

CS 589 Fall 2021 Lecture 2

**Probability ranking principle
Probability retrieval models**

**Monday 6:30-9:00
Babbio 122**

All zoom links in Canvas



photo: <https://www.scubedstudios.com/information-retrieval/>

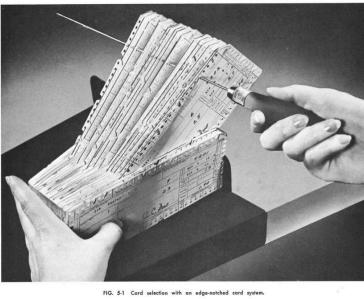
A Brief History of IR

300 BC



Callimachus: the first library catalog

1950s



Punch cards, searching at 600 cards/min

1958

Cranfield evaluation methodology; word-based indexing

1960s

building IR systems on computers; relevance feedback

1970s

TF-IDF; probability ranking principle

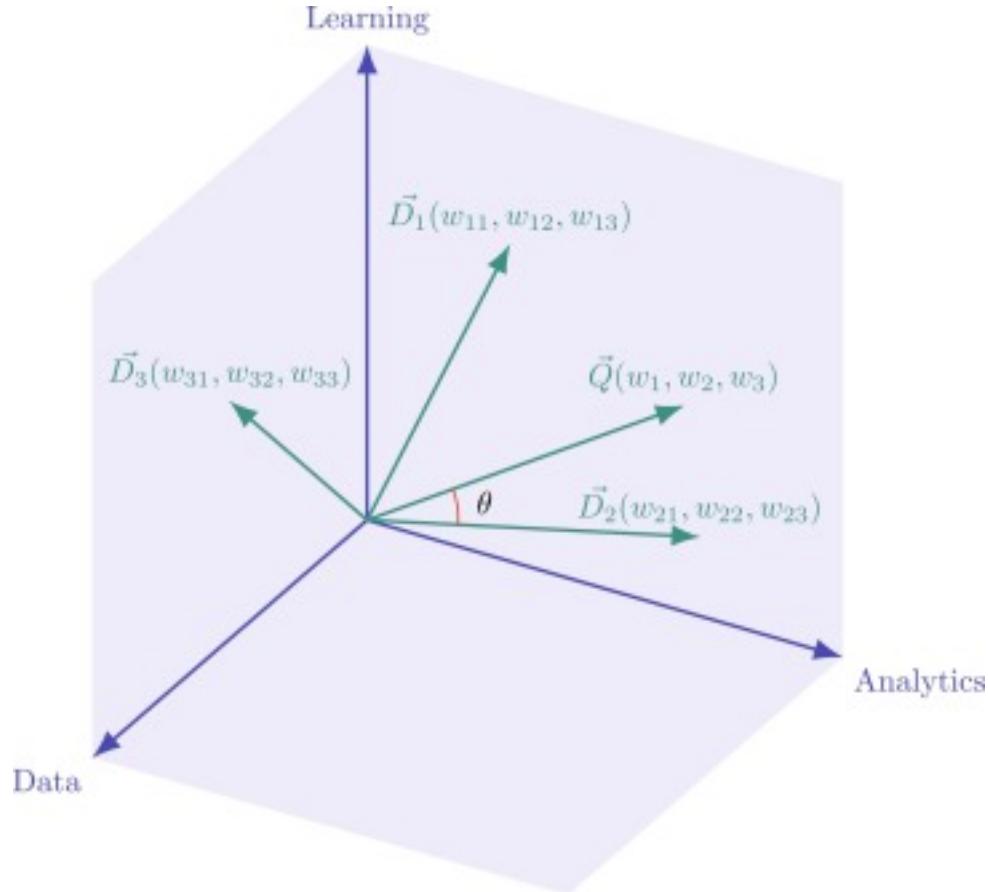
1980s

TREC; learning to rank; latent semantic indexing

1990 - now

web search; supporting natural language queries;

