

Модели информационного поиска на основе вероятностей

Средняя точность AP

- Было сказано, что это комбинированная мера, учитывает точность и полноту.
- Вопрос: где скрывается полнота

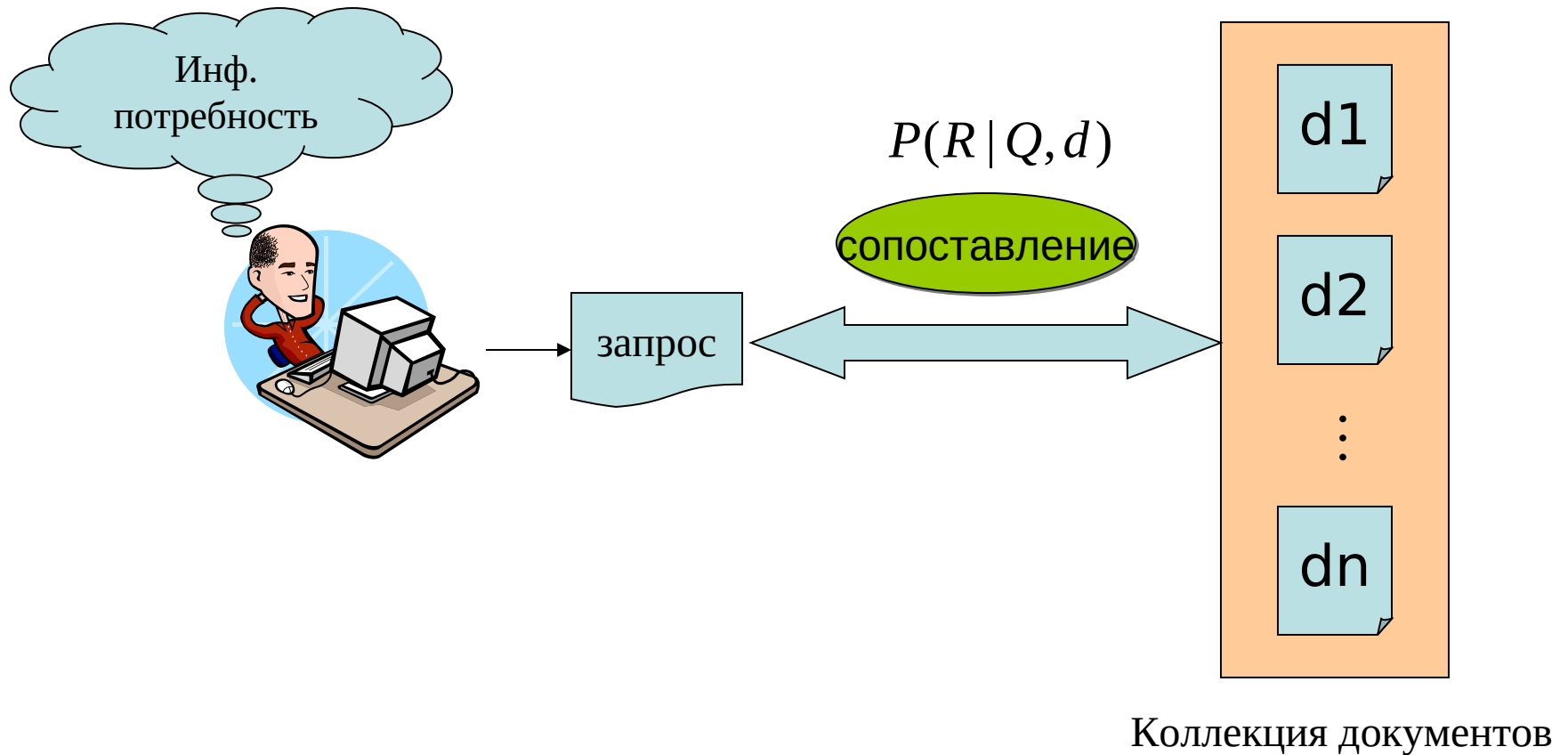
$$AP = \frac{\frac{1}{1} + \frac{2}{3} + \frac{3}{4} + \frac{4}{6} + \frac{5}{10} + \dots}{10}$$

