Кейсы применения text mining в исследованиях

Александр Бызов

Что такое text mining?

- **Text mining** – это открытие компьютером новой, ранее не известной информации путем автоматического ее извлечения из различных письменных источников. Ключевой элемент этого метода – соединение извлеченной информации таким образом, чтобы получить новые факты или гипотезы, которые в дальнейшем будут исследоваться более конвенциональными методами… (M. Hearst, professor in the School of Information, Berkley)
- **Text mining** включает в себя применение следующих техник: извлечение информации, обработка естественного языка и майнинга данных… (National Centre for Text mining)

Цит. по: Lin Y. Trans-disciplinarity and Digital Humanity: Lessons Learned from Developing Text Mining Tools for Textual Analysis// Understanding Digital Humanities/ Ed.by D. M. Berry. UK, Basingstoke: Palgrave Macmillan, 2012

Почему text mining важен?

- Один из способов работать с нереактивными данными ☺
- Позволяет обрабатывать большие неструктурированные текстовые данные;
- Результаты таких исследований воспроизводимы;
- Скорость проведения исследований;

Кейс 1: Интернет-исследования в социологических журналах

- **Исследовательские вопросы:** Положение интернет-исследований, основные темы, популярные ключевые слова и их изменение во времени, теоретические ориентации и методы интернет-исследований;
- **Выборка:** 27340 статей (авторы, названия статей и журналов, аннотации, ключевые слова и цитирования);
- **Основные методы:** кластеризация, частотный анализ

Peng T. Q. et al. Mapping the landscape of Internet studies: Text mining of social science journal articles 2000–2009 //New Media & Society. 2013. Vol. 15. №. 5. P. 644-664.

