

Introduction into NLP

Eugeny Malyutin

Rule-Based

Who am I to lecture you:

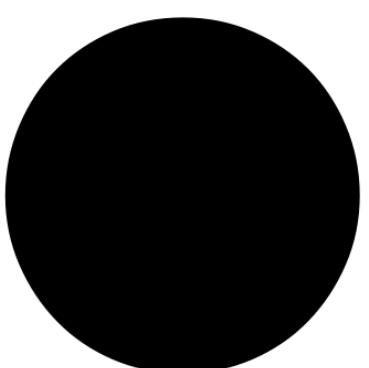
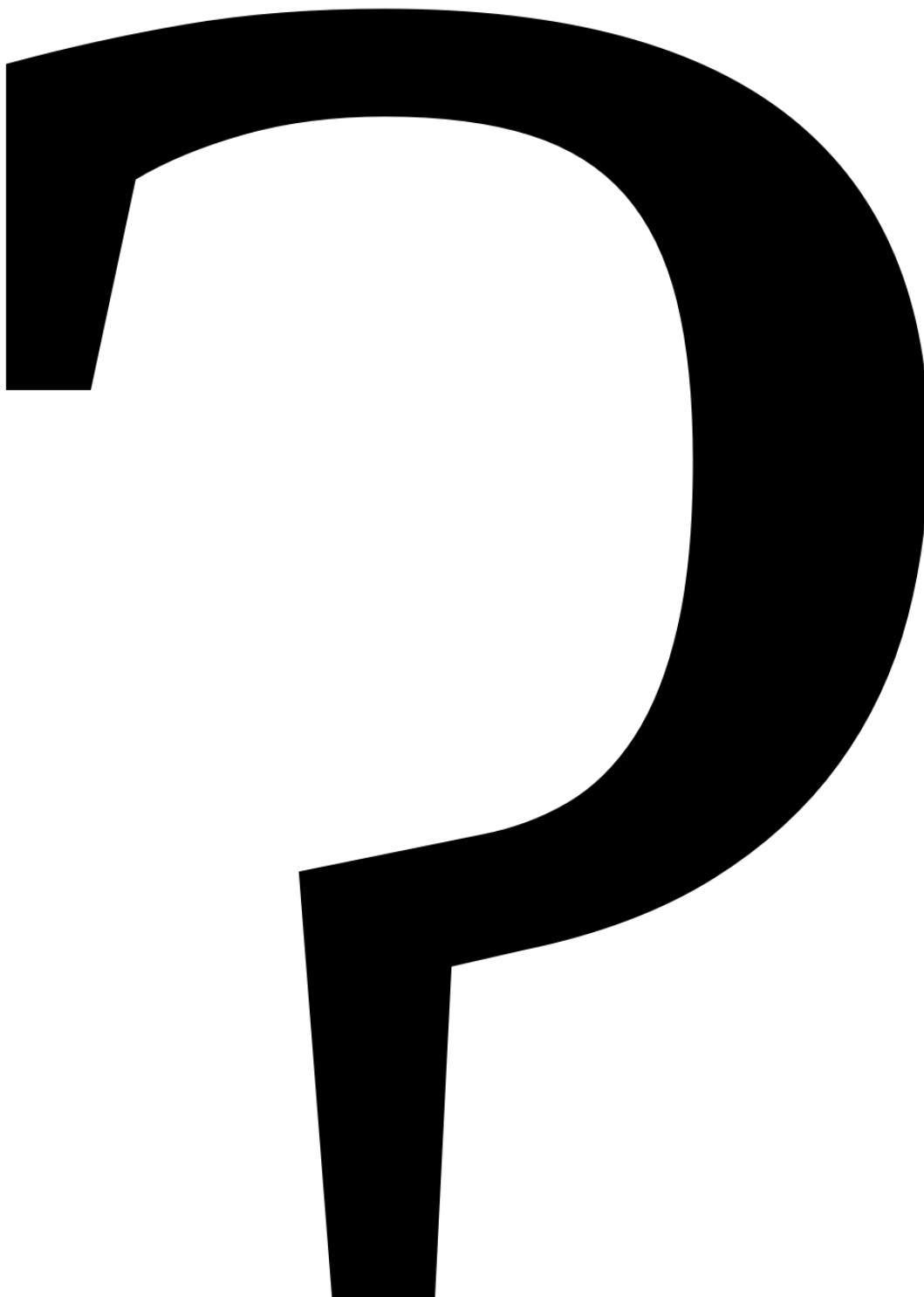
- AMCP SPbU
 - 2015 bach. (honours)
 - 2017 masters (honours),
 - 2017+ PhD st.)
- 2016 - 2020 Mail ru group:
 - Streaming textual trends detection at a scale
 - Local groups detection and topic modeling (as a supervisor)
 - Search engines

Final score:

- Your homeworks (labs) (40%)
- Exam (10%)
- CV Part (40% + 10%)

Why should someone learn NLP?

- It's fun
- It's paid
- Internet is made from text (and cats)

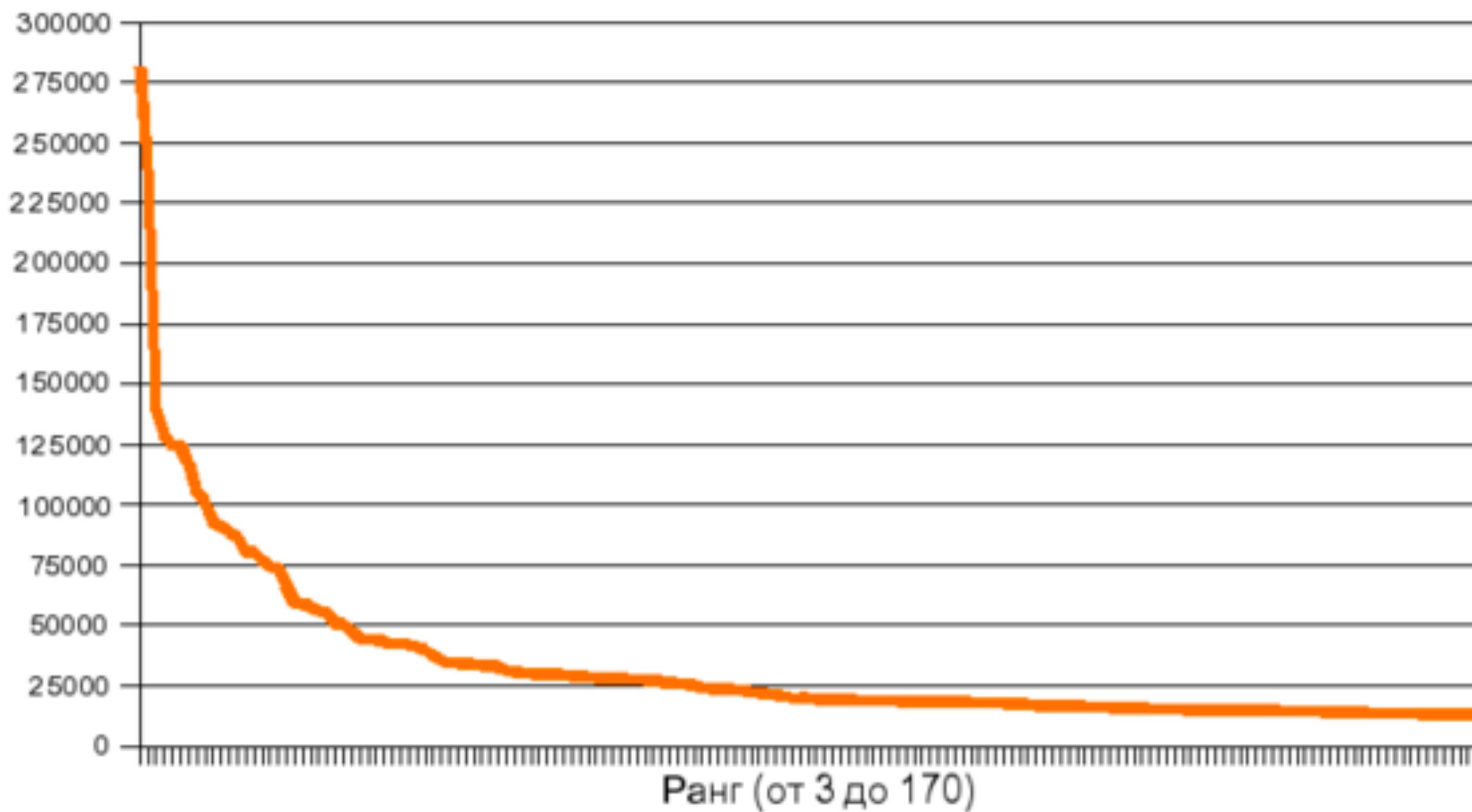


NLP tasks (incomplete list) :