Модели на основе вероятностей

- Векторная модель – выглядит как эвристика
- Модели на основе вероятностей
 - Попытка найти математическое обоснование
 - Вероятностная модель инф. поиска
 - Классическая модель BIM (Binary Independence Model)
 - Модель BM25
 - Языковая модель информационного поиска

Maron & Kuhns, 1960: т.к. поисковая система не может предсказать с уверенностью релевантность документа, то мы должны иметь дело с вероятностью

Вероятностный информационный поиск



Вероятностный инф. поиск: основная идея

Query	Doc	Rel
q	d	R
q1	d1	1
q1	d2	1
q1	d3	0
q1	d4	0
q1	d5	1
...		
q1	d1	0
q1	d2	1
q1	d2	0

$$f(q,d)=\underline{p(R=1 \mid d,q)}=\frac{\text{count}(q,d,R=1)}{\text{count}(q,d)}$$

Вероятностный инф. поиск: основная идея

Probabilistic Retrieval Models: Basic Idea

Query	Doc	Rel
q	d	R
q1	d1	1
q1	d2	1
q1	d3	0
q1	d4	0
q1	d5	1
...		
q1	d1	0
q1	d2	1
q1	d3	0

$$f(q,d)=p(R=1 | d,q)=\frac{\text{count}(q,d,R=1)}{\text{count}(q,d)}$$

$$P(R=1 | q1,d1) = 1/2$$

$$P(R=1 | q1,d2) = 2/2$$

$$P(R=1 | q1,d3) = 0/2$$

Вероятностная модель

- Предложена Робертсоном и Спарк Джонс в 1976
 - Binary independence retrieval model
 - Вероятностная модель пытается оценить вероятность, что пользователь оценит документ d_j как релевантный посредством отношения
 - $P(d_j \text{ relevant to } q)/P(d_j \text{ nonrelevant to } q)$

Вероятностная модель-2

- Определения

- Все веса слов - бинарные, т.е. $w_{i,j} \in \{0,1\}$
- Пусть R – множество документов, про которые известно, что они релевантны запросу q
- Пусть \bar{R} - оставшиеся документы
- $P(R | d_j)$ - это вероятность, что документ d_j релевантен запросу q
- $P(\bar{R} | d_j)$ - вероятность, что документ d_j нерелевантен запросу q

Вероятностная модель-3

- Сходство $\vec{sim}(d_j, q)$ документа d_j с запросом q определяется как отношение

$$\vec{sim}(d_j, q) = \frac{P(R | d_j)}{P(\bar{R} | d_j)}$$

- Используя правило Байеса

$$P(\mathcal{A} | \mathcal{B}) = \frac{P(\mathcal{B} | \mathcal{A})P(\mathcal{A})}{P(\mathcal{B})}$$

$$\vec{sim}(d_j, q) = \frac{P(d_j | R) \times P(R)}{P(d_j | \bar{R}) \times P(\bar{R})}$$

- $P(R)$ – вероятность случайно выбрать релевантный документ в коллекции

$$P(d_j | R)$$

- $P(d_j | R)$ – вероятность случайно выбрать документ d_j из множества R релевантных документов

Вероятностная модель-4

$$\text{sim}(\vec{d}_j, q) \approx \log \frac{P(\vec{d}_j | R)}{P(\vec{d}_j | \bar{R})} + \log \frac{P(R)}{P(\bar{R})}$$

- Предполагая независимость слов и имея запрос $q = (q_1, q_2, \dots, q_t)$,

$$P(\vec{d}_j | R) = \prod_{i=1}^t P(k_i = q_i | R)$$

$$P(\vec{d}_j | \bar{R}) = \prod_{i=1}^t P(k_i = q_i | \bar{R})$$

$$\text{sim}(\vec{d}_j, q) \approx \log \frac{\prod_{i=1}^t P(k_i = q_i | R)}{\prod_{i=1}^t P(k_i = q_i | \bar{R})}$$

