

Информационный поиск и
обработка естественного языка

Information Retrieval and
Natural Language Processing

Павел Браславский

ВВЕДЕНИЕ

Информационный поиск

- (Самое?) массовое приложение, успешная бизнес-модель
- Приложение-агрегатор (карты, картинки, звук, социальные сети, голос, ...)
- От минимального использования ОЕЯ до самых «продвинутых» методов
- Большие объемы текстовой информации
- Стандарты оценки (evaluation)

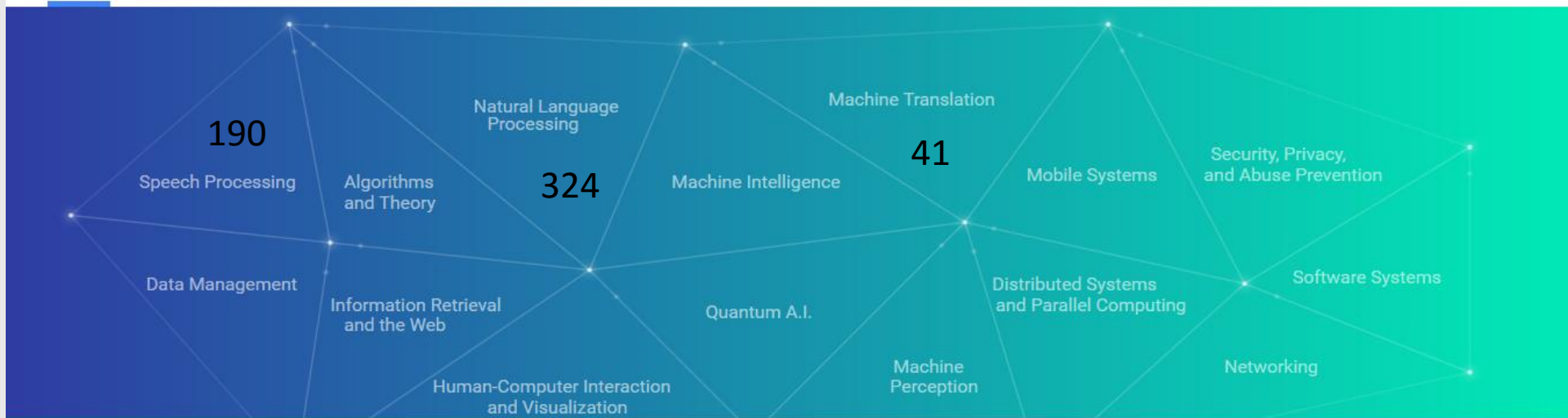
Темы SIGIR2017

- Queries and Query Analysis
- Web Search
- Mining and Modeling Search Activity
- Interactive Search
- Local and Mobile Search
- Retrieval Models and Ranking
- Social Search
- Filtering and Recommending
- Evaluation
- Document Representation and Content Analysis
- Question Answering
- Efficiency and Scalability
- Search in Structured Data
- Multimedia Search
- Other Applications and Specialized Domains

NLP@Google



[Home](#) [Publications](#) [People](#) [Teams](#) [Outreach](#) [Blog](#) [Work at Google](#)

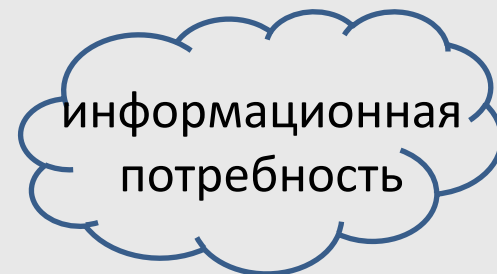
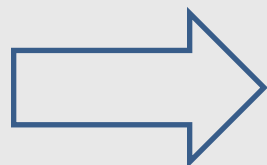


Определение

Поиск объектов (документов) в больших текстовых (=неструктурированных) коллекциях, которые удовлетворяют информационные потребности пользователей.

