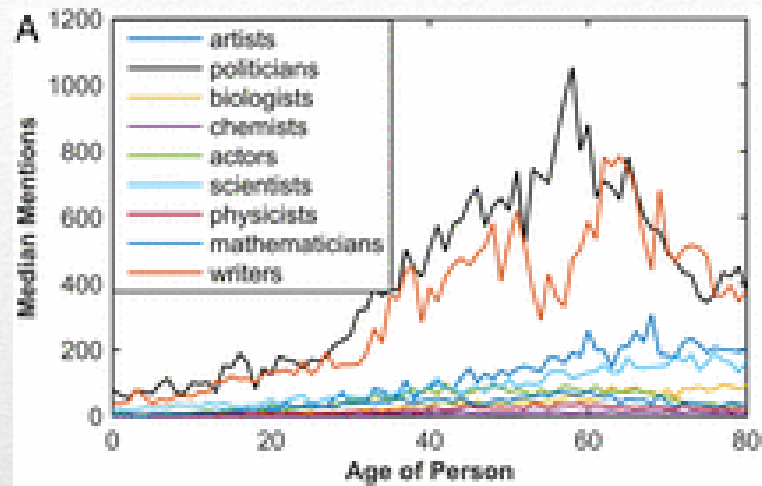
- Технологии
- Коммуникации
- Экономика
- Политика



- Права женщин
- Анархизм
- Популярная культура



- К людям, геолокациям и организациям привязываются определённые характеристики, программы
- Yago, DBpedia - это базы знаний с открытым исходным кодом. Он автоматически извлекается из Википедии и других источников



- Актуальность (вывод о гендерных стереотипах в современных СМИ)
- Аккуратность в выборке n-грамм (обладают высокой чувствительностью, но не слишком восприимчивы к семантическим сдвигам и ошибкам распознавания)
- Возможность улучшения и продолжения исследования ("шумоподавление", поиск связей с другими источниками и корпусами данных, устранение неоднозначности данных)
- Репрезентативная выборка изданий
- Использование проверенных методов
- Тщательный анализ полученных результатов
- Проверка полученных машинами данных самими исследователями (вычитка некоторых статей)

Плюсы исследования (минусов, на мой взгляд, нет)