Review of Lecture 1: Vector-Space model



Vector space model:

$$score(q, d) = \frac{q \cdot d}{\|q\| \cdot \|d\|}$$

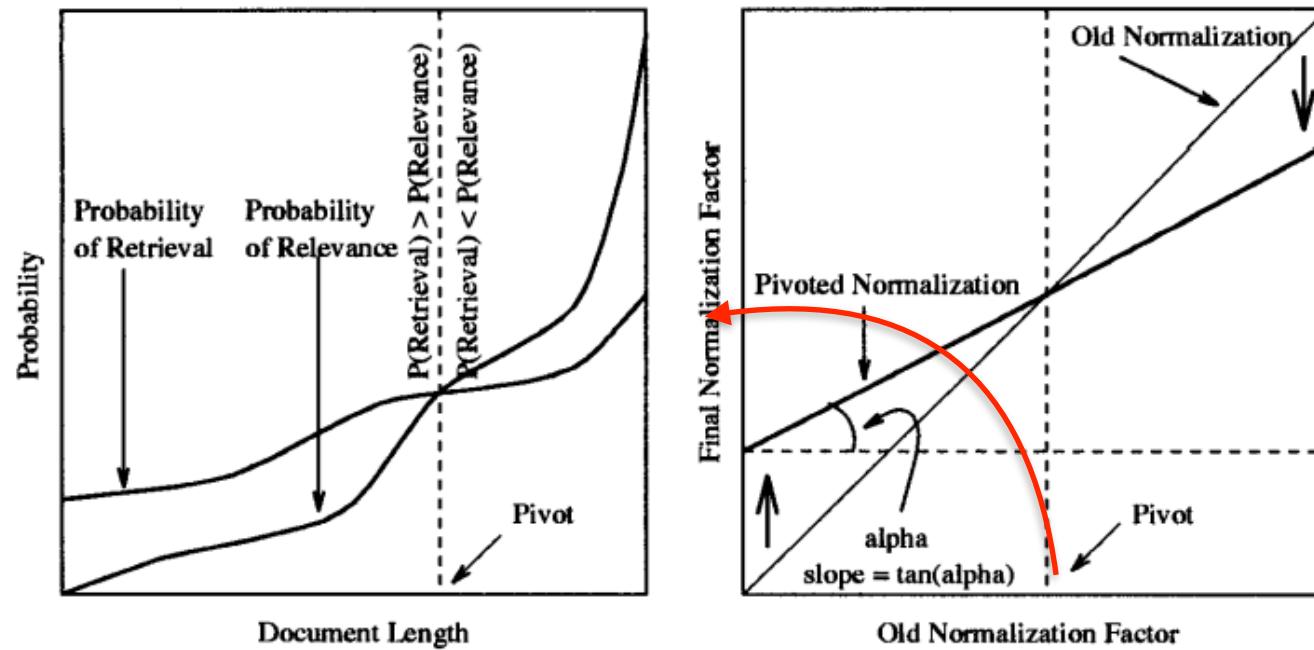
TF-IDF weighting:

$$IDF(w) = \log N / df(w)$$

$$q(d, w) = TF(d, w) \times IDF(w)$$

Document length pivoting

- Rotate the relevance score curve, such that it most closely align with the relevance judgement curve



$$y = x$$

$$\text{pivot} = \text{pivot} \times \text{slope} + \text{intercept}$$

$$\text{pivot}_\text{normalized} = (1.0 - \text{slope}) \times \text{pivot} + \text{slope} \times \text{old normalization}$$