+представление, хранение, ... документов и коллекций

Акцент на *информации*, а не *данных*



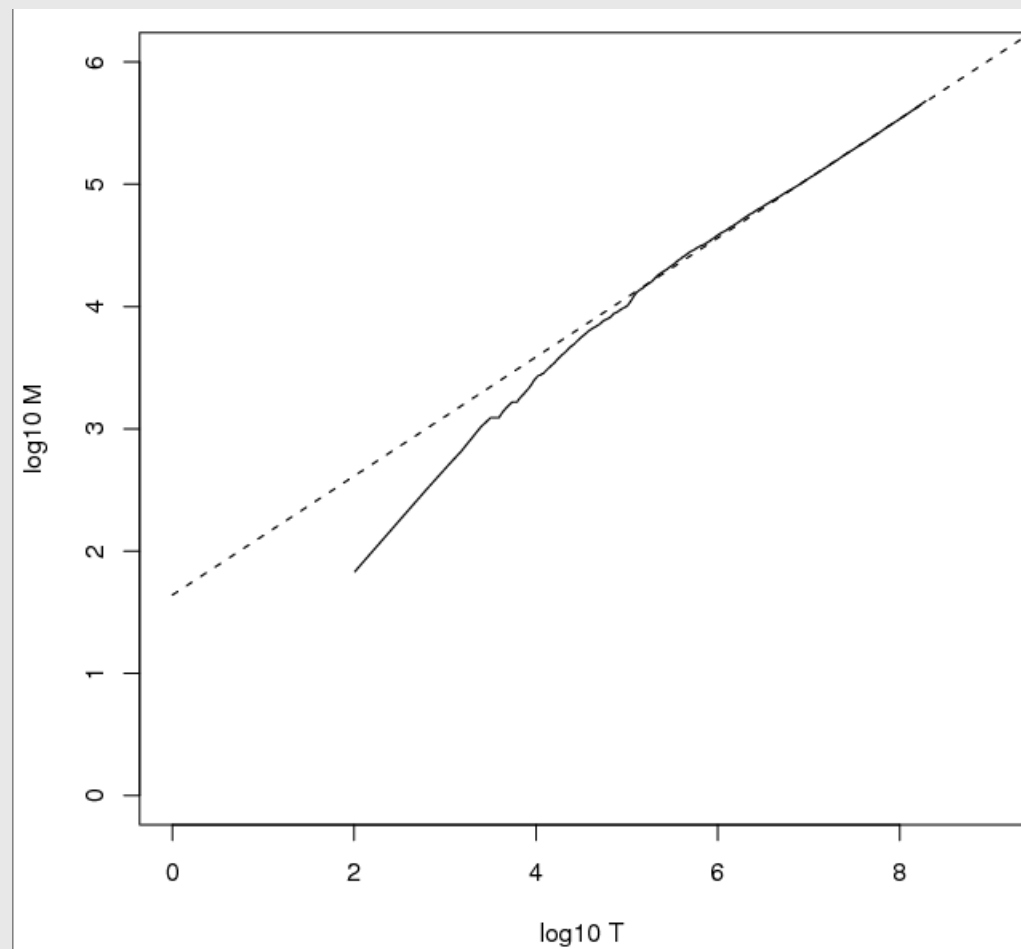
ЧАСТОТНЫЕ СВОЙСТВА СЛОВ

Рост словаря с ростом коллекции

- Закон Хипса: $M = kT^b$
- M – размер словаря (уникальные слова),
 T – словоупотребления в коллекции
- Типичные значения: $30 \leq k \leq 100$, $b \approx 0.5$
- Эмпирический закон

Закон Хипса – коллекция RCV1

- Аппроксимация методом наименьших квадратов:
- $\log_{10} M = 0.49 \log_{10} T + 1.64$
 $M = 10^{1.64} T^{0.49}$, т.е.
 $k = 10^{1.64} \approx 44$, $b = 0.49$.
- Для первых 1,000,020 слов коллекции модель предсказывает 38,323 уникальных слов;
фактически: 38,365



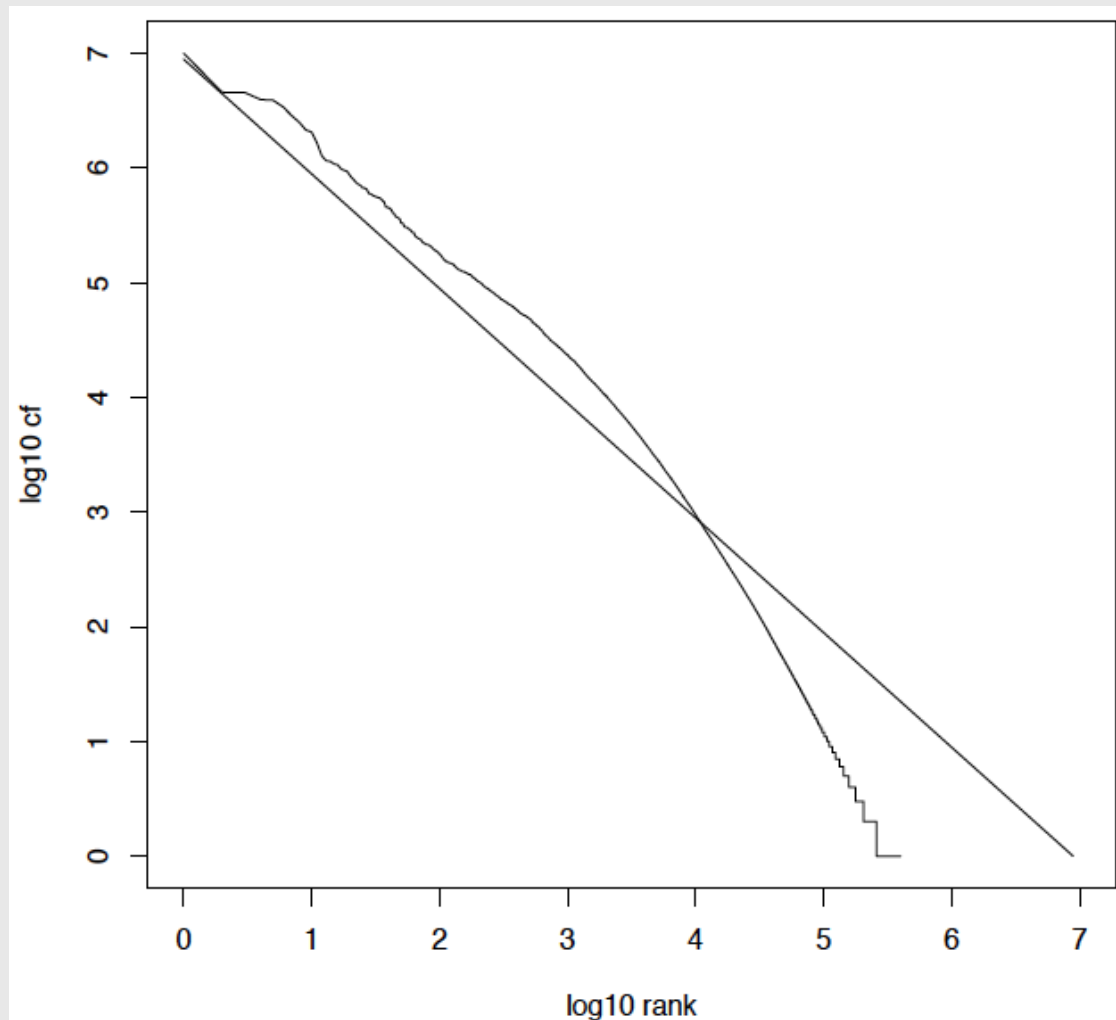
Закон Ципфа

- Частота i -го слова в частотном пропорциональна $1/i$ (первое слово – самое частотное)
- $cf_i \sim 1/i = K/i$, где K – нормализующая константа
- cf_i – частота в коллекции (*collection frequency*): сколько раз слово встретилось в коллекции.
- Эмпирический закон
- Если самое частое слово встречается cf_1 раз, то второе по частоте встречается $cf_1/2$ раз, третье $cf_1/3$ раз и т.д.