Слово из запроса может присутствовать или отсутствовать в документе. $P()$ – вероятность присутствия или отсутствия слова запроса в документах

Вероятностная модель-5

- $P(k_i | R)$ – это вероятность, что слово k_i присутствует в документе из случайно выбранного документа в множестве релевантных документов R
- $P(\bar{k}_i | R)$ – это вероятность того, что k_i не присутствует в документе, случайно выбранного из множества релевантных документов R

Вероятностная модель - 6

$$sim(\vec{d}_j, q) \approx \frac{\prod_{g_i(q_j)=1} P(k_i | R) \prod_{g_i(q_j)=0} P(\bar{k}_i | R)}{\prod_{g_i(q_j)=1} P(k_i | \bar{R}) \prod_{g_i(q_j)=0} P(\bar{k}_i | \bar{R})}$$

$$\because P(k_i | R) + P(\bar{k}_i | R) = 1$$

Вторую группу сомножителей (для слов, которых нет в документе) домножаем на такие же сомножители, которые есть в документе, и для сохранения результата делим на такие же сомножители, относя их к первой группе.

Вторая группа сомножителей может быть убрана, поскольку теперь не зависит от конкретного документа. Получаем из первой группы:

$$sim(\vec{d}_j, q) \approx RSV_d = \sum_{i=1}^t \left(\log \frac{P(k_i | R)}{1 - P(k_i | R)} + \log \frac{1 - P(k_i | \bar{R})}{P(k_i | \bar{R})} \right)$$

Оценки вероятности на практике

- Полные множества релевантных и нерелевантных документов неизвестны
 - Нужно оценивать
- Для нерелевантных документов (второе слагаемое)
- Если нерелевантные документы аппроксимировать целой коллекцией, то
- r_i (вероятность встречаемости слова в нерелевантных документах для запроса) = df/N
 - $\log (1 - r_i)/r_i = \log (N - df)/df \approx \log N/df = \text{IDF!}$

Оценки вероятности на практике

Статистика релевантных документов может быть оценена различным образом:

- ① Можно использовать статистику слов в известных релевантных документах - это основа для вероятностных подходов к relevance feedback
- ② Установить как константу. Предположим, что вероятность нахождения слова запроса в релевантном документе $P(k_i | R) = 0.5$
 - Тогда первое выражение сокращается
 - Слабая оценка, но не противоречит предположениям
 - Получается, что ранжирование документов получено просто суммированием весов idf
 - Для коротких документов (заголовков или абстрактов) работает неплохо

Вероятностные модели

- Одна из старых формальных моделей информационного поиска
- Предположения в модели BIR:
 - Булевское представление документов, запросов и релевантности
 - Независимость слов
 - Слова, не входящие в запрос, не влияют на поиск
 - Релевантность документов не зависит друг от друга

Различие между векторными моделями и вероятностными не очень велико

- В обоих случаях поисковая система строится похожим образом
- Различия: в вероятностном информационном поиске сходство между запросов и документом считается не косинусной мерой и tf-idf в векторном пространстве, а несколько другой формулой, мотивированной теорией вероятности

Окарі BM25: Небинарная модель

- Вероятностная модель ВІМ была изначально создана для поиска по записям в коротких каталогах сопоставимой длины – и работала прилично в этих условиях
- Для современного полнотекстового поиска, модель должна учитывать частоту термина в документе и длину документа
- BestMatch25 (BM25 или Okapi), развитие модели ВІМ, учитывает эти величины
- С 1994 до наших дней, модель BM25 – это одна из наиболее распространенных и устойчивых моделей информационного поиска

Окарі ВМ25: Небинарная модель

- Простейшая форма веса для документа d – это просто суммирование idf слов запроса, которые присутствуют в этом документе.
- Это формула «исправляется» учетом частоты слова в документе и длины документа:

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{df_t} \right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b \times (L_d/L_{ave})) + tf_{td}}$$

- tf_{td} : частота слова в документе d
- L_d (L_{ave}): длина документа d (средняя длина документа в коллекции)
- k_1 : параметр, контролирующий учет частоты слова
- b : параметр, контролирующий учет длины документа

Okapi BM25: A Nonbinary Model

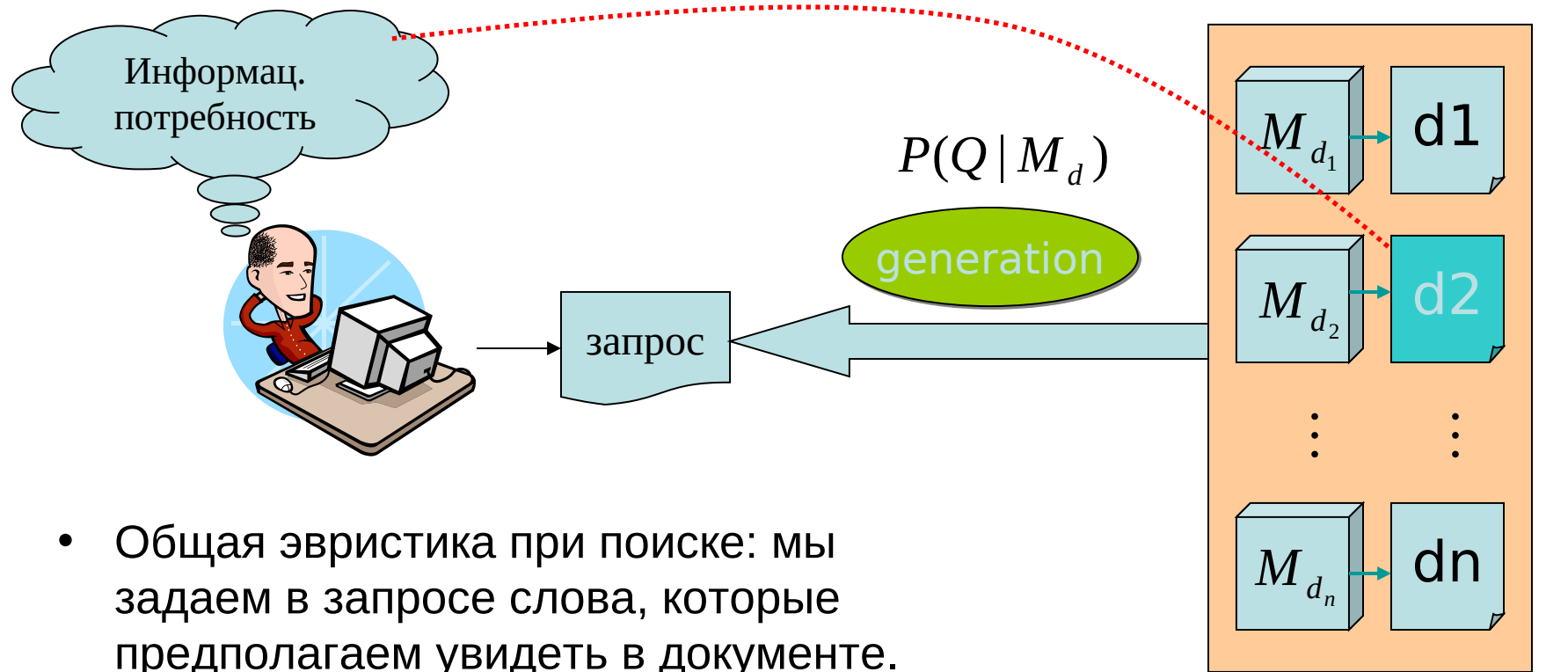
- Если запрос длинный, то можно учитывать похожее взвешивание для слов запроса

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{df_t} \right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b \times (L_d/L_{ave})) + tf_{td}} \cdot \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}}$$