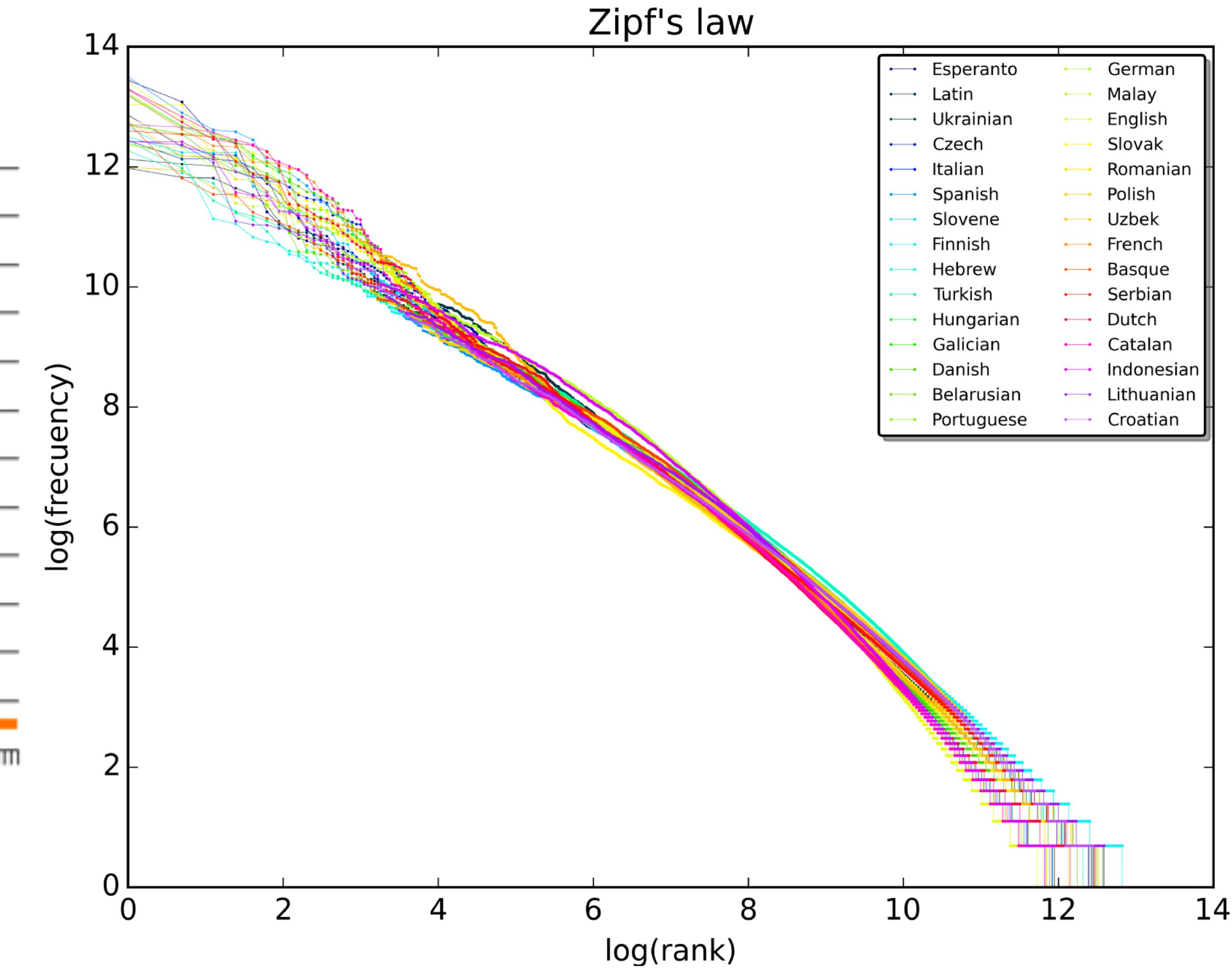
- Language modeling
- Part-of-speech tagging
- Named entity recognition
- Text (topic) classification
- Keyword extraction
- Spelling checking
- Syntax parsing
- Dependency parsing
- Machine translation
- Stemming Lemmatization
- Distributional semantics
- Text generation
- Text clustering
- Wikification
- QA systems
- dialogue systems
- Plagiarism detection
- Morpheme analysis
- Grammar check
- Relation extraction
- Entity linking
- Sentiment analysis
- Topic modeling
- Text summarization
- Semantic role labeling
- Information retrieval
- Speech Recognition / Synthesis

Zipf law [1949]: rank * freq ~ const

Частота



Russian corpus frequency



Zipf law [1949]: rank * freq ~ const

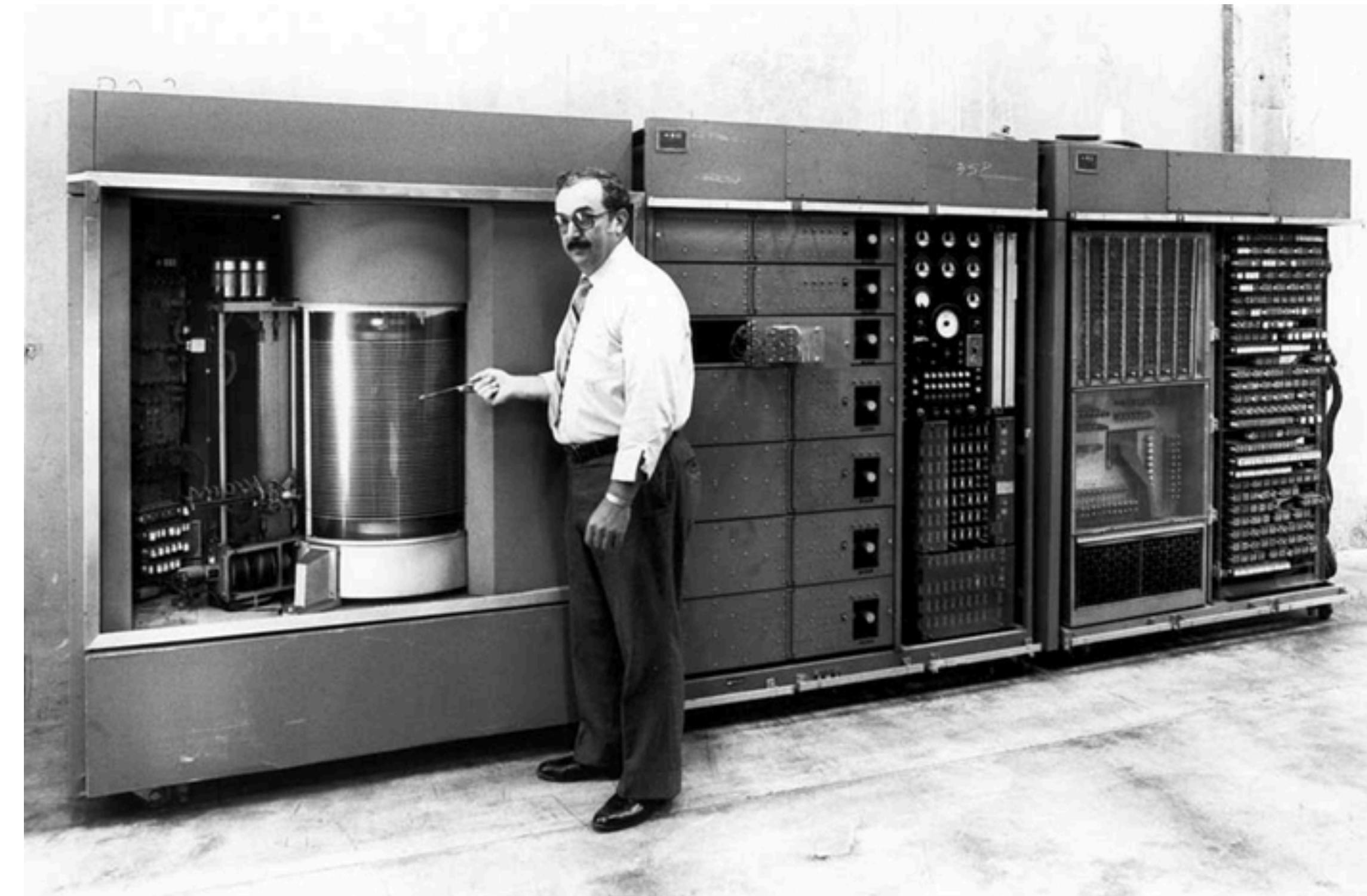
- Works for city population in modern countries. (Is it correct for Russia?)

$$P_n = P_1/n$$

- There is an opinion that Zipf's law is not about text semantics but pure statistical phenomena. Still works as good heuristic for us.

Back to the roots

- Established opinion: CompLing was born in 1950s thanks to the machine translation task (both in the USA and the USSR)
- Georgetown experiment 60+ sentences, 6 rules 1954, IBM + Georgetown University, "the task will be solved in 3-5 years"
- 20M \$ and 12 years — human's still cheaper
- In 1966 USA gov bans all grants to NLP researchers for 20 years due to ALPAC report.



Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyusu.

Magnitude of angle is determined by the relation of length of arc to radius.

Myezhdunarodnoye ponyimaniye yavlyayetsya vazhnim faktorom v ryeshenyiyi polyityicheskix voprosov.

International understanding constitutes an important factor in decision of political questions

Progress

50-e: first attempts, Information Theory, Formal Grammars

60-70-e: ‘Syntactic Structures’, AI, bayesian models, first corpora

80-e: first distributional semantics approaches, data-driven research

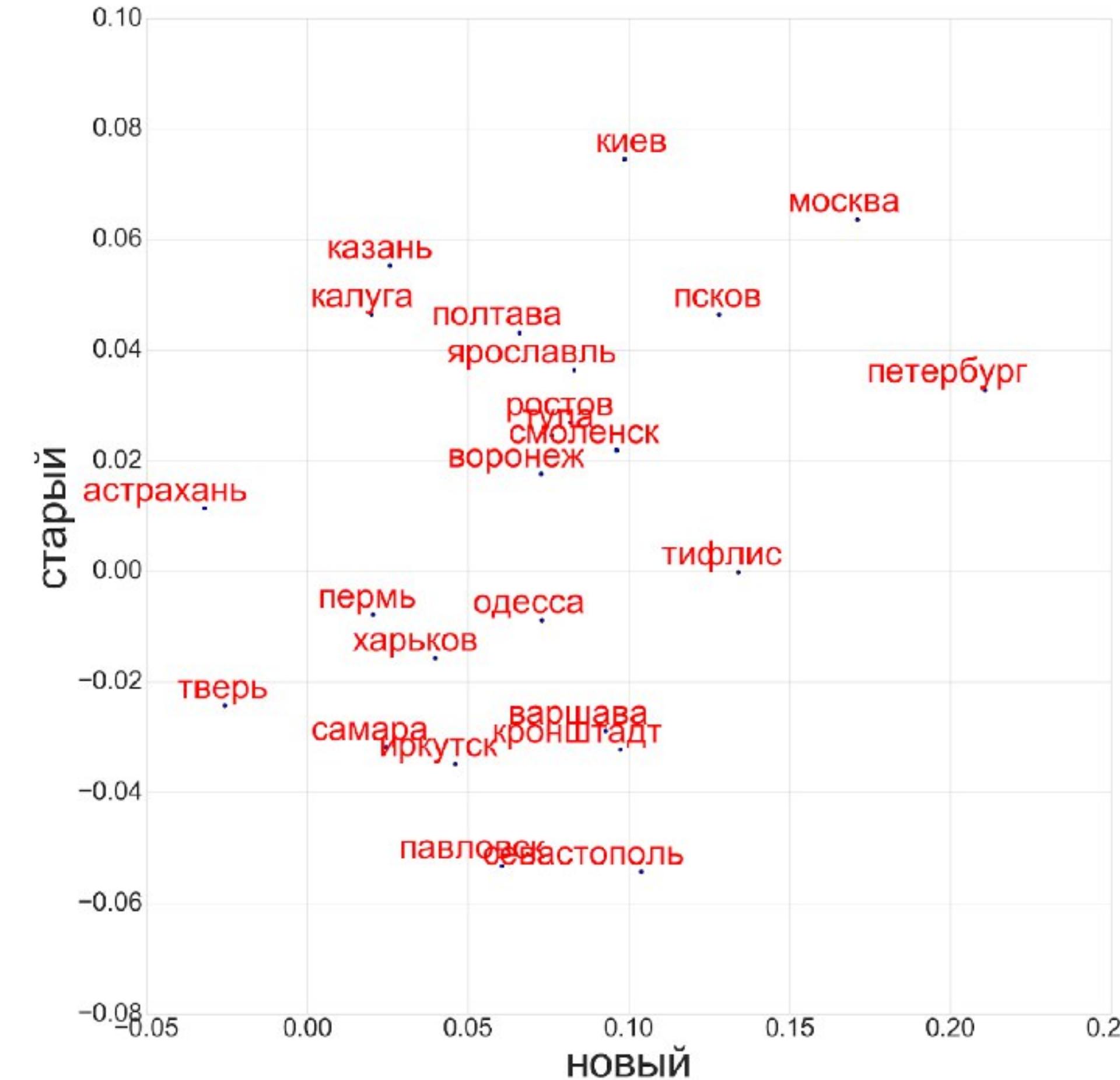
90-e: models evaluation tracks, applications for wide range of users