Кейс 1: Примеры результатов

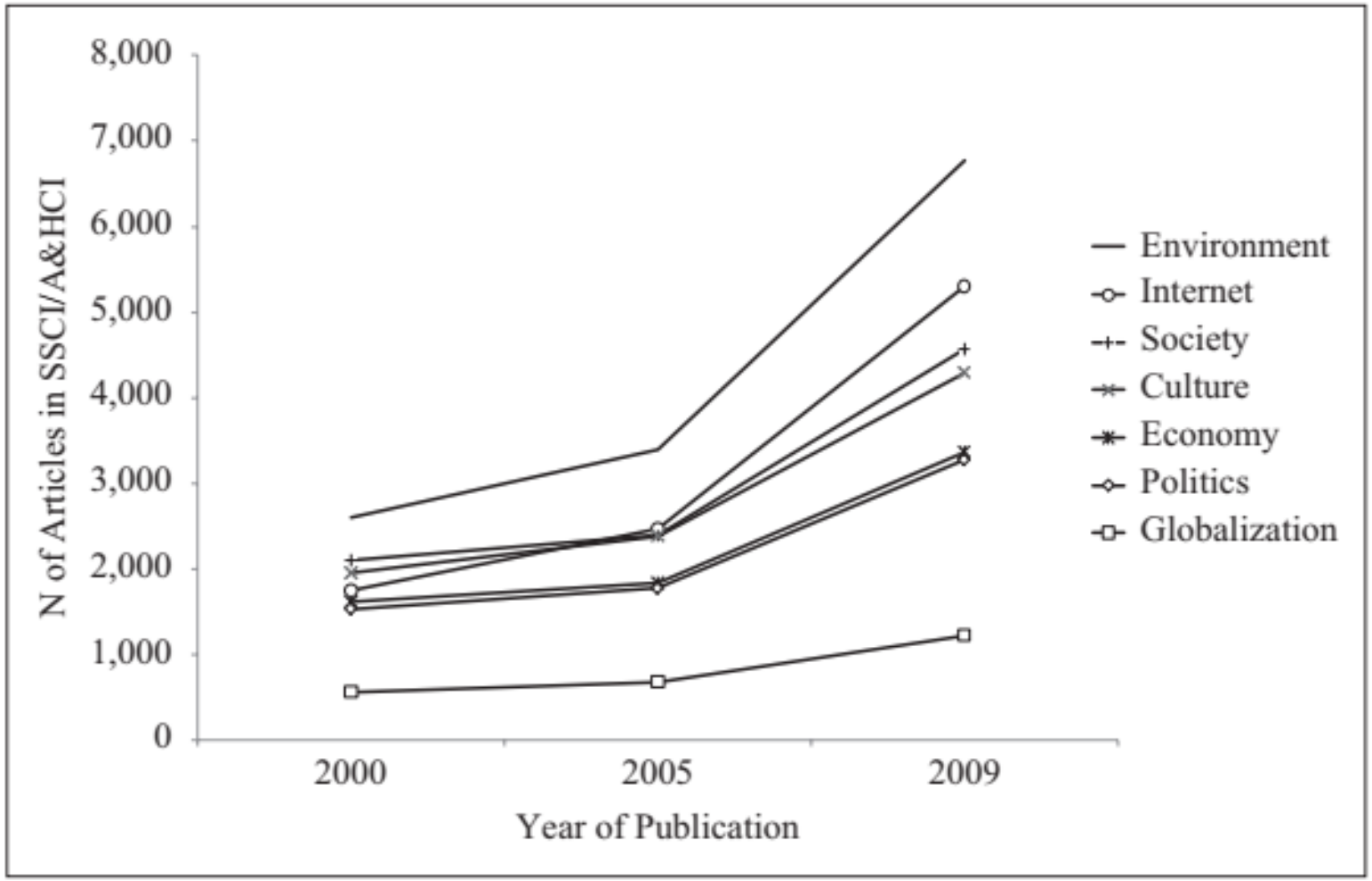