- tf_{tq} : частота слова в запросе q
- k_3 : параметр, контролирующий частоту термина в запросе
- Нет нормализации запроса по длине (поскольку поиск делается для фиксированного запроса)
- Параметры нужно настраивать на коллекции
- Если оптимизация не выполнялась, то в экспериментах получено, что величины k_1 и k_3 должны иметь значения в промежутке $[1.2, 2]$, $b = 0.75$

Языковые модели в информационном поиске

Информационный поиск, основанный на языковой модели (Language Model (LM))



- Общая эвристика при поиске: мы задаем в запросе слова, которые предполагаем увидеть в документе.
- **Подход на основе языковых моделей напрямую использует эту идею**

Коллекция документов

Пример

$$p(q = \text{"presidential campaign"} | d = \text{"... news of presidential campaign ... presidential candidate ..."})$$

Как посчитать такую вероятность? – языковая модель

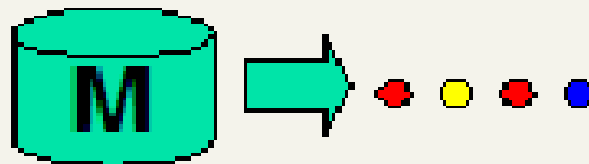
Языковые модели (language models)

- Статистические модели: определение вероятности предложений, последовательностей слов
- Как вероятна каждая последовательность?
 - $P(w_1, w_2, w_3, \dots, w_n)$
 - $P(w_5 | w_1, w_2, w_3, w_4)$
- Языковая модель – математическая модель, которая вычисляется вероятность последовательности слов или условную вероятность следования слова в контексте

Языковые вероятностные модели

- Статистическая модель порождения текста определяет вероятности строк для данного языка

- A statistical model for generating text
 - Probability distribution over strings in a given language



$$P(\text{red, yellow, red, blue} | M) = P(\text{red} | M)$$

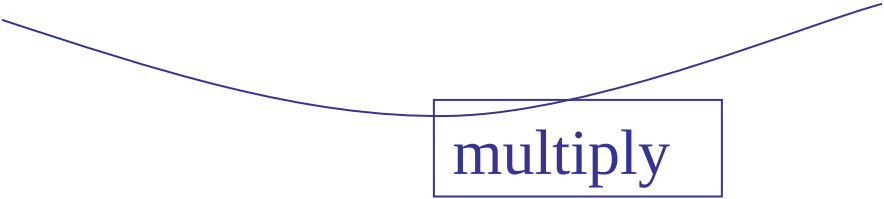
$$P(\text{yellow} | M, \text{red})$$

$$P(\text{red} | M, \text{red, yellow})$$

$$P(\text{blue} | M, \text{red, yellow, red})$$

Статистические языковые модели

- Моделируют вероятность порождения строк в языке

Model M		Униграммная модель:				
0.2	the	the	man	likes	the	woman
0.1	a	—	—	—	—	—
0.01	man	0.2	0.01	0.02	0.2	0.01
0.01	woman					
0.03	said					
0.02	likes					
...		$P(s \mid M) = 0.000000008$				

Использование языковых моделей в информационном поиске

- Трактуем каждый запрос как случайный процесс
- Подход:
 - Насчитывает языковые модели для каждого документа
 - Выводим вероятность порождения запроса на основе модели каждого документа
 - Ранжируем документы в соответствии с этими вероятностями
 - Обычно используются униграммы

Вероятность порождения запроса (1)

- Ранжирующая формула

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

- Вероятность порождения запроса на основе языковой модели документа d на основе MLE:

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

Предположение об униграммах:
При условии определенной языковой модели, слова запроса встречаются независимо