2000-e: web! data! machine learning, unsupervised approaches

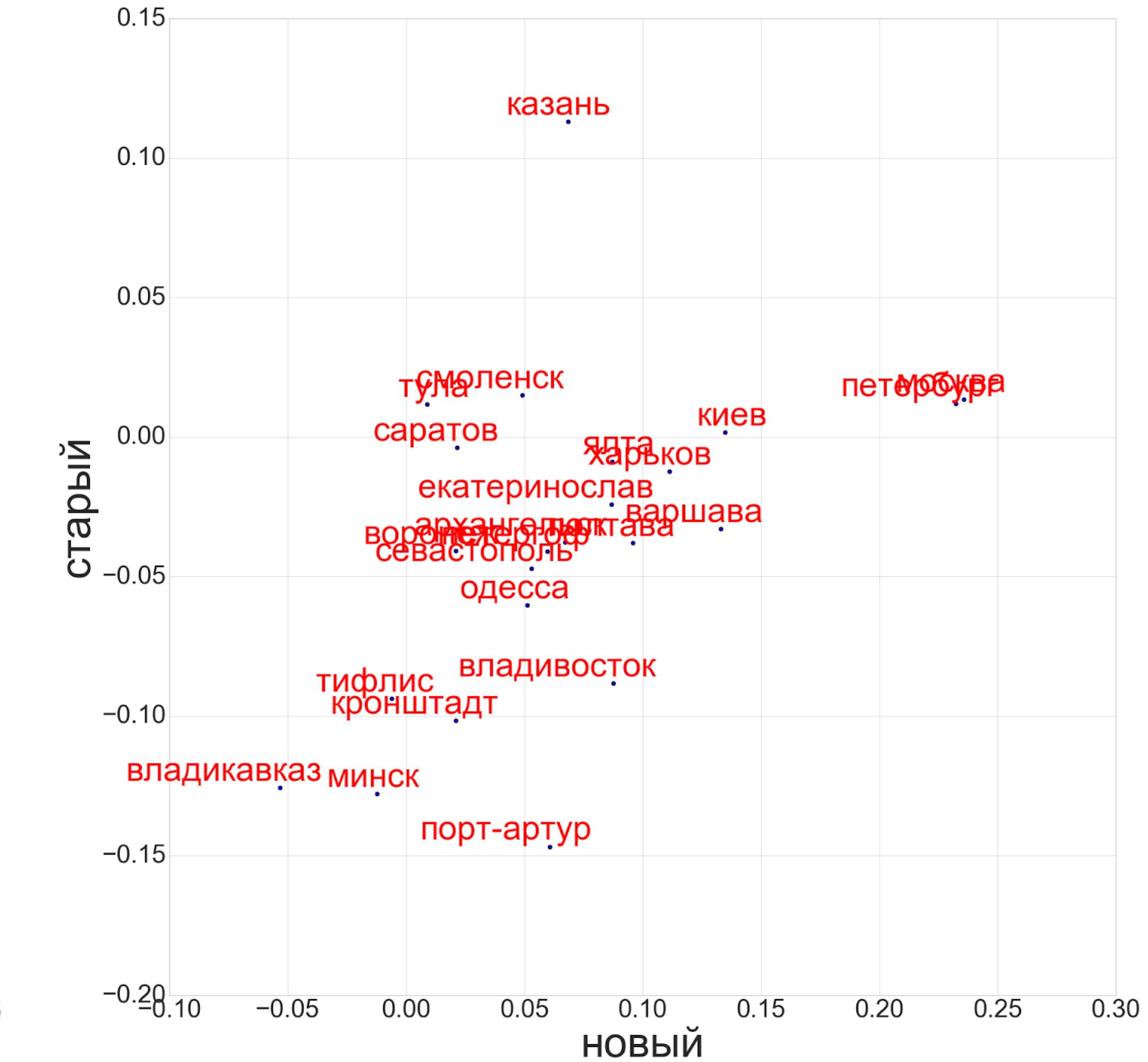
2010-e: Deep Learning feast, tons of applications in various fields