НКРЯ

№	Словоформа	Документы	Частота
1	<u>и</u>	57816	7416716
2	<u>в</u>	58555	5842670
3	<u>не</u>	49962	3385161
4	<u>на</u>	55776	2936096
5	<u>с</u>	53453	2228350
6	<u>что</u>	49428	2210373
7	<u>я</u>	24694	1592127
8	<u>а</u>	47492	1541398
9	<u>он</u>	34574	1377314
10	<u>как</u>	44897	1300577
11	<u>к</u>	46631	1132463
12	<u>по</u>	51068	1071698
13	<u>но</u>	41552	1048321
14	<u>его</u>	39492	983462
15	<u>это</u>	40341	957828
16	<u>из</u>	46975	836230
17	<u>все</u>	39105	817619
18	v	37148	798746

Закон Ципфа – коллекция RCV1



[Nayak & Raghavan]

ВЕКТОРНАЯ МОДЕЛЬ

Матрица «термин-документ»

Основная структура данных в ИП

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

[Nayak and Raghavan]

Частота термина в документе

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

«мешок слов» (bag of words) – не важен порядок,
взаимное расположение слов, только частота

[Nayak and Raghavan]

Сглаживание частоты

- Сглаженный вес термина t в документе d
$$w_{t,d} = \begin{cases} 1 + \log_{10} \text{tf}_{t,d}, & \text{if } \text{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$
- $0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 10 \rightarrow 2, 1000 \rightarrow 4$, etc.
- Соответствие запросу: сумма по терминам t , которые есть и в запросе q , и в документе d :

$$\sum_{t \in q \cap d} (1 + \log \text{tf}_{t,d})$$

[Nayak and Raghavan]

Документная частота

- df_t документная частота t : количество документов, содержащих t
 - df_t обратная мера «информативности» t
 - $df_t \leq N$
- idf (inverse document frequency):
$$idf_t = \log_{10} (N/df_t)$$
 - Логарифм – для сглаживания

[Nayak and Raghavan]

Весы tf-idf

- Вес tf-idf термина t в документе d : комбинация частоты по документу и частоты в коллекции

$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \times \log_{10}(N / \text{df}_t)$$

- Пожалуй, самая известная формула в информационном поиске
- Используется в классификации документов, выделении ключевых слов, ...
 - $\text{tf-idf} \leftarrow$ это не минус, а дефис (ВИП ☹);
альтернативные обозначения: tf.idf , $\text{tf} \times \text{idf}$
- Тем больше, чем чаще термин встречается в документе
- Тем больше, чем реже встречается в коллекции

[Nayak and Raghavan]

Представление tf.idf

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

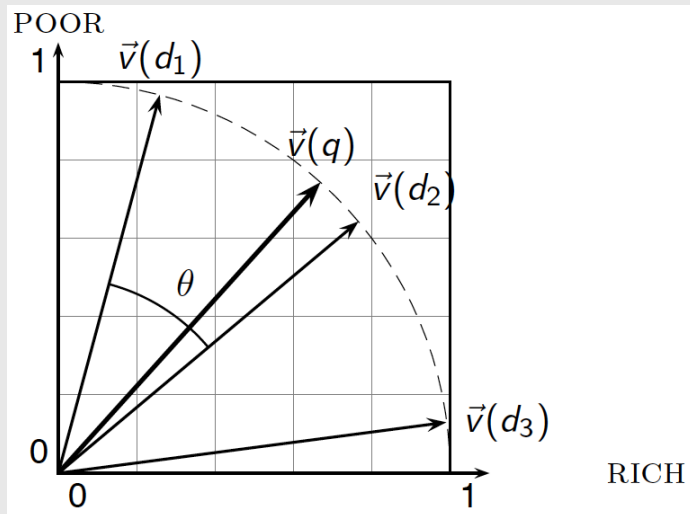
Документы – векторы в пространстве словаря, компоненты векторов – веса tf.idf

[Nayak and Raghavan]

Близость между векторами

- Документы, запросы – векторы
- Для векторов единичной длины

$$\cos(\vec{q}, \vec{d}) = \vec{q} \bullet \vec{d} = \sum_{i=1}^{|V|} q_i d_i$$



[Nayak and Raghavan]

Пример

Документ: *последняя точно последняя чашка*

термин	tf	log_tf	df	cf	idf	wt	norm.
последняя	2	1.3	300	700	0.82	1.07	0.41
точно	1	1.0	400	1 500	0.70	0.70	0.27
чашка	1	1.0	10	15	2.30	2.30	0.87

$$|d| = \text{sqrt}(1.07^2 + 0.7^2 + 2.3^2) = 2.63$$

ЯЗЫКОВЫЕ МОДЕЛИ В ИП

Идея

- Предположение: пользователи имеют представление о релевантном документе и формируют запрос из слов, которые могут встретиться в таком документе
- Каждый документ → униграммная языковая модель
- Ранжировать документы по убыванию вероятности того, что модель документа сгенерировала запрос

Реализация

$$p(Q, d) = p(d)p(Q | d) \approx p(d)p(Q | M_d)$$

$$\hat{p}(Q | M_d) = \prod_{t \in Q} \hat{p}_{ml}(t | M_d) = \prod_{t \in Q} \frac{tf_{(t,d)}}{dl_d}$$

Что делать с нулями?