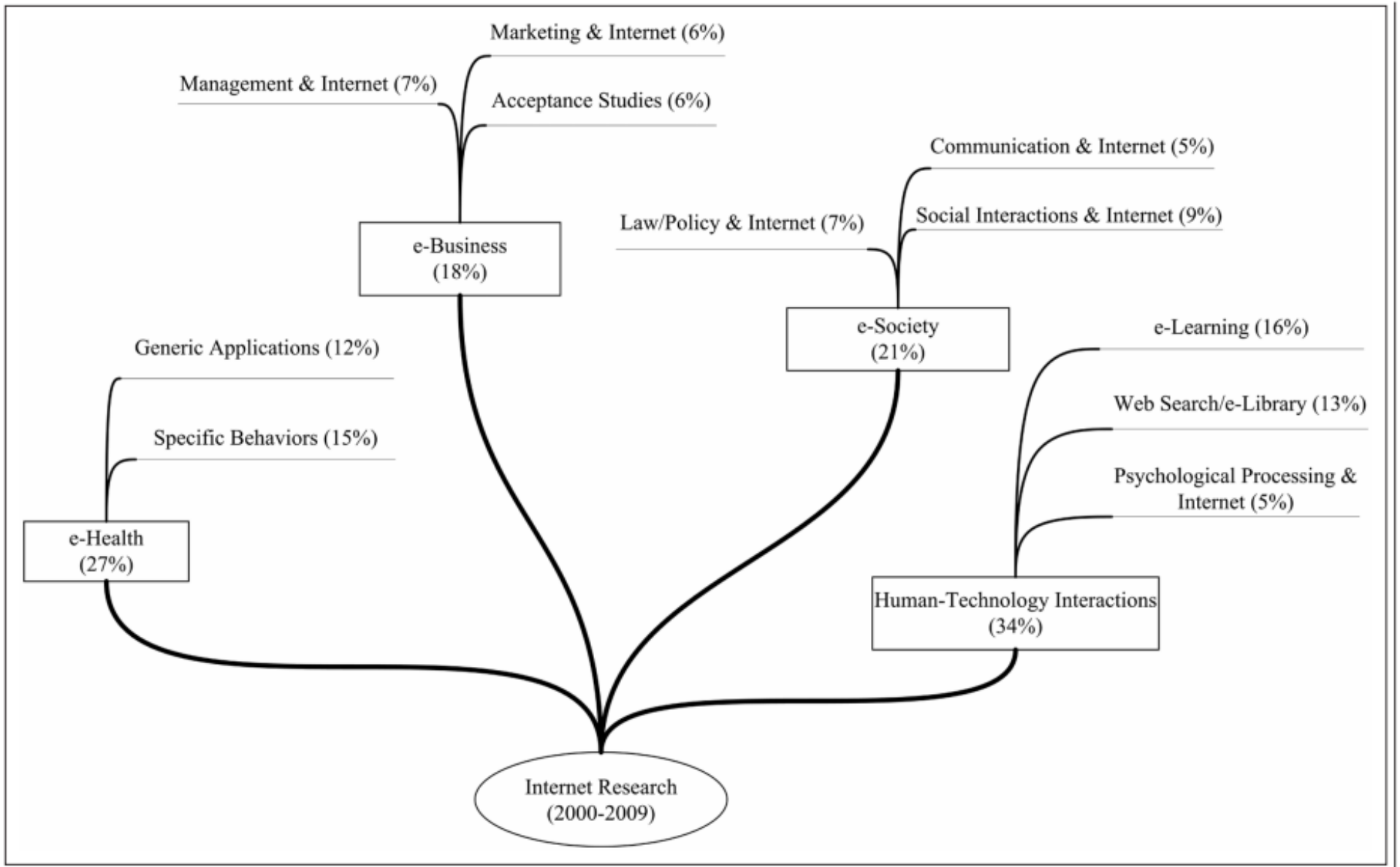