4

Document length pivoting

- Rotate the relevance score curve, such that it most closely align with the relevance judgement curve

$$\frac{tf \cdot idf \text{ weight}}{(1.0 - slope) \times pivot + slope \times old \text{ normalization}}$$

$$\frac{tf \cdot idf \text{ weight}}{1 + \frac{slope}{(1.0 - slope) \times pivot} \times old \text{ normalization}} \quad \leftarrow \text{fixing pivot to avgdl}$$

$$(1.0 - slope) + slope \times \frac{old \text{ normalization}}{\text{average old normalization}} \quad \leftarrow \quad 1 - b + b \frac{|dl|}{|avgdl|}$$

standard formulation for doc length normalization

5

Pop quiz

- Suppose we have one query and two documents:
 - $q = \text{"covid 19"}$
 - $d_1 = \text{"covid patient"}$
 - $d_2 = \text{"19 99 car wash"}$
 - $d_3 = \text{"19 street covid testing facility is reopened next week"}$
- If we use the vector space model **without** IDF, which document has the highest and the lowest score?
 - A. d_1, d_2
 - B. d_1, d_3
 - C. d_3, d_1
 - D. d_3, d_2

Lecture 2

- Review of fundamentals in stats
 - Random variables, Bayes rules, maximum likelihood estimation
- Probabilistic ranking principle
- Probability retrieval models
 - Model 1: Robertson & Spark Jones model (RSJ model)
 - Model 2: BM25 model
 - Model 3: Language model based retrieval model

Random variables

- Random variables: biased coin



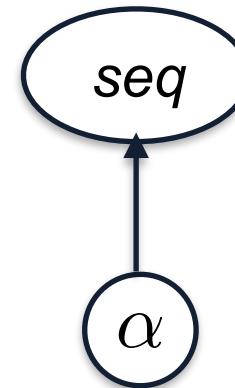
sequence $\neq 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0$

Observation

Bernoulli distribution:

$$p(\text{up}) = \alpha, p(\text{down}) = 1 - \alpha$$

Parameter: α



Random variables

- Random variables: unbiased coin



sequence $\neq 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0$

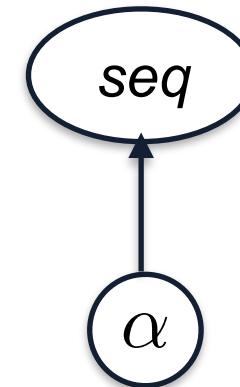
Observation

Bernoulli distribution:

$$p(\text{sequence}) = \alpha \times (1 - \alpha) \times \dots \times (1 - \alpha) \times (1 - \alpha)$$

$$= \alpha^{\#\text{up}} \times (1 - \alpha)^{\#\text{down}}$$

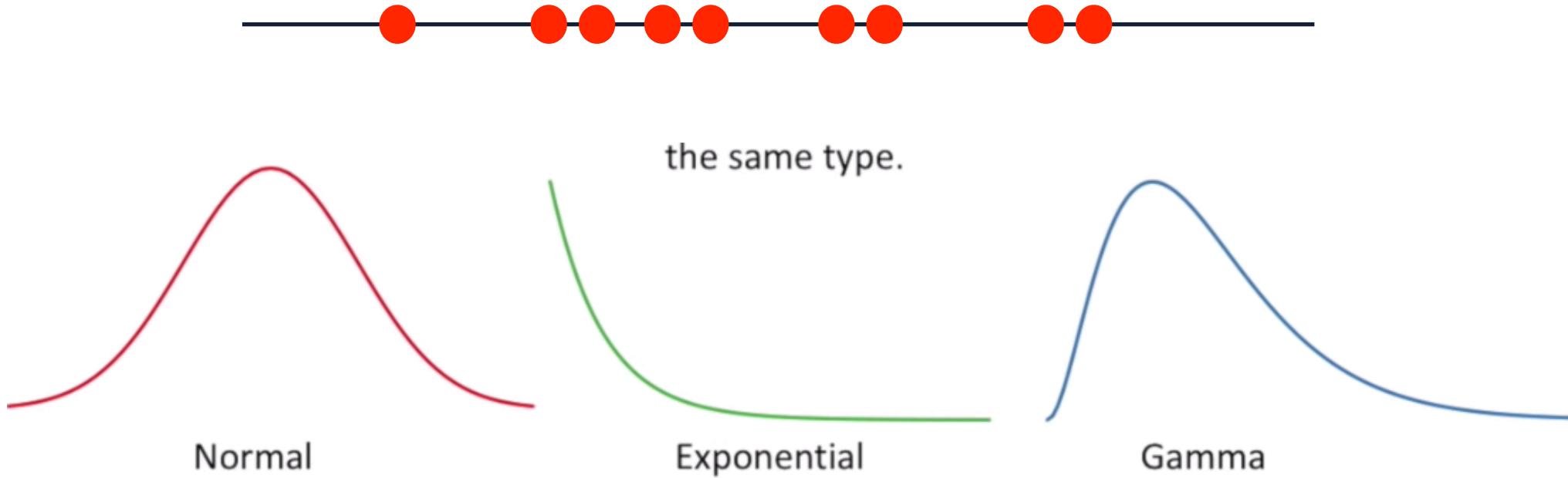
$$\Rightarrow \alpha = \frac{\#\text{up}}{\#\text{up} + \#\text{down}}$$