Смешанная модель:

$$P(w|d) = \lambda P_{mle}(w|M_d) + (1 - \lambda)P_{mle}(w|M_c)$$

Подбор λ критичен для качества

Можно настраивать в зависимости от длины запроса

$$p(Q, d) = p(d) \prod_{t \in Q} ((1 - \lambda) p(t) + \lambda p(t | M_d))$$

Пример

Коллекция:

d1: красный синий зеленый желтый охра

d2: красный белый серый голубой лазоревый

q: красный синий

$$\lambda = 0.5$$

$$p(q | d1) = [(0.2+0.2)/2] * [(0.2+0.1)/2] = 0.03$$

$$p(q | d2) = [(0.2+0.2)/2] * [(0+0.1)/2] = 0.01$$

Результаты [Ponte & Croft, 1998]

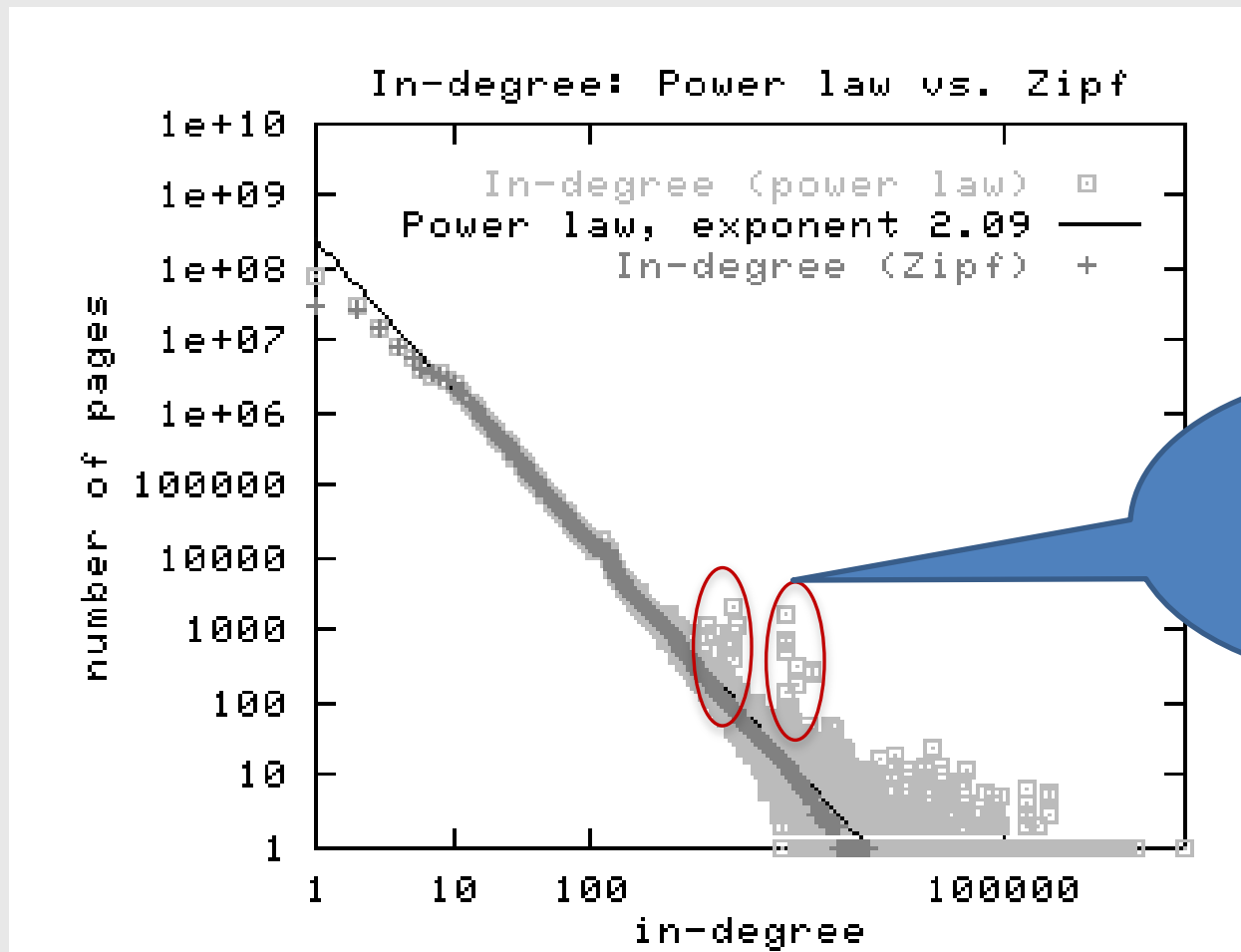
	tf.idf	LM	%chg	I/D	Sign	Wilc.
Rel:	6501	6501				
Rret.:	3201	3364	+5.09	36/43	0.0000★	0.0002★
Prec.						
0.00	0.7439	0.7590	+2.0	10/22	0.7383	0.5709
0.10	0.4521	0.4910	+8.6	24/42	0.2204	0.0761
0.20	0.3514	0.4045	+15.1	27/44	0.0871	0.0081★
0.30	0.2761	0.3342	+21.0	28/43	0.0330★	0.0054★
0.40	0.2093	0.2572	+22.9	25/39	0.0541	0.0158★
0.50	0.1558	0.2061	+32.3	24/35	0.0205★	0.0018★
0.60	0.1024	0.1405	+37.1	22/27	0.0008★	0.0027★
0.70	0.0451	0.0760	+68.7	13/15	0.0037★	0.0062★
0.80	0.0160	0.0432	+169.6	9/10	0.0107★	0.0035★
0.90	0.0033	0.0063	+89.3	2/3	0.5000	undef
1.00	0.0028	0.0050	+76.9	2/3	0.5000	undef
Avg:	0.1868	0.2233	+19.55	32/49	0.0222★	0.0003★
Prec.						
5	0.4939	0.5020	+1.7	10/21	0.6682	0.4106
10	0.4449	0.4898	+10.1	22/30	0.0081★	0.0154★
15	0.3932	0.4435	+12.8	19/26	0.0145★	0.0038★
20	0.3643	0.4051	+11.2	22/34	0.0607	0.0218★
30	0.3313	0.3707	+11.9	28/41	0.0138★	0.0070★
100	0.2157	0.2500	+15.9	32/42	0.0005★	0.0003★
200	0.1655	0.1903	+15.0	35/44	0.0001★	0.0000★
500	0.1004	0.1119	+11.4	36/44	0.0000★	0.0000★
1000	0.0653	0.0687	+5.1	36/43	0.0000★	0.0002★
RPr	0.2473	0.2876	+16.32	34/43	0.0001★	0.0000★