Figure 1. Number of publications in Social Sciences Citation Index (SSCI)/Arts & Humanities Citation Index (A&HCI) journals with seven central keywords in respective fields.

Кейс 1: Примеры результатов



Кейс 1: Ограничения

> Аннотации не обязательно отражают все темы статьи;

> «Эффект картотеки» - не учитываются, напр., аннотации конференций

Кейс 2: Использование текстов для исследования структуры и характеристики сетей

- **Исследовательские вопросы:** можно ли раскрыть сетевую структуру радикальных активистских групп, участвующих в энвайроменталистских движениях? Какие группы (подсети) более склонны к радикальным или мейнстримным формам протеста?
- **Выборка:** Журнал Do or Die, 10 выпусков (разбиты по 12 предложений)
- **Основные методы:** кластеризация, сетевой анализ

Almquist Z. W., Bagozzi B. E. Using radical environmentalist texts to uncover network structure and network features //Sociological Methods & Research. 2015.

Кейс 2: Примеры результатов

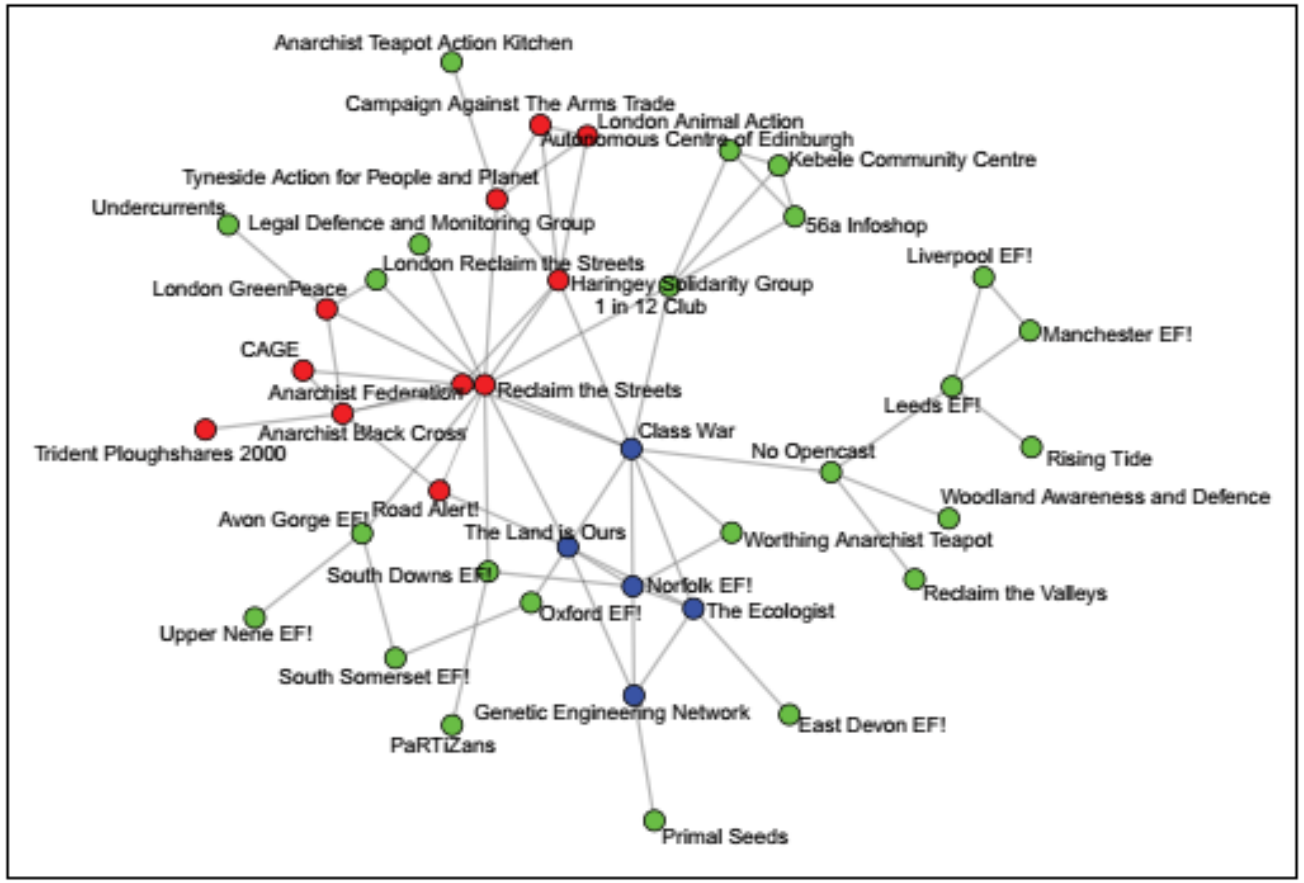


Figure 3. Network plot of the connected component of the 12-sentence network colored for cluster assignment.

Кейс 2: Примеры результатов

Topic	Top 20 Words	Labels
1	one, made, anoth, everi, side, time, hand, mani, turn, left, open, togeth, enough, around, given, reach, great, strong, run, join	Inspirational Language
2	role, polit, organis, ideolog, non, activ, idea, movement, radic, mainstream, question, organi, opinion, individu, media, violent, problem, engag, potenti, posit	Group Identity Debates
3	oil, compani, mine, crop, farmer, fish, indigen, papua, corpor, genet, govern, zapatista, shell, bougainvill, peasant, mexico, engin, industri, western, indonesian	Neocolonialism
4	book, isbn, publish, box, http, zine, magazin, guid, press, copi, write, articl, send, email, read, page, list, web, review, mail	Eco−Literature