Maximum likelihood estimation

Maximum likelihood estimation

- Fitting a distribution model to the data
 - Assumes mouse weights follow an underlying distribution



Bayes' rules

Chain rule: joint distribution

$$P(A, B) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

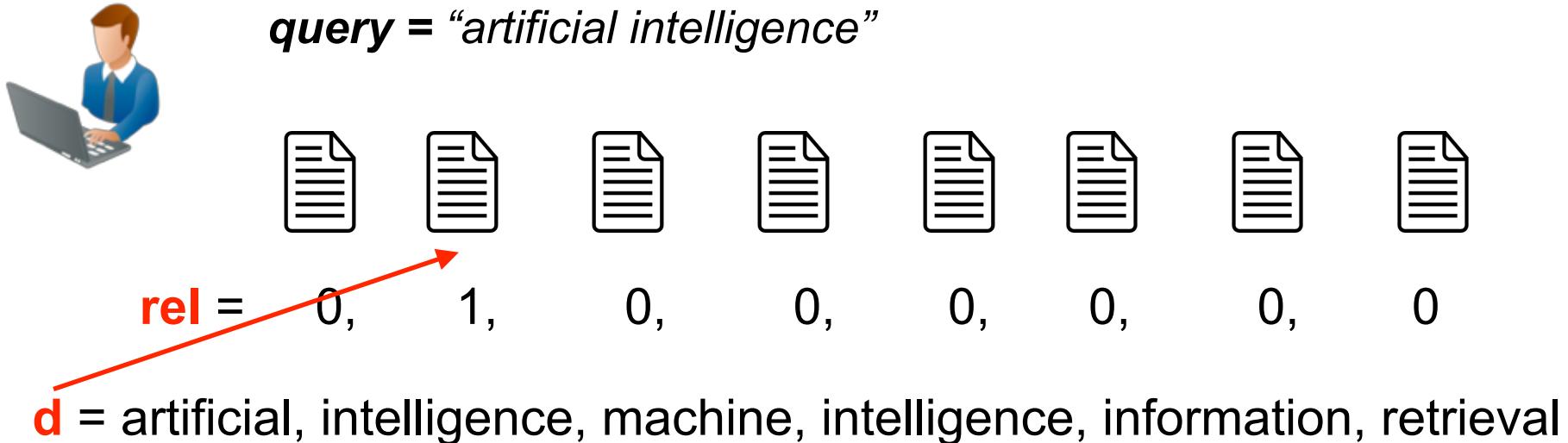
Bayes' rule:

posterior likelihood prior

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \left[\frac{P(B|A)}{\sum_{X \in \{A, \bar{A}\}} P(B|X)P(X)} \right] P(A)$$

$$P(A|B) \propto P(B|A)P(A)$$

Random variables in information retrieval



$q = \text{artificial, intelligence}$

Notations: in future slides, q denotes the query, d denotes the document, rel denotes the relevance judgment

Probability ranking principle

- Assume documents are labelled by 0/1 labels (i.e., the relevance judgement is either 0 or 1), given query q , documents should be ranked on their probabilities of relevance (*van Rijsbergen 1979*):

PRP: rank documents by $p(\text{rel} = 1|q, d)$

- **Theorem.** *The PRP is optimal, in the sense that it minimizes the expected loss (Ripley 1996)*

Estimating $p(rel = 1|q, d)$

$$\begin{aligned} p(rel = 1|q, d) &= \frac{p(rel = 1, q, d)}{p(q, d)} \\ &= \frac{\text{count}(rel = 1, q, d)}{\text{count}(q, d)} \end{aligned}$$

Problems with this estimation?