АНАЛИЗ ССЫЛОК

Анализ ссылок

- Веб-документ: текст + внешние ссылки
- Классификация по темам, сниппеты, близкие слова, ...
- Веб как граф:
 - Анализ авторитетности документа
 - Создание/выявление спама

Пример: входящие ссылки



спам:
нарушение
степенного
распределения

[Nayak and Raghavan]

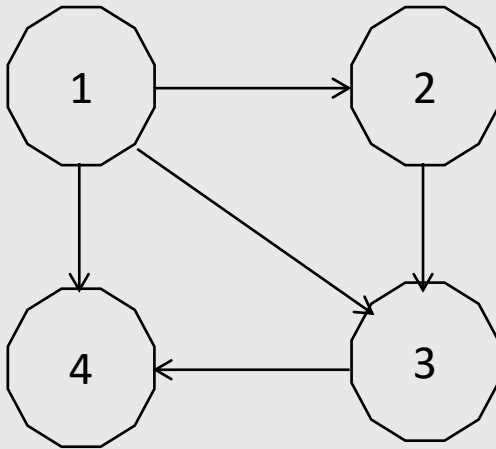
PageRank

- Авторитетность страницы на основе анализа графа ссылок
- ~Анализ цитирования научных публикаций
- Модель случайного блуждания по вебу:
 - с равной вероятностью переходим по любой исходящей ссылке
 - телепортация: с вероятностью 0.15 переходим на случайную страницу

Марковский случайный процесс

- Матрица переходов $P_{n \times n}$
- Эргодический случайный процесс: есть стационарное распределение вероятностей
- x – вектор вероятностей состояний
- $xP, xP^3, xP^2 \dots$ – последовательность состояний
- Для стационарного состояния $a = aP$
- a – главный собственный вектор P

Пример



0	1	1	1
0	0	1	0
0	0	0	1
0	0	0	0

0	0,33	0,33	0,33
0	0	1	0
0	0	0	1
0,25	0,25	0,25	0,25

0,038	0,321	0,321	0,321
0,038	0,038	0,888	0,038
0,038	0,038	0,038	0,888
0,250	0,250	0,250	0,250

Пример (продолжение)

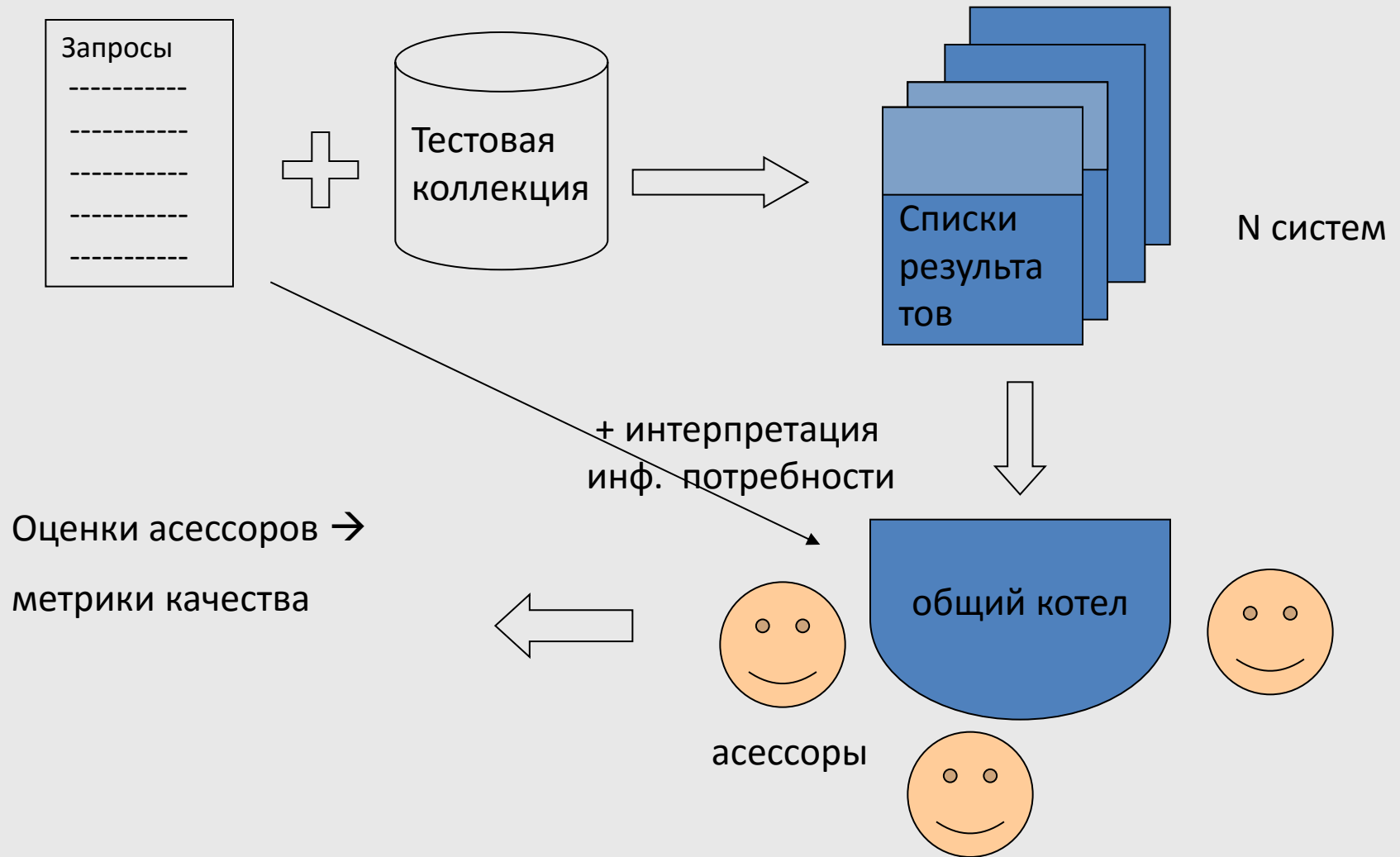
x_0	x_1	x_2	x_3	x_4	x_5
1	0,04	0,11	0,12	0,14	0,12
0	0,32	0,12	0,15	0,17	0,16
0	0,32	0,39	0,25	0,30	0,31
0	0,32	0,39	0,48	0,39	0,42