5	women, law, case, footbal, men, legal, game, evid, court, school, terrorist, privat, famili, properti, terror, intellig, record, team, agenc, custom	News and Culture
6	speci, habitat, restor, wildlif, peat, bird, ecosystem, biodivers, plant, soil, extinct, woodland, highland, conserv, hotspot, moor, natur, garden, biolog, wild	Species Conservation
7	let, away, will, back, voic, danger, keep, put, take, look, etc, earth, thank, mother, first, ever, fill, eye, die, sound	General Concern
8	forest, protect, log, water, environment, mountain, dam, north, nativ, area, east, island, region, rainforest, timber, ago, urban, northern, land, river	Land Conservation
9	polic, arrest, cop, crowd, march, banner, vehicl, window, offic, smash, demo, confer, riot, hour, demonstr, mask, bank, pragu, car, camera	Violent Protest
10	action, direct, campaign, group, involv, network, sabotag, anti, reclaim, tactic, meet, success, event, sab, act, target, rts, opencast, media, aim	Direct Action/Ecotage
11	camp, evict, site, road, quarri, twyford, tunnel, sit, council, tree, climb, squat, set, hill, fenc, hous, build, tarmac, built, trash	Occupation/Camps
12	cultur, must, societi, exist, life, learn, live, desir, civilis, understand, human, relationship, skill, can, experi, planet, process, sens, domin, alien	Sustainable Societies
13	pirat, prison, sentenc, black, panther, bomb, murder, jail, ship, imprison, kill, shepherd, sent, africa, gun, white, death, frame, fbi, trial	International Terror
14	know, dont, realli, want, thing, that, think, lot, someth, get, someon, sure, there, say, just, bad, theyr, didnt, ask, thought	Admonishments
15	capit, capitalist, labour, revolut, class, struggl, global, economi, union, counter, worker, social, econom, elit, resist, globalis, autonomi, wage, spanish, market	Anti−Capatalist Left