M_d языковая модель документа d

$tf_{(t,d)}$: количество упоминаний tf терма t в документе d

dl_d : общее число слов в документе d

Нехватка данных

- Нулевая вероятность $p(t | M_d) = 0$
- Модель не должна приписывать нулевую вероятность запросу, который содержит слово, отсутствующее в документе: **сглаживание**
- Общий подход в информационном поиске
 - Не встречавшееся в документе слово возможно, но не должно быть более вероятно, чем случайно встреченное в коллекции
 - Если $tf_{(t,d)} = 0$ то
$$p(t | M_d) = \frac{cf_t}{cs}$$
 - где cf_t – количество вхождений слова в коллекцию
 - cs – количество слов в коллекции

Сглаживание

- Необходимо уйти от нулевых вероятностей => техники сглаживания
- Методы
 - Можно использовать: добавить 1, $\frac{1}{2}$ или ϵ к частотам упоминания
 - Здесь: смешивание статистики от документа и коллекции

Смешанная модель

- $P(w|d) = \lambda P_{\text{mle}}(w|M_d) + (1 - \lambda)P_{\text{mle}}(w|M_c)$
- P_{mle} – это просто вероятность, посчитанная без сглаживания (Maximum likelihood estimation)
- Соединяет вероятность слова в документе и вероятность в коллекции целиком
- Необходимо корректный подбор λ
- Высокое значение λ делает поиск более похожий на булевский (требуется упоминание всех слов в запросе), больше подходит для коротких запросов
- Низкое значение – больше подходит для длинных запросов
- Можно настраивать λ для оптимизации качества поиска

Итоговая смешанная модель

- Общая формулировка языковой модели для информационного поиска

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

Модель коллекции

Индивидуальная модель документа

- Пользователь имеет документ в уме и порождает запрос из этого документа
- Равенство выражает вероятность, что документ, который имел в виду пользователь именно этот

Пример

- Коллекция – 2 документа
 - d_1 : Xerox reports a profit but revenue is down
 - d_2 : Lucent narrows quarter loss but revenue decreases further
- Модель: MLE по униграммам из документа; $\lambda = 1/2$
- Запрос: *revenue down*
 - $P(Q|d_1) = [(1/8 + 2/16)/2] \times [(1/8 + 1/16)/2]$
 $= 1/8 \times 3/32 = 3/256$
 - $P(Q|d_2) = [(1/8 + 2/16)/2] \times [(0 + 1/16)/2]$
 $= 1/8 \times 1/32 = 1/256$
- Ранжирование: $d_1 > d_2$

Эксперименты Ponte&Croft (1998)

- Данные:
 - Топики TREC 202-250 (диски 2, 3)
 - Запросы на естественном языке
 - Топики TREC 51-100 (диск 3 с использованием указанных концептов)
 - Список хороших терминов

<num>Number: 054

<dom>Domain: International Economics

<title>Topic: Satellite Launch Contracts

<desc>Description:

... </desc>

<con>Concept(s):

1. Contract, agreement

2. Launch vehicle, rocket, payload, satellite

3. Launch services, ... </con>

Precision/recall: топики 202-250

	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*

Precision/recall: топики 51-100

	tf.idf	LM	%chg	I/D	Sign	Wilc.
Rel	10485	10485				
Rret.:	5818	6105	+4.93	32/42	0.0005*	0.0003*
Prec.						
0.00	0.7274	0.7805	+7.3	10/22	0.7383	0.2961
0.10	0.4861	0.5002	+2.9	26/44	0.1456	0.1017
0.20	0.3898	0.4088	+4.9	24/45	0.3830	0.1405
0.30	0.3352	0.3626	+8.2	28/47	0.1215	0.0277*
0.40	0.2826	0.3064	+8.4	25/45	0.2757	0.0286*
0.50	0.2163	0.2512	+16.2	26/40	0.0403*	0.0007*
0.60	0.1561	0.1798	+15.2	20/30	0.0494*	0.0025*
0.70	0.0913	0.1109	+21.5	14/22	0.1431	0.0288*
0.80	0.0510	0.0529	+3.7	8/13	0.2905	0.2108
0.90	0.0179	0.0152	-14.9	1/4	0.3125	undef
1.00	0.0005	0.0004	-11.9	1/2	0.7500	undef
Avg:	0.2286	0.2486	+8.74	32/50	0.0325*	0.0015*
Prec.						
5	0.5320	0.5960	+12.0	15/21	0.0392*	0.0125*
10	0.5080	0.5260	+3.5	14/30	0.7077	0.1938
15	0.4933	0.5053	+2.4	14/28	0.5747	0.3002
20	0.4670	0.4890	+4.7	16/34	0.6962	0.1260
30	0.4293	0.4593	+7.0	20/32	0.1077	0.0095*
100	0.3344	0.3562	+6.5	29/45	0.0362*	0.0076*
200	0.2670	0.2852	+6.8	29/44	0.0244*	0.0009*
500	0.1797	0.1881	+4.7	30/42	0.0040*	0.0011*
1000	0.1164	0.1221	+4.9	32/42	0.0005*	0.0003*
RPr	0.2836	0.3013	+6.24	30/43	0.0069*	0.0052*