ОЦЕНКА

Оценка

- Релевантность (смысл слова, синтаксический разбор,...) – «в голове», мы не можем обойтись без человека при оценке
- Необходимо бороться с субъективностью и смещением
- Желательно делать результаты переиспользуемыми

Метод общего котла



Оценка качества поиска

Основа – понятие *релевантности* (соответствие информационной потребности)

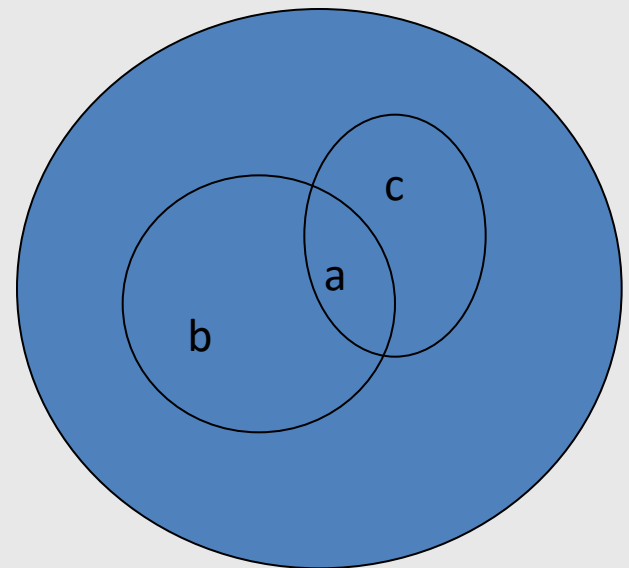
- Точность (precision)

$$p=a/b$$

- Полнота (recall)

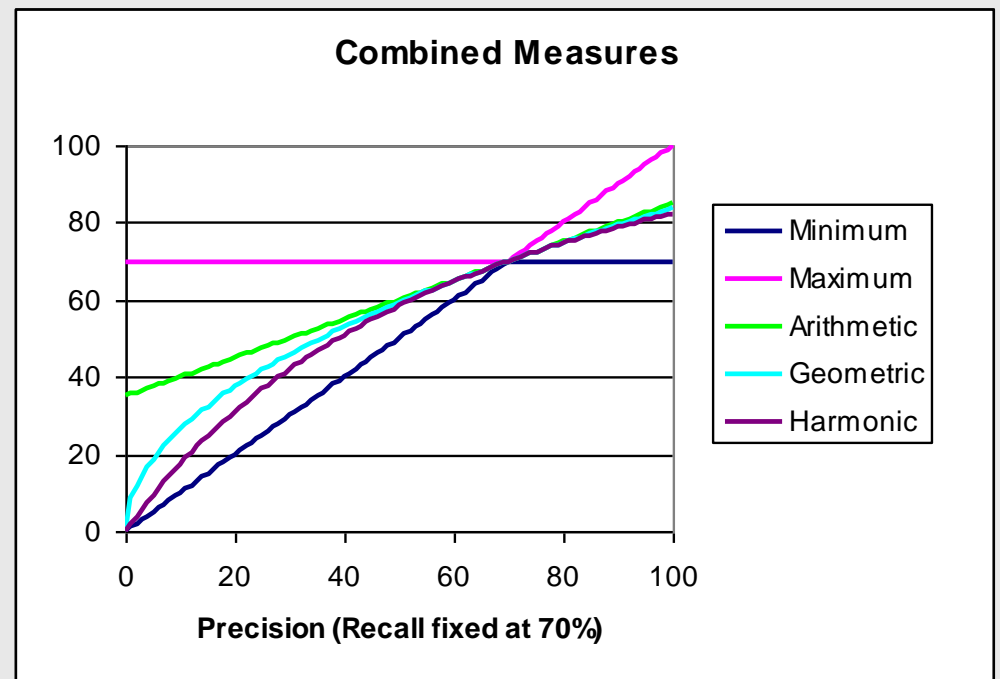
$$r=a/c$$

a – релевантные в отклике,
 b – всего в отклике,
 c – всего релевантных.