Cases: digital humanities

**Word2Vec Rus Corpora
projection on axes
represented by words
«new» and «old»**



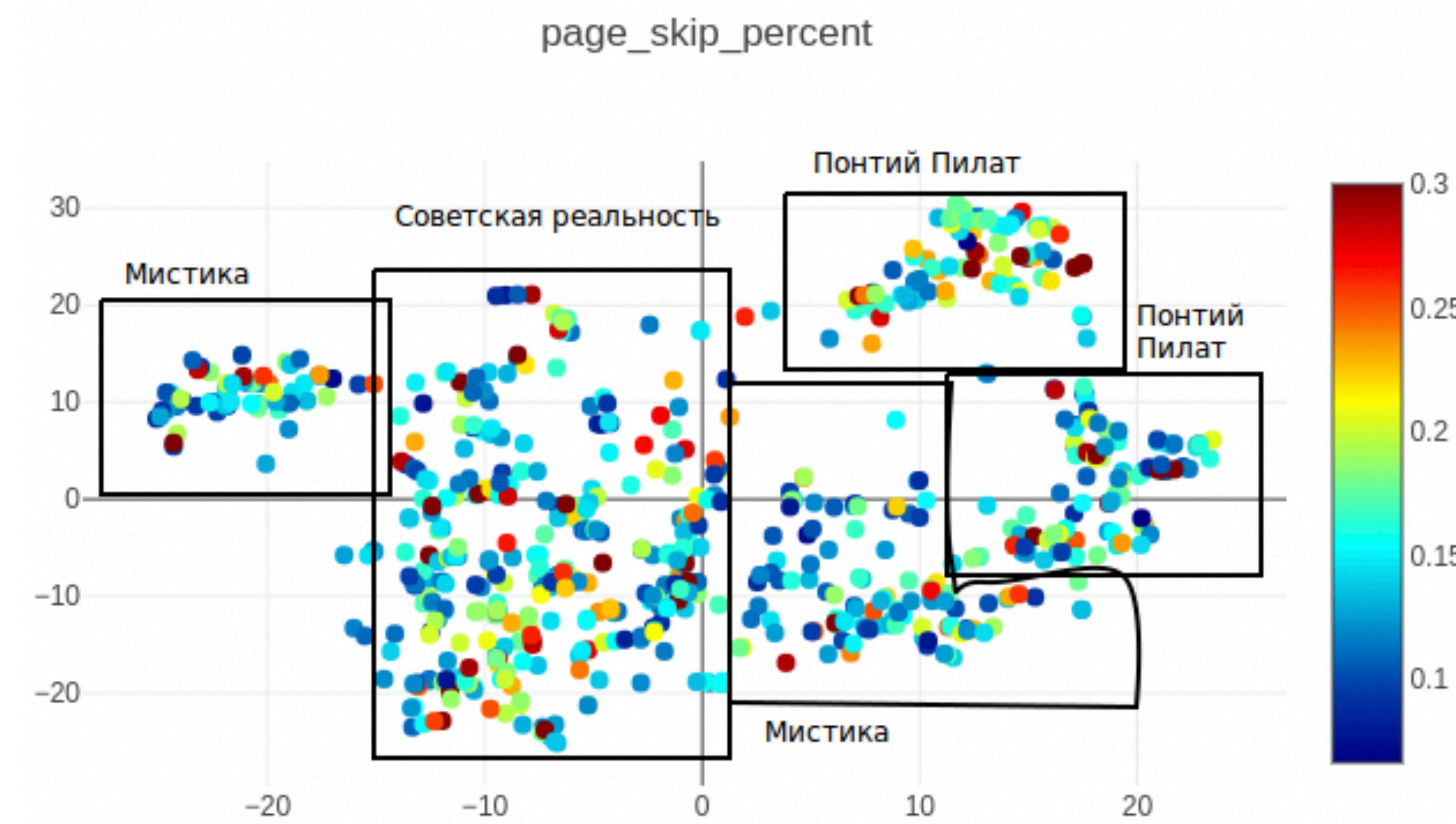
НКРЯ 1897-1916



НКРЯ 1917 -1929

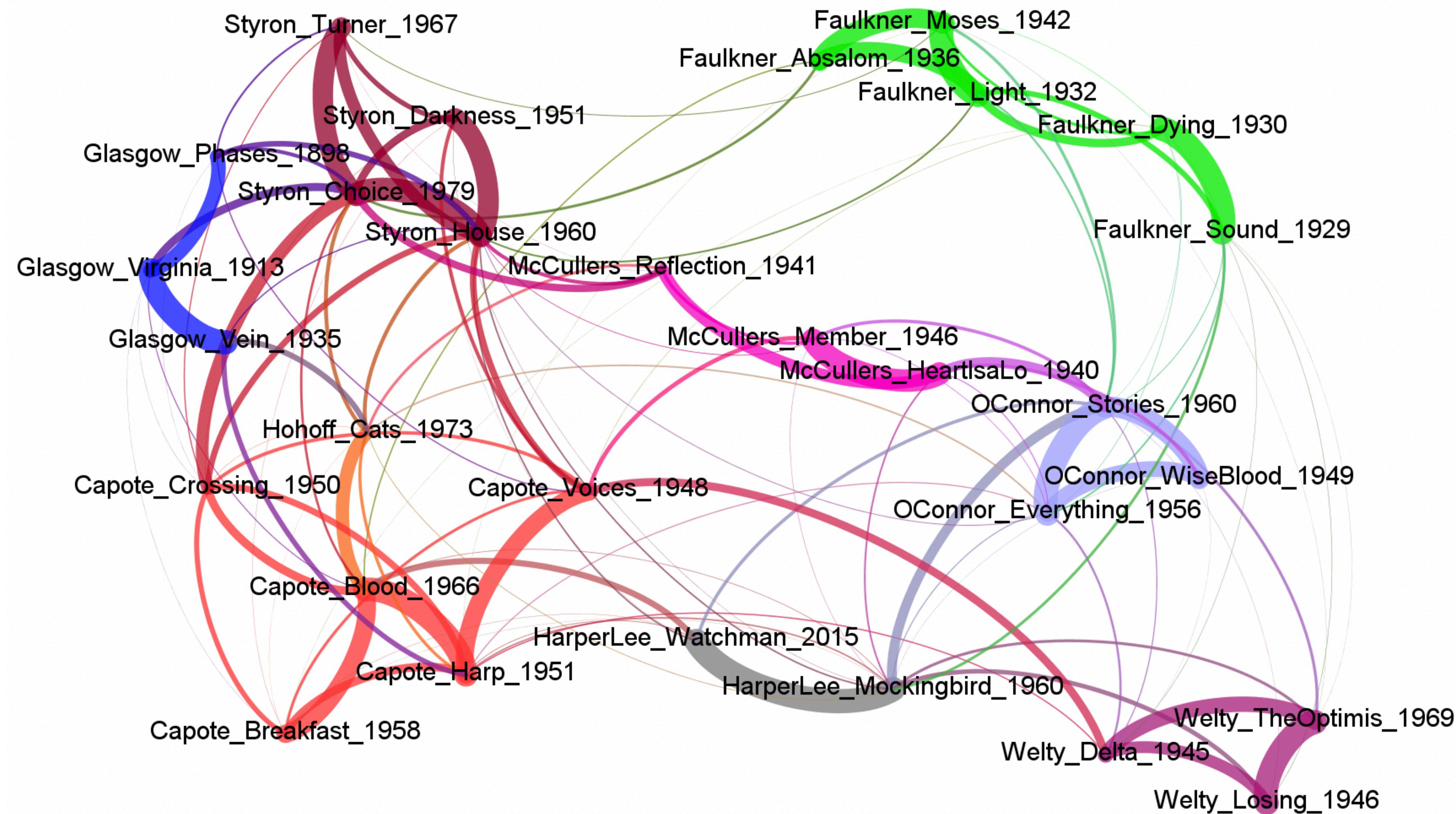
https://t.me/dh_hse

Cases: digital humanities



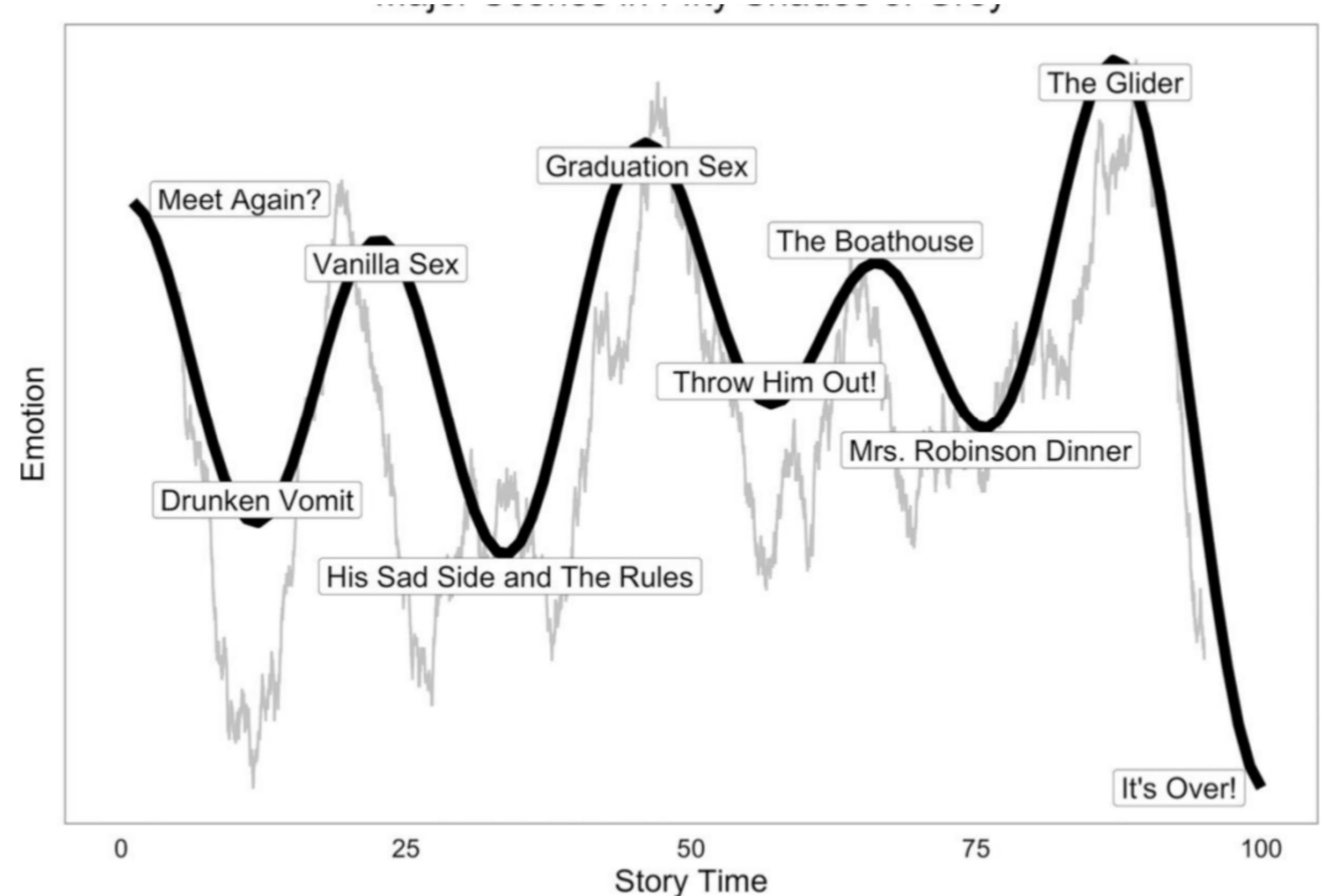
T-SNE clustering for word2vec embeddings for pages of «The Master and Margarita»

Cases: stylometry

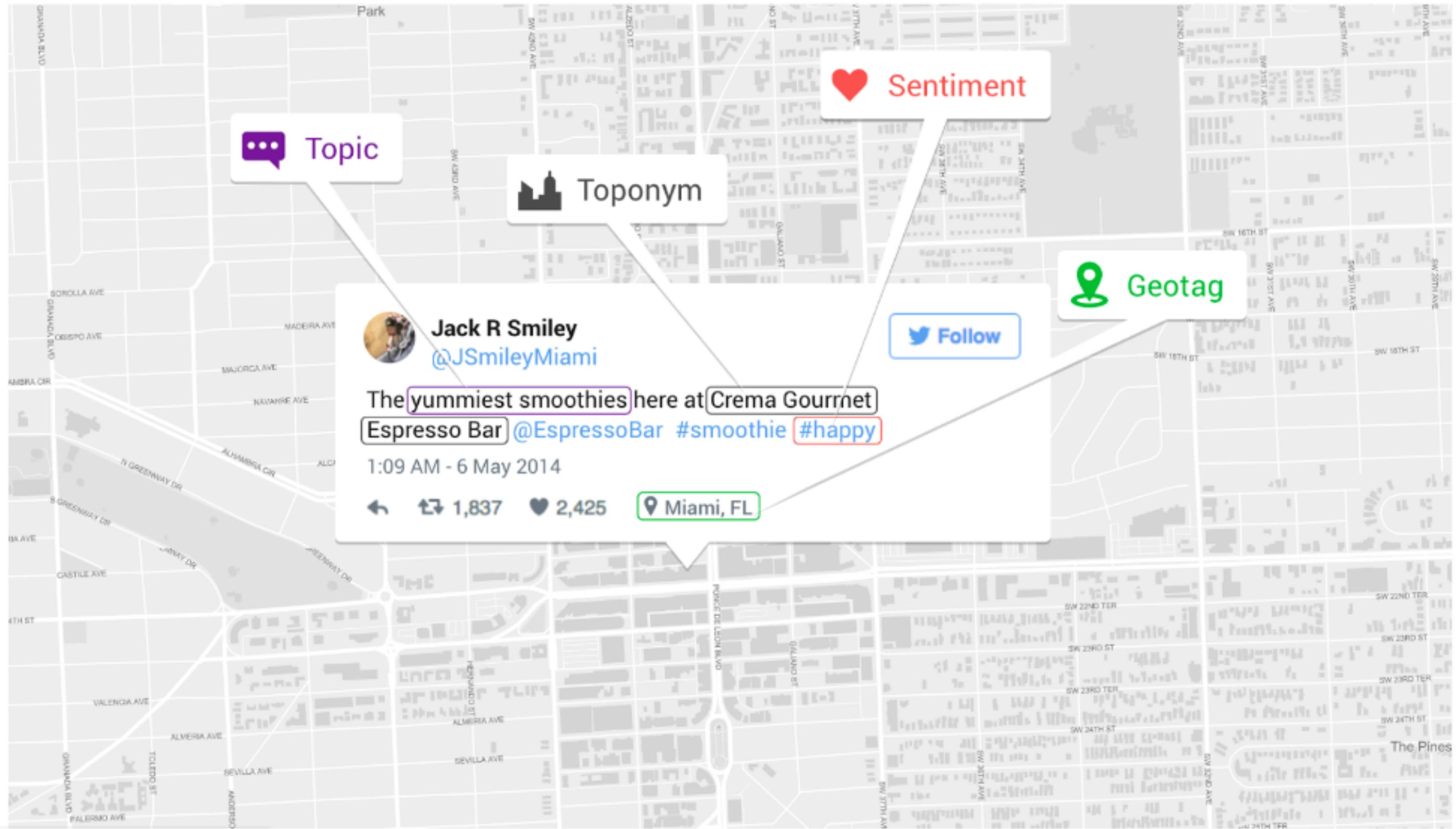


Cases: tonality plot

Matthew Jockers and «50 shades of black» tonality graph.



Cases: urban studies



Cases: urban studies

MUSIC WORK INTERNET

FOOD ENTERTAINMENT

WATCHING SPORTS EVENTS

 enjoying some amazing #macarons at #janetteandco #nutella #venezuelanchocolate

 at burgerfi for awesome burgers! #burgerfi

 can't wait to dig into this yummy food @carnaval_miami !