1. not enough data
2. cannot adapt to new q

$$p(rel = 1|q, d) \propto p(d|rel = 1, q)p(rel = 1) \quad \text{Bayes rule}$$

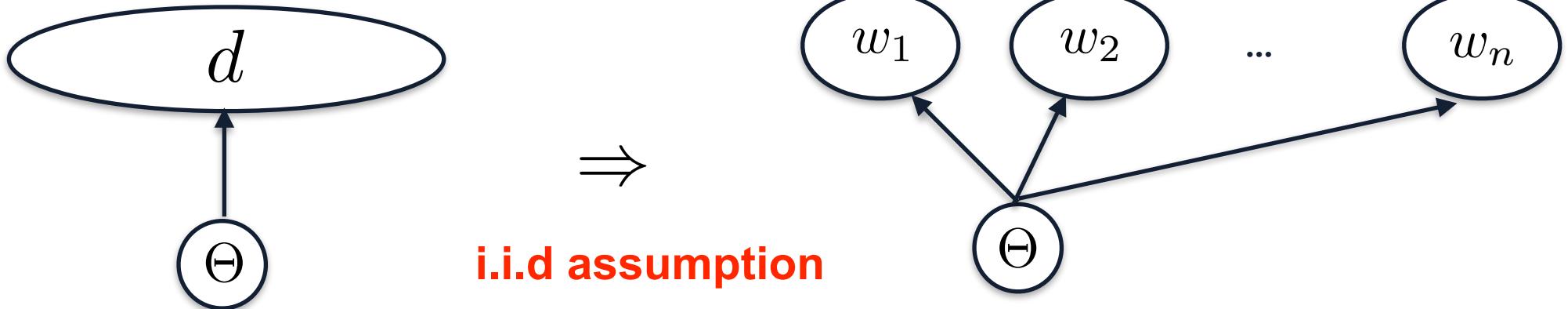
Estimating $p(rel = 1|q, d)$

$$p(rel = 1|q, d) \propto p(d|rel = 1, q)p(rel = 1)$$

$$\begin{aligned} O(rel = 1|q, d) &= \frac{p(rel = 1|q, d)}{p(rel = 0|q, d)} \\ &= \frac{p(d|rel = 1, q)p(rel = 1)}{p(d|rel = 0, q)p(rel = 0)} \end{aligned}$$

↑
odds

Estimating the generative model $p(d|rel = 1, q)$



$$p(d|rel = 1, q) = \prod_i p(w_i|rel = 1, q)$$

$$O(rel = 1|q, d) = \prod_i \frac{p(w_i|rel = 1, q)}{p(w_i|rel = 0, q)} \times \frac{p(rel = 1)}{p(rel = 0)}$$

$$\stackrel{rank}{=} \prod_{w_i=1} \frac{\alpha_i}{\beta_i} \times \prod_{w_i=0} \frac{(1 - \alpha_i)}{(1 - \beta_i)} \quad (1)$$

$$\stackrel{rank}{=} \prod_{w_i=1} \frac{\alpha_i}{\beta_i} \times \prod_{w_i=1} \frac{(1 - \beta_i)}{(1 - \alpha_i)} \times const \quad (2)$$

$$= \prod_{w_i=1} \frac{\alpha_i(1 - \beta_i)}{\beta_i(1 - \alpha_i)} \quad (3)$$

$$\stackrel{rank}{=} \sum_{w_i=1} \log \frac{\alpha_i(1 - \beta_i)}{\beta_i(1 - \alpha_i)} \quad (4)$$

$$\alpha_i = p(w_i = 1|rel = 1, q)$$

$$\beta_i = p(w_i = 1|rel = 0, q)$$

Model 1: RSJ model

$$O(rel = 1|q, d) \stackrel{rank}{=} \sum_{w_i=1} \log \frac{\alpha_i(1 - \beta_i)}{\beta_i(1 - \alpha_i)}$$