F-measure

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$



[Nayak and Raghavan]

Ранжированные результаты

- precision@n
- Mean Average Precision (MAP)
- Mean Reciprocal Rank (MRR)
- (normalized) Discounted Cumulative Gain (DCG/nDCG)

Пример

	Q1	Q2	Q3	Q4	Q5
1	0	1	1	1	0
2	1	0	1	0	0
3	1	1	0	0	1
4	0	0	1	0	1
5	1	1	0	0	1
Кол-во релевантных	4	3	5	2	5

$P = 0.52$

$R = 0.69$

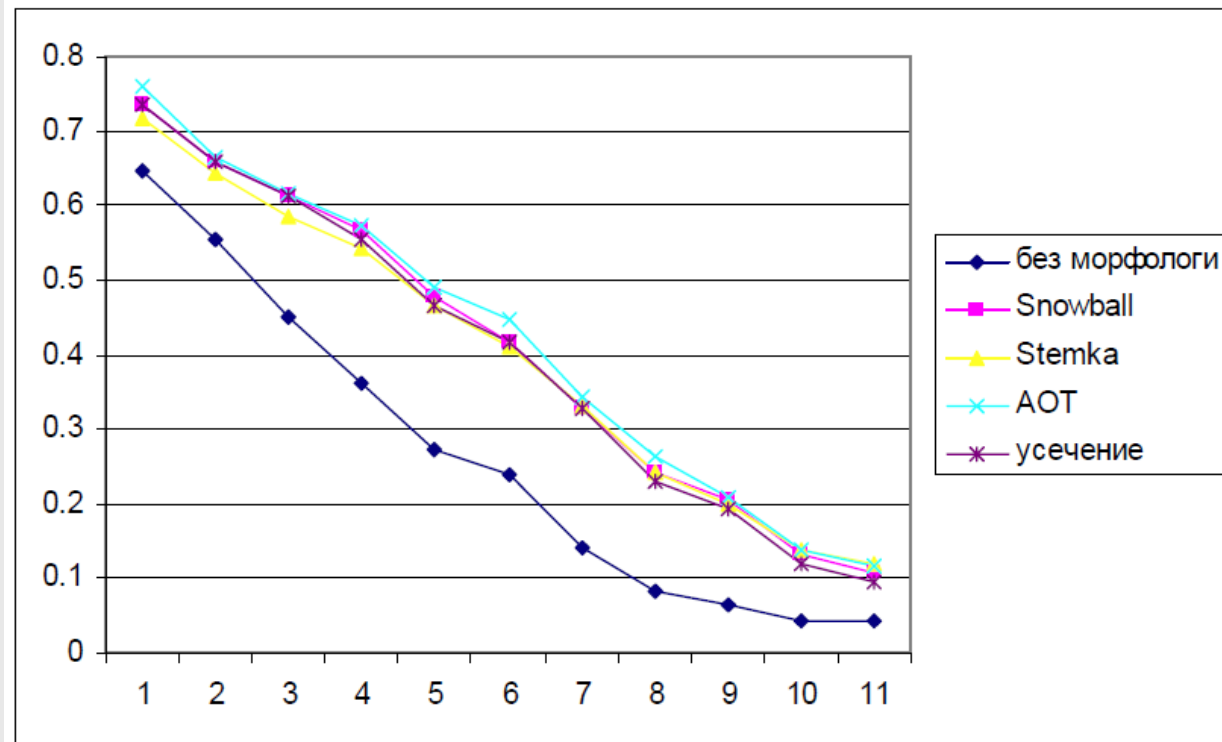
$p@3 = 0.53$

$MAP = 0.75$

$MRR = 0.77$

Морфология в поиске

Stemming helped markedly for Finnish (30% improvement) and Spanish (10% improvement), but for most languages, including English, the gain from stemming was in the range 0–5%, and results from a lemmatizer were poorer still. [IIR]



Согласие ассессоров

Каппа-статистика Коэна:

$$\kappa = [P(A) - P(E)] / [1 - P(E)]$$

$P(A)$ – доля совпадений

$P(E)$ – доля ожидаемых случайных совпадений

+вариант с весами

попарное согласие для 3 и более ассессоров

Пример

Number of docs	Judge 1	Judge 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

$$P(A) = 370/400 = 0.925$$

$$P(\text{nonrelevant}) = (10+20+70+70)/800 = 0.2125$$

$$P(\text{relevant}) = (10+20+300+300)/800 = 0.7878$$

$$P(E) = 0.2125^2 + 0.7878^2 = 0.665$$

$$\text{Kappa} = (0.925 - 0.665)/(1-0.665) = 0.776$$

Инициативы по оценке

- TREC <http://trec.nist.gov/>
- NTCIR <http://research.nii.ac.jp/ntcir/>
- CLEF <http://www.clef-initiative.eu/>
- SemEval <http://alt.qcri.org/semEval2016/>
- TAC <https://tac.nist.gov/>
- WMT <http://www.statmt.org/wmt17/>
- РОМИП <http://romip.ru/>
- Dialog <http://www.dialog-21.ru/evaluation/>

Коллекции РОМИП

Коллекция	Документы	Размер (compressed)	Запросы	Оценено
Legal	~300,000	2 Gb	14,794	220
By.Web	1,524,676	8 Gb	~ 60,000	1 500+
KM.RU	3,010,455	13 Gb	~ 60,000	~250

Дорожки РОМИП

- Поиск по запросу (ad-hoc text retrieval)
- Классификация документов
- Генерация сниппетов
- Вопросно-ответный поиск и извлечение фактов
- Кластеризация новостей
- Поиск по документу-образцу
- Анализ тональности
- Машинный перевод