Языковые модели vs. вероятностные модели-1

- Основное различие состоит в том, фигурирует ли понятие релевантности эксплицитно в модели или нет
 - Подход, основанный на языковой модели, пытается избежать моделирования релевантности
- Подход, основанный на языковых моделях предполагает, что документы и запросы представляются собой сущности одного типа
- Невысокая вычислительная сложность, интуитивно понятно

Языковые модели vs. вероятностные модели-2

- Проблемы базового подхода на основе языковых моделей
 - Очень простая языковая модель
 - Трудно интегрировать пользовательскую разметку (Relevance feedback), пользовательские предпочтения, и др.
 - Не может легко интегрировать фразы, абзацы, булевские операторы
 - Текущие исследования посвящены вопросам оптимальной интеграции дополнительной информации

Сравнение с векторной моделью

- Имеется сходство с традиционными tf.idf моделями:
 - Частота слова непосредственно присутствует в модели
 - Вероятности нормализуют по длине частоты слов
 - Эффект смешанной модели похож на idf: слова, редкие в коллекции, но частые в документе имеют большое воздействие на ранжирование

Сравнение с векторной моделью-2

- **Сходство**
 - Веса слов базируются на частотах
 - Слова трактуются как независимые
 - Используется обратная частота по коллекции или документам коллекции
 - Используется некоторая форма нормализации длины
- **Различие**
 - Основывается на вероятности, а не сходстве; аналогии скорее вероятностные, чем геометрические
 - Детали использования длины документа, частот слова в документе/коллекции различаются

Домашнее задание-7 (на неделю)

- Запрос к поисковой системе состоит из двух слов: a b
- В коллекции имеются следующие документы:
-
- a b c d
- a a a
- b b c
- a b b c
-
- Других документов в коллекции нет.
- Примените языковую модель к этой коллекции.
- Сравните $\lambda=0.5$ и $\lambda=0.9$
- Как упорядочатся документы при этих значениях λ ? Какая выдача кажется более правильной?

Домашние задания 8-9.

- 8. Реализовать языковую модель информационного поиска для поиска предложений в Википедии
 - $\lambda=0.5$ и $\lambda=0.9$
 - Прислать 04 ноября
- 9. Оценить NDCG языковой модели по вашим трем запросам и сравнить с предыдущими моделями
 - 11 ноября

Пояснение к заданию 9

- Запросы по Википедии
- Для каждого запроса сделать идеальную разметку, т.е. разметить предложения, насколько они релевантны запросу
 - Использовать трехбалльную шкалу $\{2, 1, 0\}$
 - 2 – предложение содержит полный факт
 - 1 – предложение содержит часть факта
- Оценить качество выдачи, используя NDCG
- Представить отчет по почте
 - Идеальное расположение предложений с оценками,
 - Расчет NDCG для каждого запроса и варианта выдачи
 - Найти среднее для каждой из моделей (2 векторные модели, 2 языковые модели). Сделать выводы