(Robertson & Sparck Jones 76)

$$\begin{aligned}\alpha_i &= p(w_i = 1|q, rel = 1) \\ &= \frac{count(w_i = 1, rel = 1) + 0.5}{count(rel = 1) + 1}\end{aligned}$$

Probability for a word to appear in a relevant doc

$$\begin{aligned}\beta_i &= p(w_i = 0|q, rel = 0) \\ &= \frac{count(w_i = 0, rel = 0) + 0.5}{count(rel = 0) + 1}\end{aligned}$$

Probability for a word to appear in a non-relevant doc

RSJ model: Summary

- Uses only **binary word** occurrence (binary inference model), does not leverage TF information
 - RSJ model was designed for retrieving short text and abstract!
- Requires relevance judgment
 - No-relevance judgment version: [Croft & Harper 79]
- Performance is not as good as tuned vector-space model

How to improve RSJ based on these desiderata?

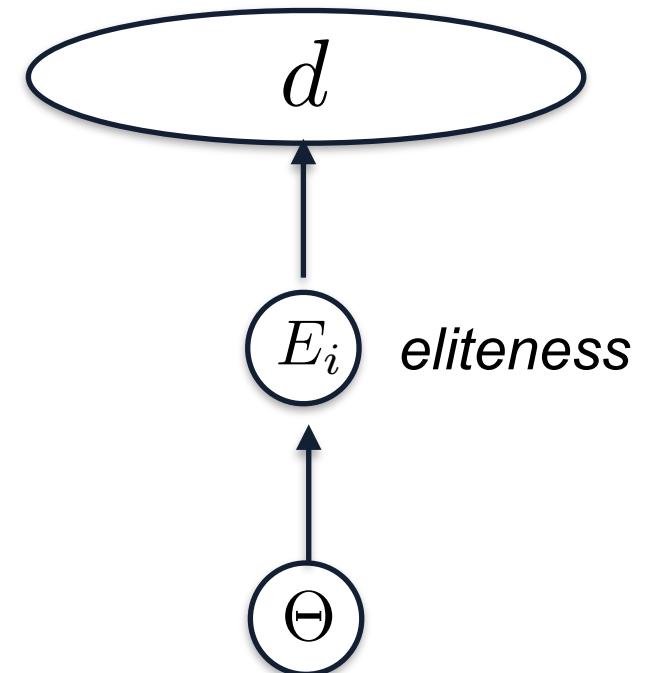
Desiderata of retrieval models

- Recall the desiderata of a retrieval models:
 - The importance of TF is sub-linear
 - Penalizing term with large document frequency using IDF
 - Pivot length normalization

How to improve RSJ based on these desiderata?

Model 2: Okapi/BM25

- Introduced in 1994
 - SOTA non-learning retrieval model
- Estimate probability using **eliteness**
 - *Hidden variable between each document-term pair*
 - *A document-term pair is elite if the document is **about the concept denoted by the term**, e.g., for “**artificial intelligence**”, an elite word is “**learning**”*
 - *Eliteness is binary*
 - *Term occurrence depends on eliteness*