MUSIC WORK INTERNET

FOOD ENTERTAINMENT

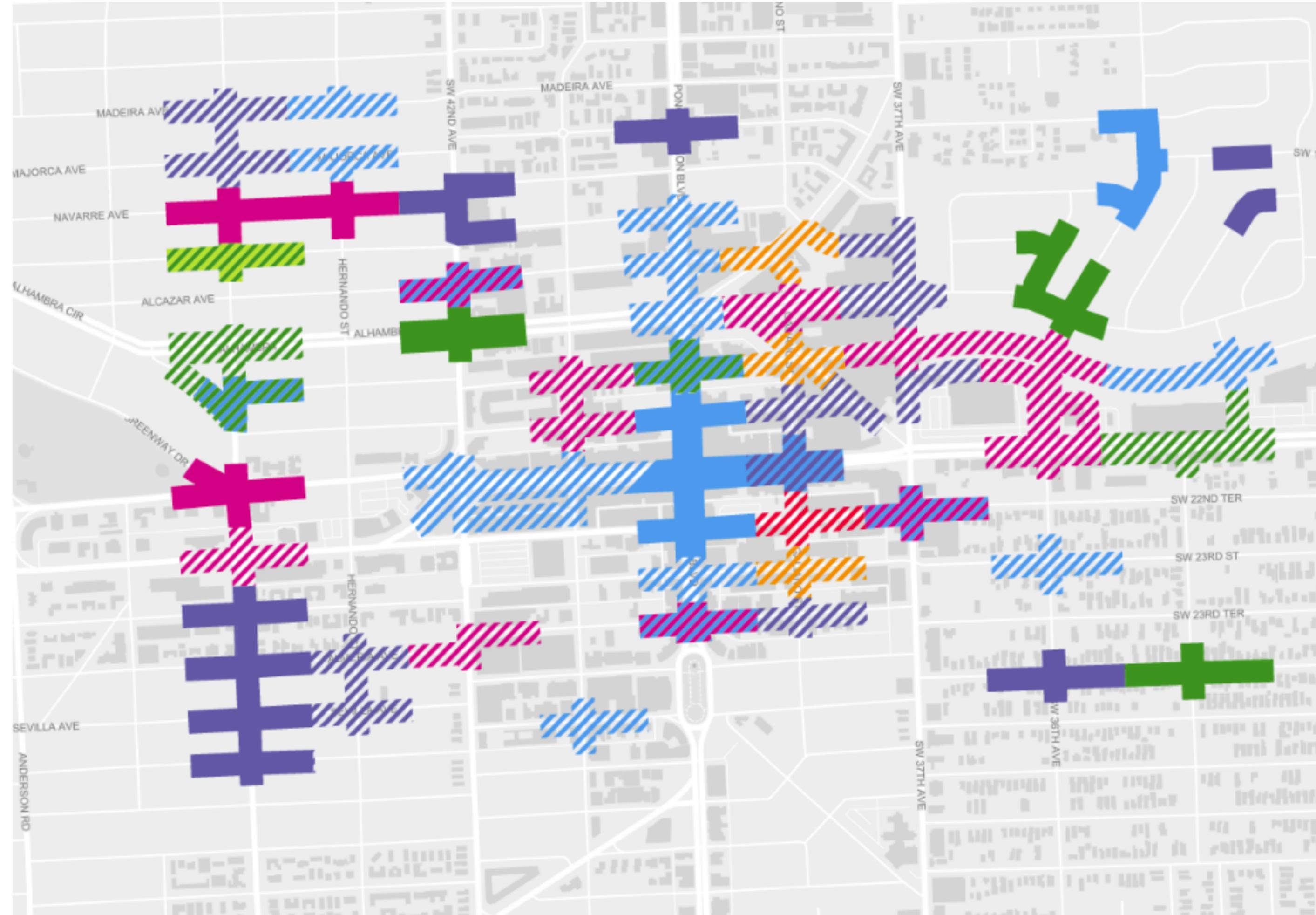
WATCHING SPORTS EVENTS

 @arionnation he has been our best bench guy this year (welp), but he has a lot of passed balls, which piss people off. also, he isn't gattis

 anybody who has a choice this week and chooses chip/joe over vin scully to watch braves at dodgers is doing baseball all wrong

 @mikenewmanrs lol, the best humor comes from pain (fortunately here, that is just baseball fan pain, not real life pain)

▶ ▶ Слайд 20 ⏪ ⏩ ⚙

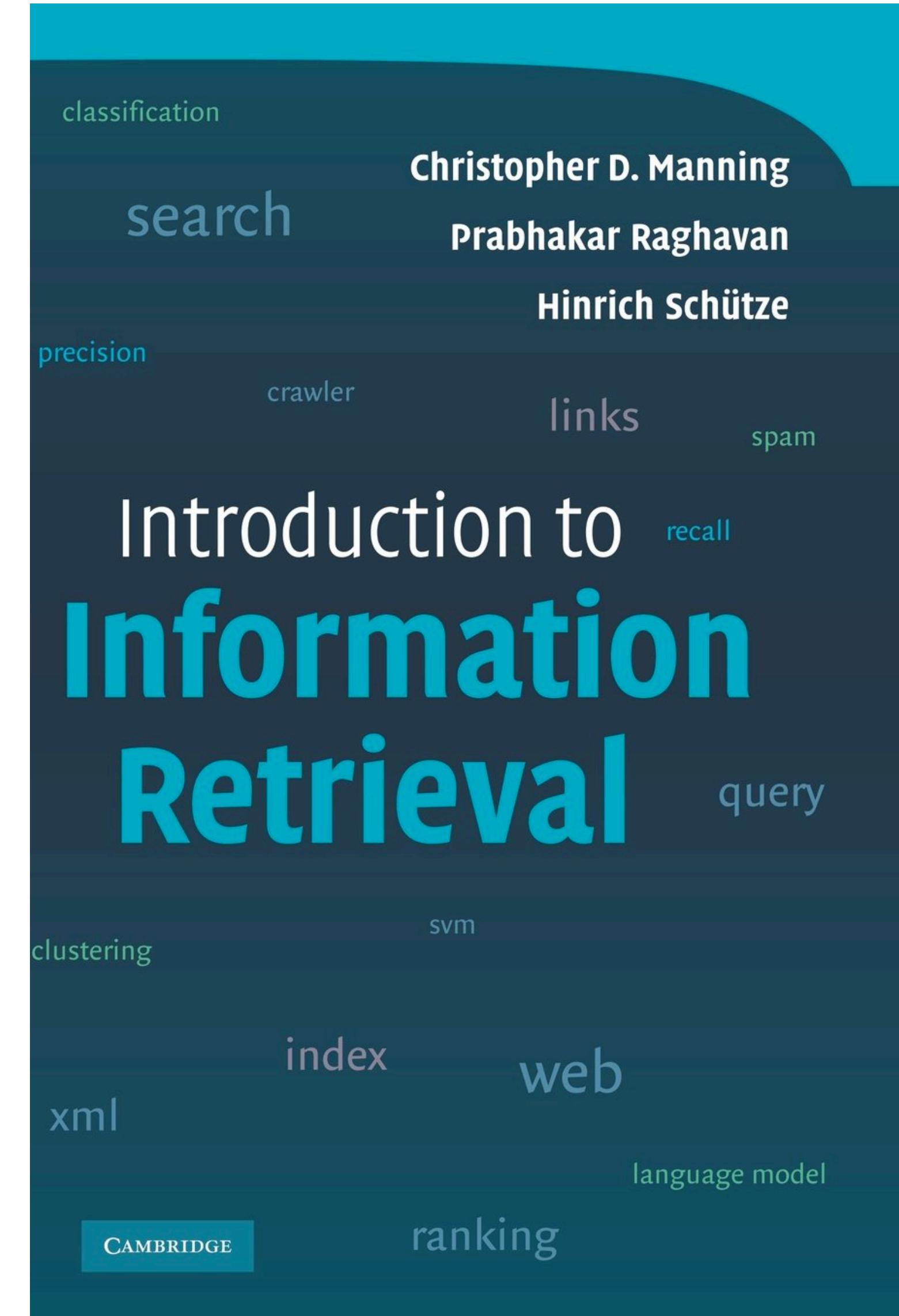


HARIDATUM

Information retrieval

Why?

- Huge intersection
- NLP is made from IR
- Good enough to start



Information retrieval formalism:

D - a set of documents (that is, texts + possibly some metadata)

Q - a set of queries, also a sequence of terms

We believe there is **Rel**: $Q \times D \rightarrow R$ – relevance score for document and query (a measure to what extent the user's information need (represented as a query) is satisfied by the document)

The goal is **to find best-matching documents** (in terms of Rel). On the fly

Information retrieval:

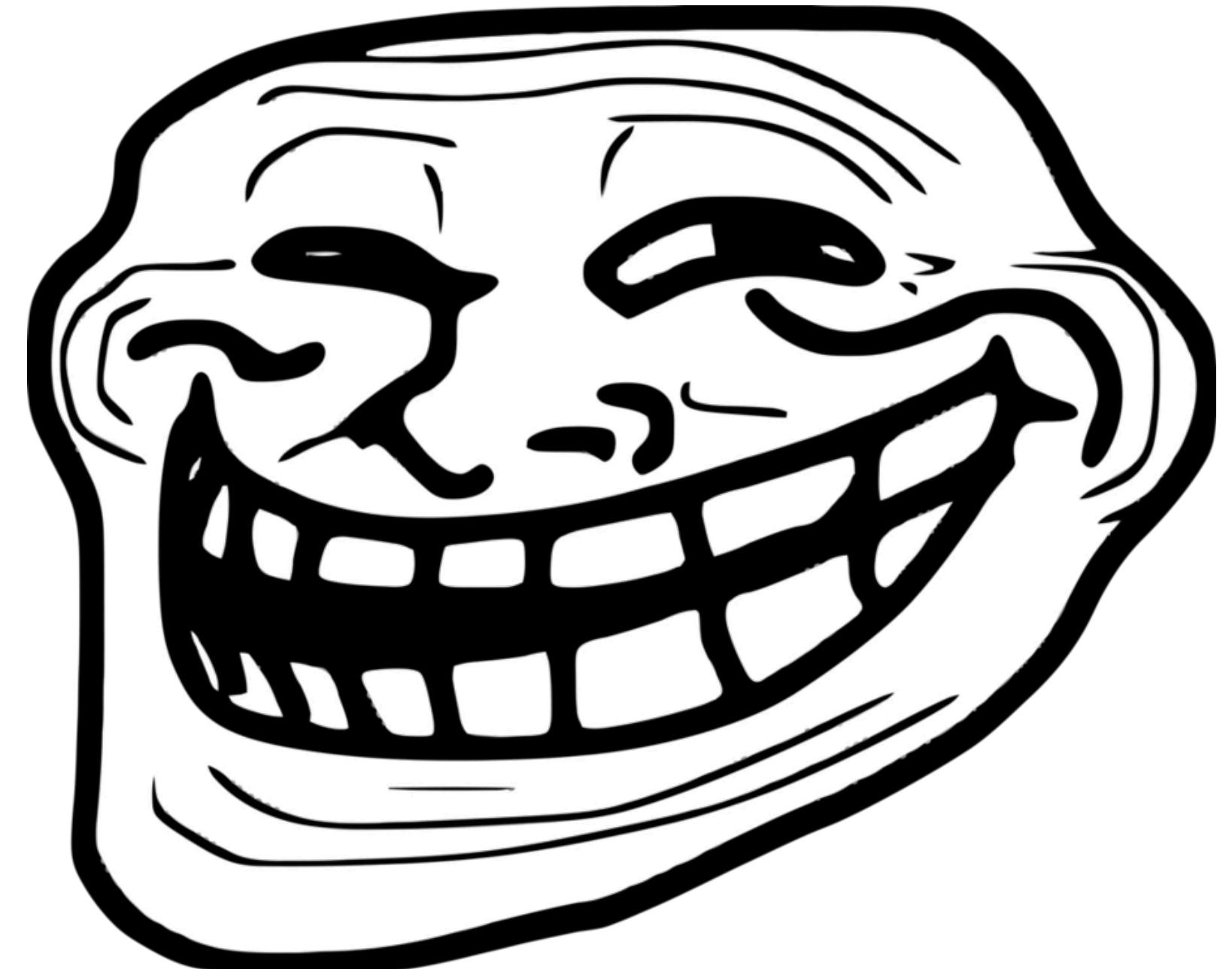
Plan:

1. Download whole internet
2. Invent Rel
3. Scan the database for every query
and compute Rel for every query-document pair
4. ???
5. PROFIT!

Information retrieval:

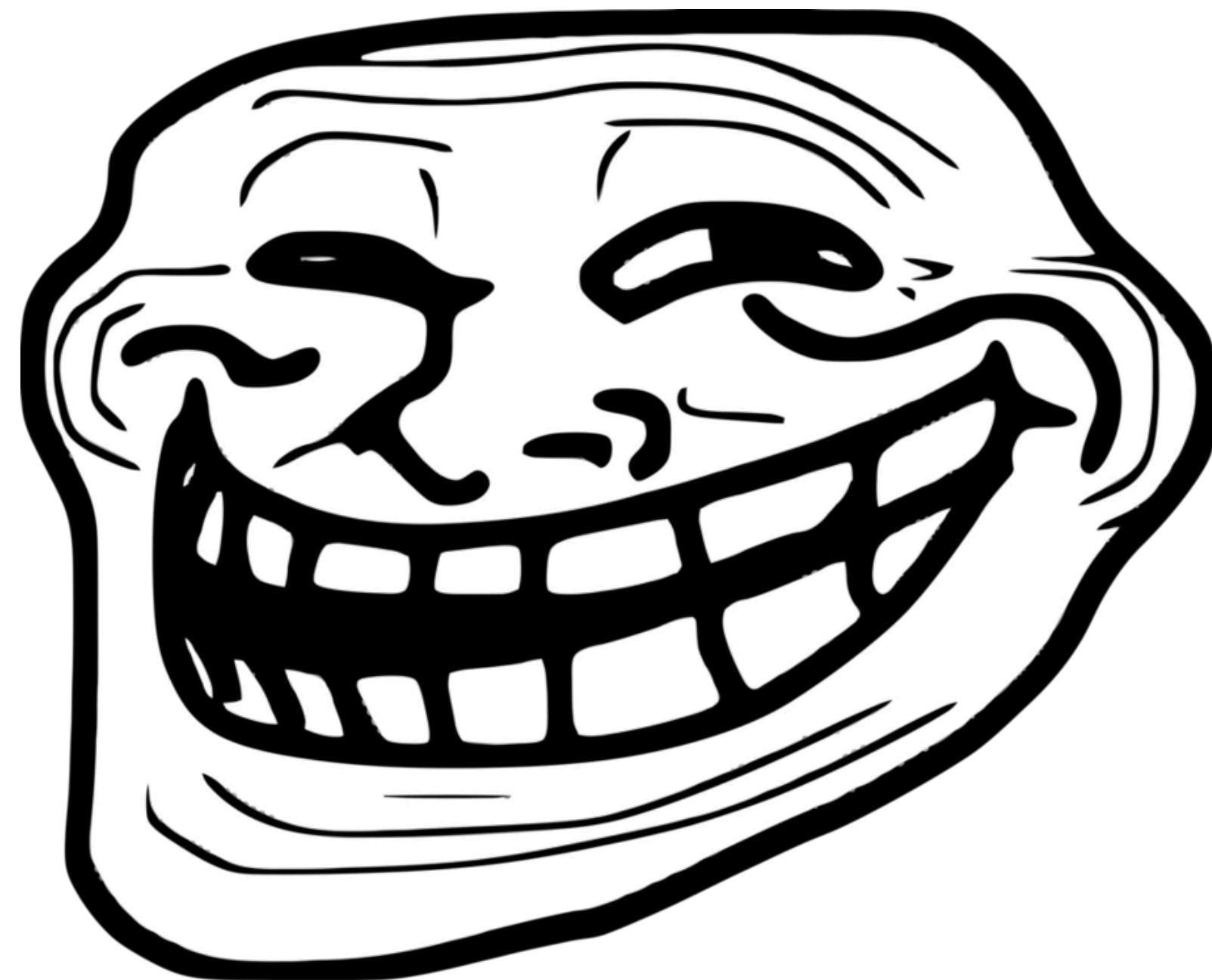
Plan:

1. Download whole internet
2. Invent Rel
3. Scan the database for every query
and compute Rel for every query-document pair
4. ????
5. PROFIT!
6. PROBLEMS?



PROBLEM?

Information retrieval:

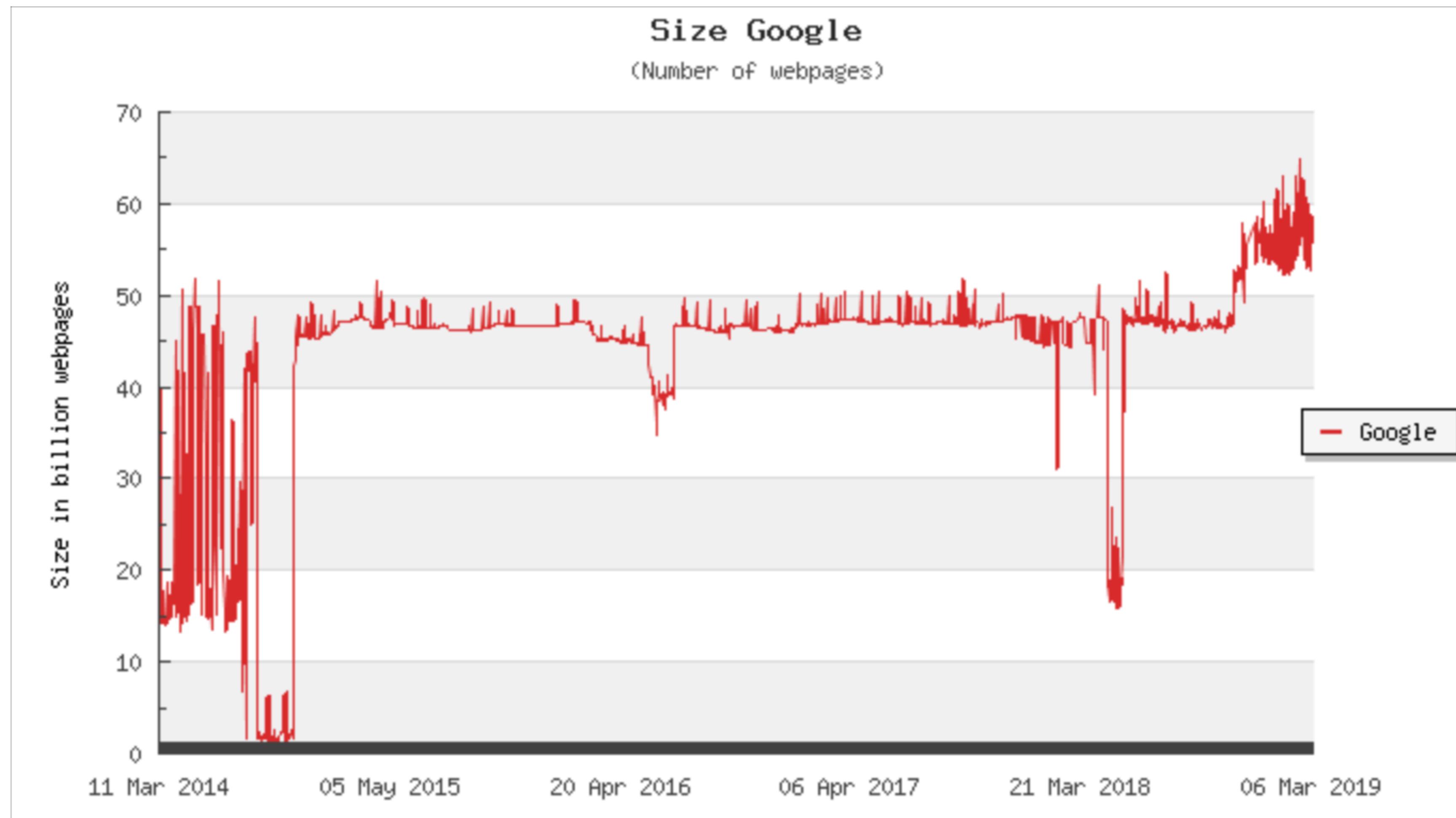


PROBLEM?