Dialog Evaluation

2016

Анализ тональности

Выделение сущностей

Исправление опечаток

2015

Анализ тональности

Семантическая близость

2014

Разрешение анафоры

2013

Анализ тональности

Машинный перевод

2012

Анализ тональности

Синтаксис

2010

Морфология

Современные тенденции

- Краудсорсинг (crowdsourcing)
- Онлайн-оценка

ЕЩЕ БОЛЬШЕ ОЕЯ В ЗАДАЧАХ ИП

Синонимы

The screenshot shows a Google search interface with the query 'apartment rent london' in the search bar. The results page displays four search results, each with a title, URL, and a brief description. The word 'rent' is underlined in the search bar and in the descriptions of the first three results. The word 'flats' is underlined in the description of the third result. The word 'rent' is underlined in the description of the fourth result.

Google apartment rent london 🔍

All Maps News Shopping Images More ▾ Search tools

About 74,200,000 results (0.50 seconds)

Property to rent in London - Houses & Flats to rent in London
www.rightmove.co.uk/property-to-rent/London.html ▾ Rightmove ▾
Find property to rent in London. Search over 250000 properties to rent from the top lettings agents in the UK - Rightmove.
[Flats to rent in London](#) — [Property to rent in East London](#)

London property - Flats and houses for sale or to rent in ...
www.rightmove.co.uk/property/London.html ▾ Rightmove ▾
Find property in **London**. We have a wide range of **London** houses and flats for sale or to rent from top UK estate agents - Rightmove.

Flats & Houses to Rent | Property to Rent in London - Gumtree
<https://www.gumtree.com/flats-and-houses-for-rent/london> ▾ Gumtree ▾
Find the latest property to rent in London on Gumtree. See classified ads for flats, houses, office space, parking, storage, offers and more for rent in London.

Flats to rent in London - Zoopla
www.zoopla.co.uk/to-rent/flats/london/ ▾ Zoopla ▾
Find flats to rent in London with Zoopla. See apartments to let in London on a map.
[1 bedroom flats to rent in London](#) — [2 bedroom flats to rent in London](#)

Машинный перевод



<http://www.glanzundelend.de/konstanteseiten/Goethe/faust-prolog.htm>




Translate

From: German

To: Russian

критика

Классический Архив



Фауст - Обзор

кулак

Трагедия первая часть

Посвящение.

Вы снова к вам, колеблющиеся формы,
На ранней стадии после того , как тусклый вид показано на рисунке.
Я стараюсь хорошо, вы держите в этот раз?
Я чувствую мое сердце все же , что заблуждением склонен?
Ваши порывы к вам! ну, так что вы можете осуществлять,
Как выйти из тумана и тумана вокруг меня;
Моя грудь чувствует себя непоколебимое подростков
Волшебное прикосновение, окутывавшее ваш поезд.

Вы принесли с собой изображения более счастливые дни,
А некоторые любят тень на повышение;
Как старый, halbverklungenen прогноз
Приходите первую любовь и дружбу с планом;
Боль является новым, он повторил действие
лабиринтообразная запустить ERR Жизнь,
И называется хорошим, чтобы хорошие времена
Обманутые судьбы, подальше исчез передо мной.




HD 2400×1350



Ford Mustang.
k-punkt.com



Ответы



[All](#) [News](#) [Shopping](#) [Books](#) [Images](#) [More ▾](#) [Search tools](#)

About 66,100,000 results (0.58 seconds)

The iPhone 6 measures 5.44 x 2.64 x 0.27 inches (138.1 x 67 x 6.9mm) and weighs **4.55 ounces (129g)** – a weight increase that is roughly proportional to its 16% volume increase compared to the iPhone 5S. Sep 9, 2014

[iPhone 6 vs iPhone 6 Plus: The Differences Between The ...](#)
[www.forbes.com/.../iphone-6-vs-iphone-6-plus-what-is-the-differenc...](#) Forbes ▾

Feedback

[iPhone 6 - Technical Specifications - Apple](#)
[www.apple.com](#) › [iPhone](#) › [iPhone 6 ▾](#) Apple ▾

Weight: **4.55 ounces (129 grams)** 6.22 inches (158.1 mm) 3.06 inches (77.8 mm) 0.28 inch.

Поиск сущностей

дочь первого космонавта

✕ ⚙

Найти

Логин

поиск КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ

Интерес к семье первого космонавта (жене Валентине...)
vlasti.net > news/84780 ▾
Воспоминаниями о том, как сложилась судьба семьи первого космонавта, с ...
Старшая дочь Гагариных Елена так вспоминала жизнь в Звездном городке: «Там...

Галина Юрьевна Гагарина, младшая дочь первого...
fishki.net > 1562133-jurii-i-pervogo-kosmonavta.html =

папа леонардо ди каприо

🔍

All Videos Images News Maps More ▾ Search tools

About 283,000 results (0.40 seconds)

Leonardo DiCaprio / Father

George DiCaprio



George Paul DiCaprio is an American writer, editor, publisher, and distributor, known for his work in the realm of underground comix, where he collaborated with such notables as Timothy Leary and Laurie Anderson. [Wikipedia](#)

More about George DiCaprio

Feedback



Елена Юрьевна Гагарина

Генеральный директор Государственного историко-культурного музея-заповедника «Московский Кремль», искусствовед. Старшая дочь первого космонавта планеты Юрия Гагарина. [Википедия](#)

Родилась: 17 апреля 1959 г. (57 лет), Заполярный, Мурманская область, РСФСР, СССР

Родители: Валентина Ивановна Гагарина, Юрий Гагарин

Дети: Екатерина Караваева

56