Кейс 2: Ограничения

> Ограничения выборки;

> Случайность выбора критерия разделения текстов на документы;

Кейс 3: Исследования стереотипов

- **Исследовательские вопросы:** можно ли использовать модели укоренения слов (word embeddings) для исследования гендерных и этнических стереотипов?
- **Выборка:** Google News dataset, Corpus of Historical American English, New York Times Annotated Corpus
- **Основные методы:** word embeddings

Garg N. et al. Word embeddings quantify 100 years of gender and ethnic stereotypes //Proceedings of the National Academy of Sciences. 2018. Vol. 115. №. 16. P. E3635-E3644.

Кейс 3: Примеры результатов

Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

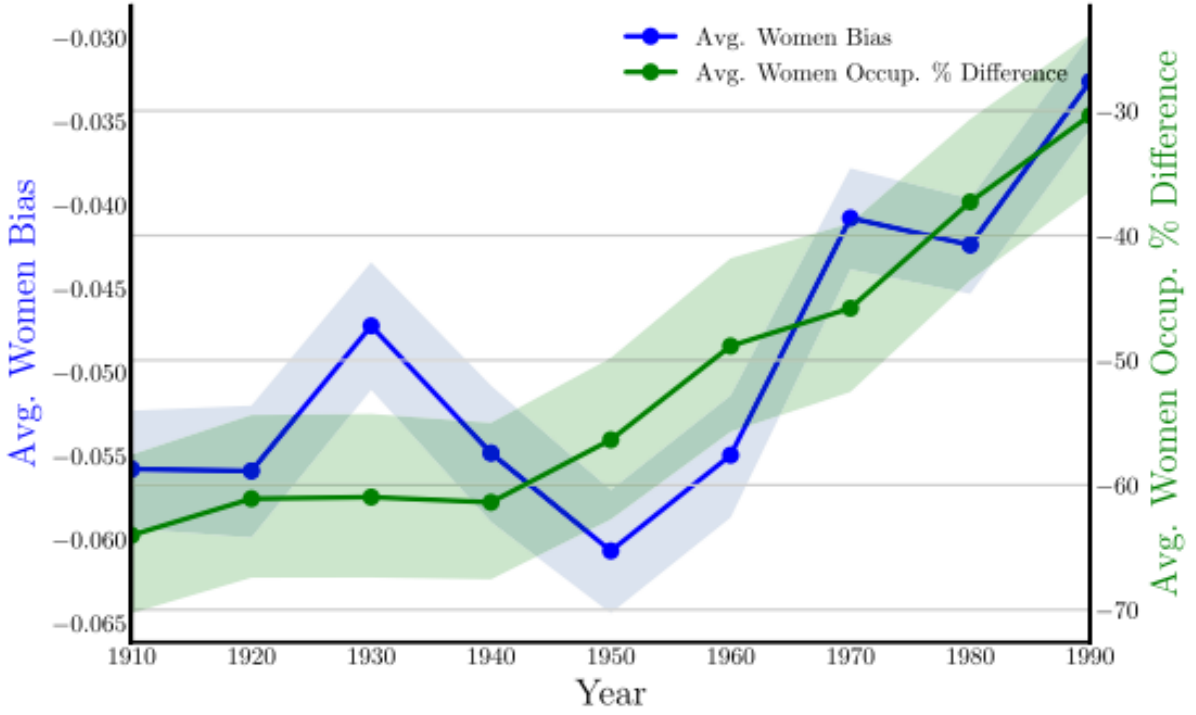
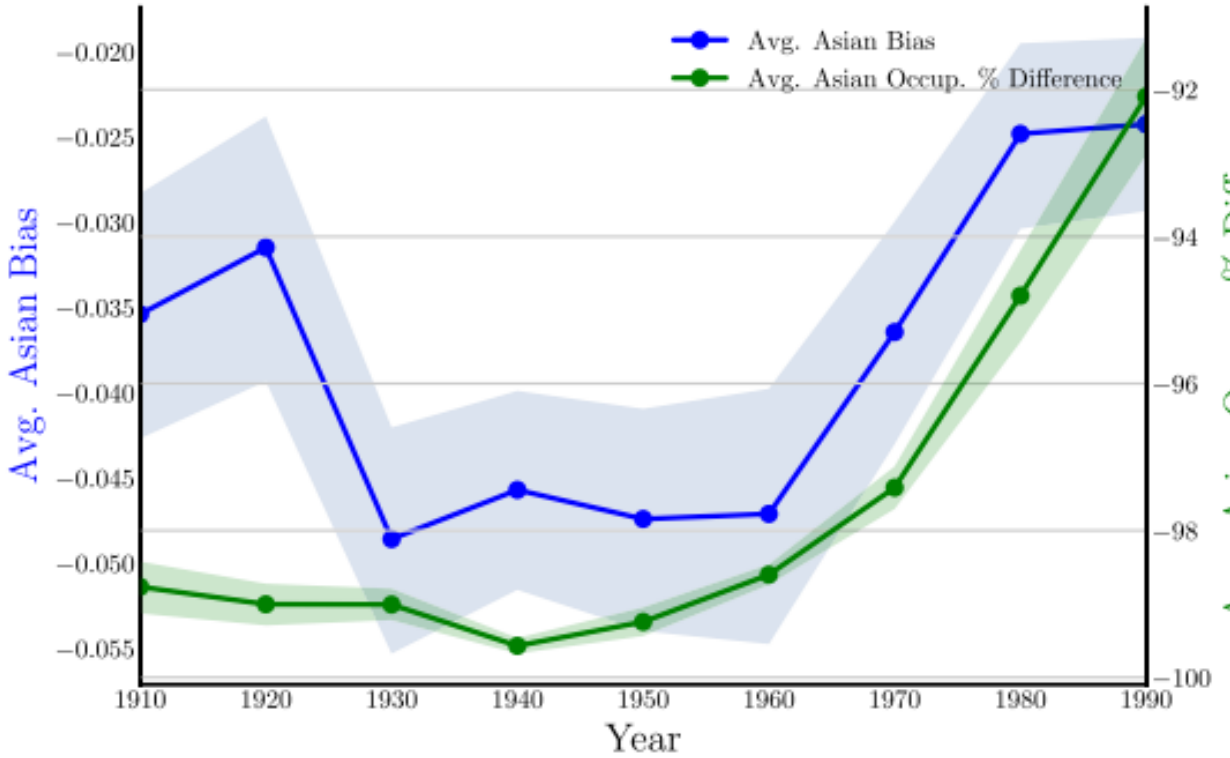


Fig. 2. Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.

Кейс 3: Примеры результатов

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer



Кейс 3: Примеры результатов

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Table 3. Top Asian (vs. White) adjectives in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

Кейс 3: Ограничения исследования

> Выбор метрик;

> Список использованных слов;

> Word embeddings – метод «черный ящик»

Как помирить text mining с социологической методологией?

- Использование дихотомий и метатеоретических схем?
- Ignatow: реалистская конструкционистская онтология. Дискурсы – это онтологически реальные эмерджентные социальные сущности (entities) и имеют каузальную связь с недискурсивными социальными и когнитивными процессами;
  - Человеческое кодирование;
  - Анализ метафор;
  - Анализ сентиментов, основанный на словарях;
- Натуралистская модель объяснения в социологии