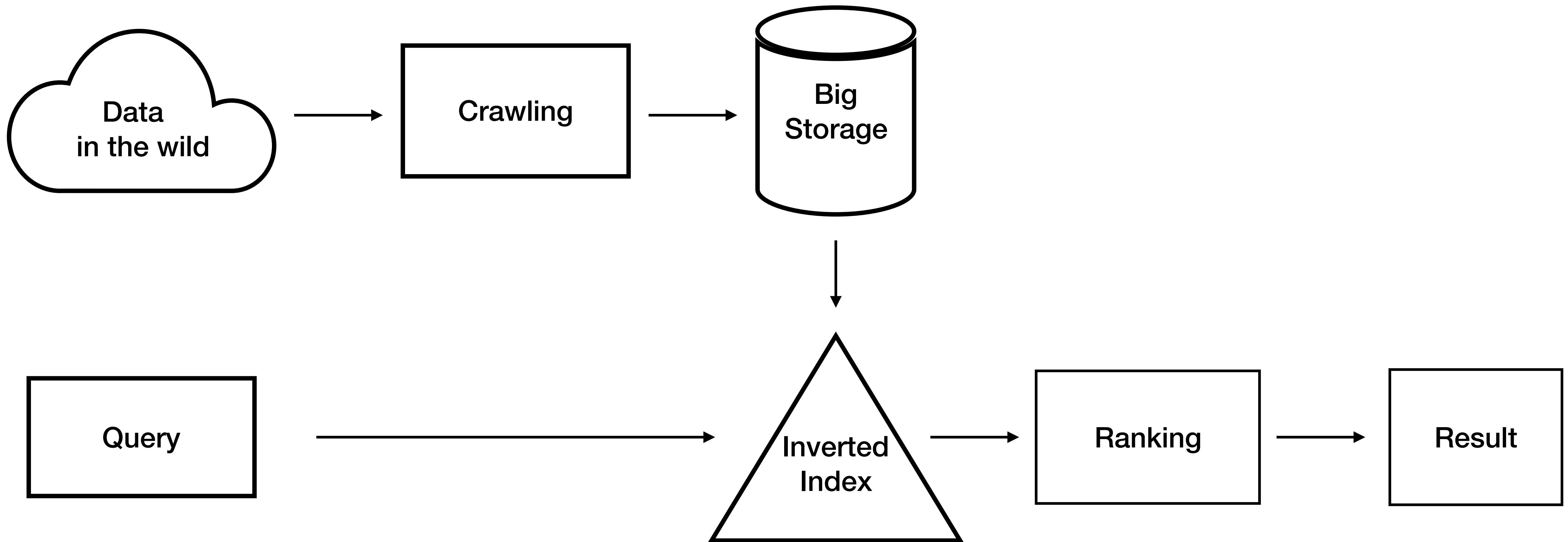
The goal is to find best-matching documents (in terms of Rel). **On the fly**

It's not so simple

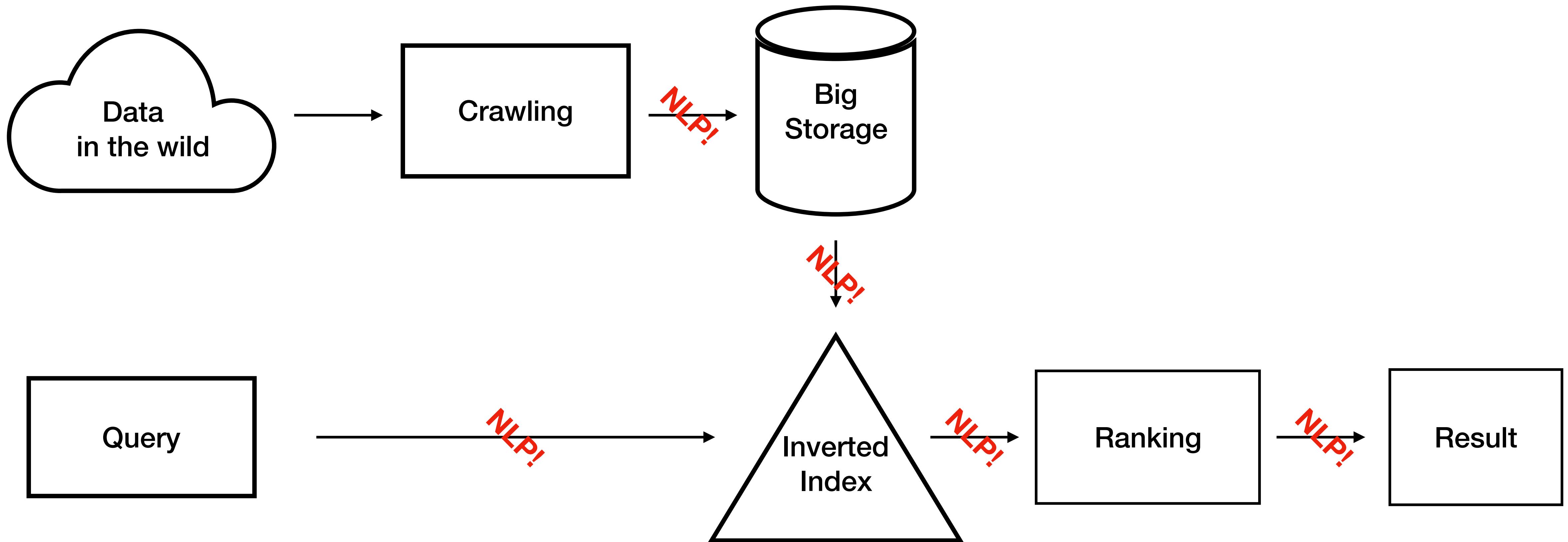
Google have to perform ~60M billion records per query!



General scheme:



General scheme:



By stages:

1. Collect data
2. Prepare documents
3. Process query
4. Retrieve by query
5. Rank documents
6. Prepare results
7. Personalize experience
8. Suggest query
9. Add wizard
10. ...

Data preparation:

Dirty text:

«На берегу пустынных волн»</p>

на берегу пустынных волн

<на берегу пустынных волн, RU>

на берег пустынныи волна

(или: на берег пустын волн)

Cleaning

Language detection

Lemmatization

Stemming

Data preparation: stemming

- Morphing words so that all possible forms of the word would turn into a single item, stem. Can be solved as a «language-independent» task.
- Usually when we say ‘stemming’ we mean cutting words (removing suffixes, prefixes, etc.) so that only the common part of all forms of the target word remains
- More widely spread and well-known is this one C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980. *New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587)*

The **rules** for removing a suffix will be given in the form

(condition) S1 -> S2



The ‘condition’ part may also contain the following:

- *S - the stem ends with S (and similarly for the other letters).
- *v* - the stem contains a vowel.
- *d - the stem ends with a double consonant (e.g. -TT, -SS).
- *o - the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

Step 1a

SSES -> SS	caresses -> caress
IES -> I	ponies -> poni

Data preparation: lemmatisation

- Preparing documents: lemmatisation
 - conversion to infinitive forms of words
- Usually: dictionary-based approach + morphological tricks!
- Russian: **mystem 3.0 / pymorphy2**
- Stemming: yandex myStem/nltk/spacy/ pattern (python) / apache lucene (Java)

By stages:

1. Collect data
2. Prepare documents
- 3. Process query**
4. Retrieve by query
5. Rank documents
6. Prepare results
7. Personalize experience
8. Suggest query
9. Add wizard

Indexing (for boolean search)

So one does not simply scan the web

The minimal entity in search are **terms** -- so let us first learn how to retrieve documents containing certain terms and term combinations, that is, “boolean retrieval”

Information need = “would love to read wiki page on manul or maine coon”

Boolean query = “wiki” AND (“maine coon” OR “manul”)



Indexing

«wiki» and («manul» OF «maine coon»)

Long bit vectors + bitmasking for search!

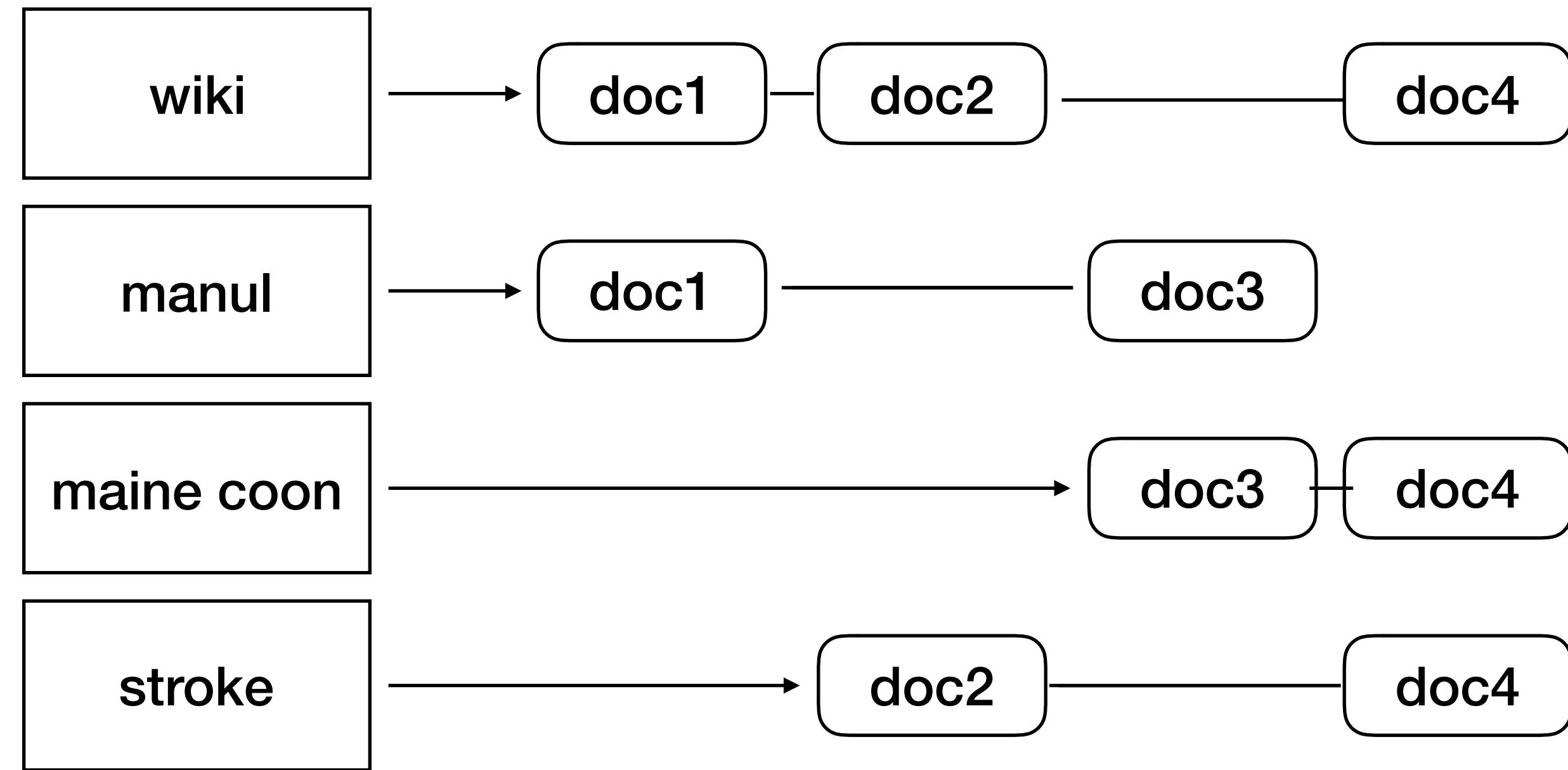
Is this an acceptable solution?

	/	Wiki	manul	Maine coon	Stroke
doc1		1	1	0	0
doc2		1	0	0	1
doc3		0	1	1	0
doc4		1	0	1	1

Indexing

- «wiki» and «manul» or «maine coon»

	/	Wiki	manul	Maine coon	Stroke
doc1	1	1	0	0	0
doc2	1	0	0	1	1
doc3	0	1	1	0	0
doc4	1	0	1	1	1



A dictionary, with linked lists of document meta information records (position in text, true word form, frequency of the term in concern in the document etc.) **ordered by document IDs**

Inverted index

- Inverted indices
 - **AND:** intersection of sorted lists
 - **OR:** union of sorted lists
- Multiple natural tricks for performance and quality boost:
 - store lists compressed and unpack on the fly when reading
 - add forward-links once in a while (skip-lists)
 - store positions of terms in the doc (allows to use distances between terms from query)

Query preprocessing:

Preparing the query the way we did with the document -- and we have the boolean retrieval system all set and ready.

Also a good idea in practice:

- add extra similar and relevant terms (query expansion)
- classification of query intention to understand which index to use (there may be many specific ones)

However, there are many more tricks, which are out of scope of this course

Ranking - is hard



Why is that hard?

jaguar

jaguar напиток

jaguar

jaguar xf

9 [Jaguar в Санкт-Петербурге - отзывы, фото, телефоны,...](#)

[maps.yandex.ru › jaguar](#)

Jaguar в Санкт-Петербурге - отзывы, фото, телефоны, адреса с рейтингом, отзывами и фотографиями. Адреса, телефоны, часы работы, схема проезда.

■ [Jaguar — смотрите картинки](#)

[yandex.ru/images › jaguar ›](#)



■ [Jaguar — подержанные и новые авто в Санкт-Петербурге](#)

[Запчасти](#) [Объявления](#) [Отзывы](#) [Каталог](#) [Дилеры](#)

[auto.ru › Jaguar](#)

Большая база объявлений о продаже автомобилей Jaguar. Полная информация об автомобилях — фотографии, отзывы, характеристики и цены.

Why is that hard?

One have to solve multiple problems simultaneously:

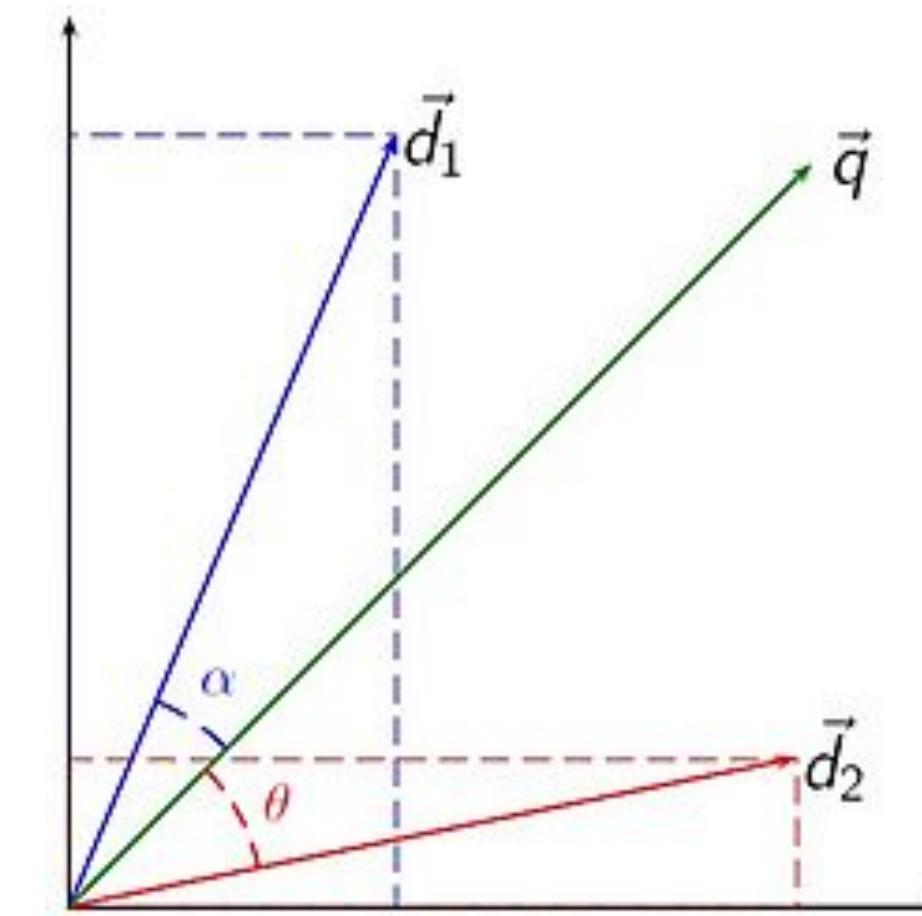
1. Matching document and query
2. Document quality
3. Matching user's interests and behavioral patterns*
4. Search results diversification (one of the “Jaguar case” solutions)
5. ...

Classics: Vector Space Model (VSM)

Every text and every query is represented as a vector of the same fixed number of dimensions, then documents vector representations are sorted by the distance/closeness to the query vector

$$d_j = (w_{j1}, w_{j2}, \dots, w_{jn})$$

$$q = (w_{1q}, w_{2q}, \dots, w_{nq})$$



$$\cos \theta = \frac{d \cdot q}{||d|| \cdot ||q||}$$

$$\text{sim}(d_j, q) = \frac{d_j \cdot q}{||d_j|| \cdot ||q||} = \frac{\sum_{i=1}^N w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^N w_{i,j}^2} \sqrt{\sum_{i=1}^N w_{i,q}^2}}$$

Ranking: the task

The goal is to sort search results so that the most relevant would be at the top of the list

We can treat it as machine learning task to learn relevance function: $\text{Rel}(D, Q) \rightarrow R$

Approaches:

- **Elementwise** approach:

We have relevance scores defined in train set; fitting to them

- **Pairwise** approach:

$\text{Rel}(d, q) < \text{Rel}(d', q)$, fitting the function given pairs

- **Listwise** approach:

Lists of documents $\{d^i\}$, sorted by relevance for q



Results preparation

Google search results for "ranking ml memes":

ranked ml memes

Все Картинки Новости Видео Покупки Ещё Настройки Инструменты

Результатов: примерно 5 520 000 (0,61 сек.)

Картинки по запросу ranking ml memes

→ Другие картинки по запросу "ranking ml memes"

Пожаловаться на картинки

Ranked Lists | Know Your Meme
<https://knowyourmeme.com/memes/ranked-lists> ▾ Перевести эту страницу

Ranked Lists is a phrasal template in which people use the format of a ranked list to express the pointlessness of the exercise. However, the number one slot ...
Не найдено: ml

Best Memes of 2018: Most Popular Memes of Last Year - Thrillist
<https://www.thrillist.com/entertainment/.../best-memes-2018> ▾ Перевести эту страницу

2018's memes had big shoes to fill -- it's tough to beat a year of blinking guy, Salt Bae, and disrespectful MLB Insider Dinger (@atf13atf) February 12, 2018 ...

The 15 best ranking memes, ranked - Mashable
<https://mashable.com/article/ranked-memes-ranking/> ▾ Перевести эту страницу

26 июн. 2018 г. - A ranked ranking of the best ranking memes. Rank rank rank rank rank.
Не найдено: ml

ranking ml memes

Web Images Video News More Anytime

The Phenomena Behind Popular Memes: How Ranking Algorithms ...
www.analyticsindiamag.com/the-phenomena-behind... ▾

The Phenomena Behind Popular Memes: How Ranking Algorithms Are Making Only Some Posts Richer. ... Tags algorithms machine learning popularity based ranking ranking.

The Week's Best Memes, Ranked - Digg
digg.com/2019/meme-ranking-really-choking-sasuke ▾

The Official Josco™ 2019 Meme Power Ranking As the weeks go by, we'll keep ranking memes, and the 2019 Power Ranking will take shape. For now, we have this week's memes.

Ranking MI Memes - Image Results



More Ranking MI Memes images

This Week's Best Memes, Ranked - Digg
digg.com/2019/meme-ranking-mulaney-shaggy-thotiana ▾

The Official Josco™ 2019 Meme Power Ranking As year goes on, we'll keep ranking memes, and the 2019 Power Ranking will take shape. For now, we have this week and last week's memes.

The ranking meme is the ultimate meme for superfans ...
www.someecards.com/life/lifestyle/ranking-memes ▾

The ranking meme is the ultimate meme for superfans. Orli Matlow. Jun 27, 2018 @ 4:39 PM. Advertising. The latest hot meme since yesterday's hottest meme (shout out ...

How to build your own IR engine:

- **Python:**
 - Scrapy, requests/urllib + BeautifulSoup
 - Pylucene, Xapian
- **Java:**
 - Crawling: Nutch, StormCrawler, ...
 - Retrieval: Lucene (Java library), ElasticSearch (incl. Lucene), Solr



Other tasks from IR:

- Query correction/augmentation
- Query expansion: morphology, semantics
- Search/suggestions personalization
- Dealing with queries semantic ambiguity
- Fact extraction
- Smart suggestions
- Queries classification
- Document clustering
- Duplicate detection
- Virus / spam detection
- Events detection

Remarks about IR + NLP

- It is mainly thanks to IR community's efforts that evaluation of data processing algorithms became a common practice (esp. in NLP)
- IR != NLP, but IR + NLP = <3
interconnections & common tricks
- Web search is probably the most successful case of applying NLP in production
(though the number grows)

Used/Recommended materials:

- Anton Alekseev's lectures at NRU ITMO, St Petersburg
- Introduction to Information Retrieval,
Chr. Manning, Pr. Raghavan and H. Schütze. Cambridge University Press. 2008
- ...