Okapi/BM25

$$\alpha_i = p(w_i = 1 | rel = 1, q)$$

$$O(rel = 1 | q, d) \stackrel{rank}{=} \sum_{w_i=}$$

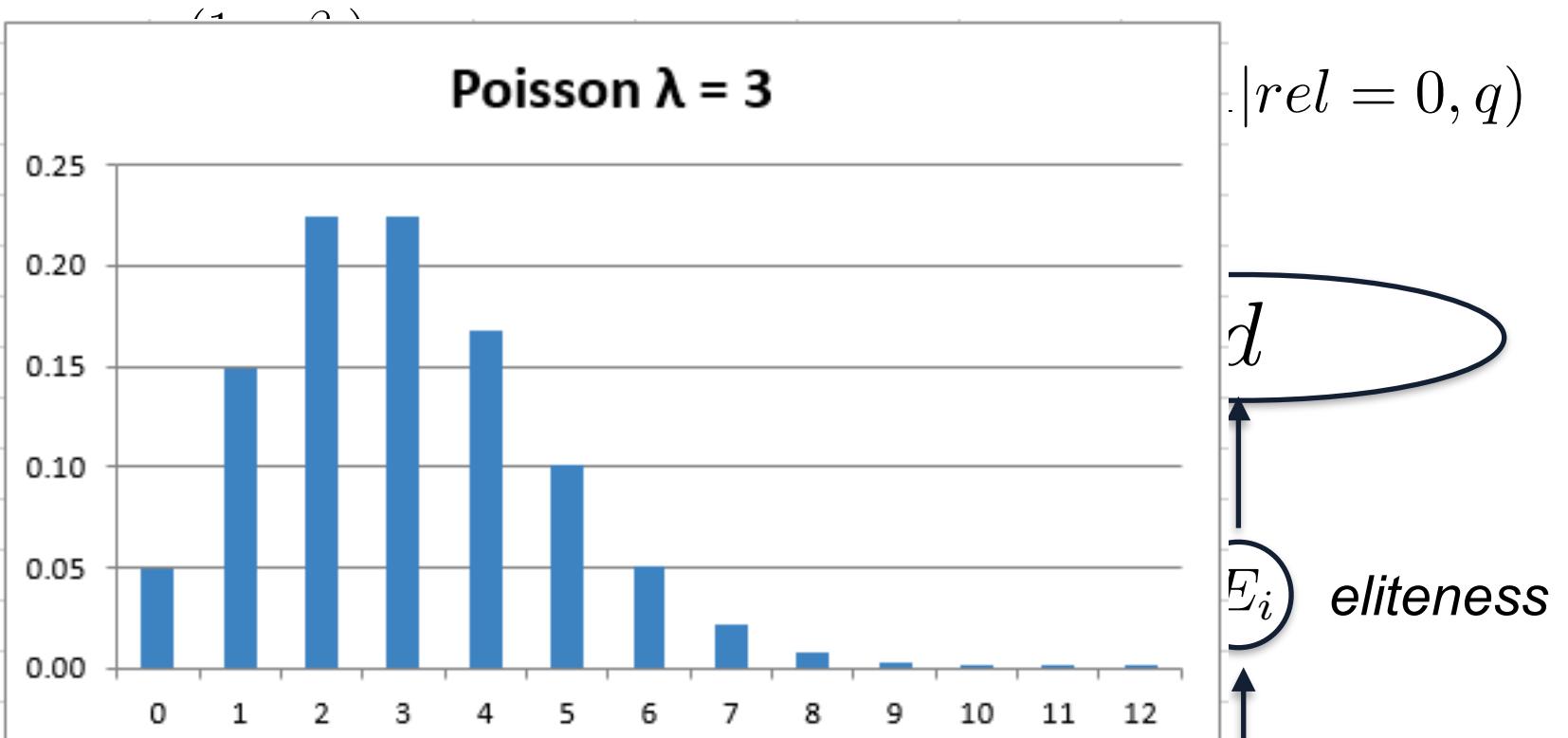
$$score(q, d) = \sum_{i \in q} c_i^{elite}$$

$$c_i^{elite}(tf_i) = \log \frac{p(w)}{p(w)}$$

$$p(w_i = tf_i | q, rel = 1)$$

$$+ p(w_i = tf_i | E_i = 0) p(E_i = 0 | q, rel)$$

$$= \pi \frac{\lambda^{tf_i}}{tf_i!} e^{-\lambda} + (1 - \pi) \frac{\mu^{tf_i}}{tf_i!} e^{-\mu} \quad (2 \text{ Poisson model})$$

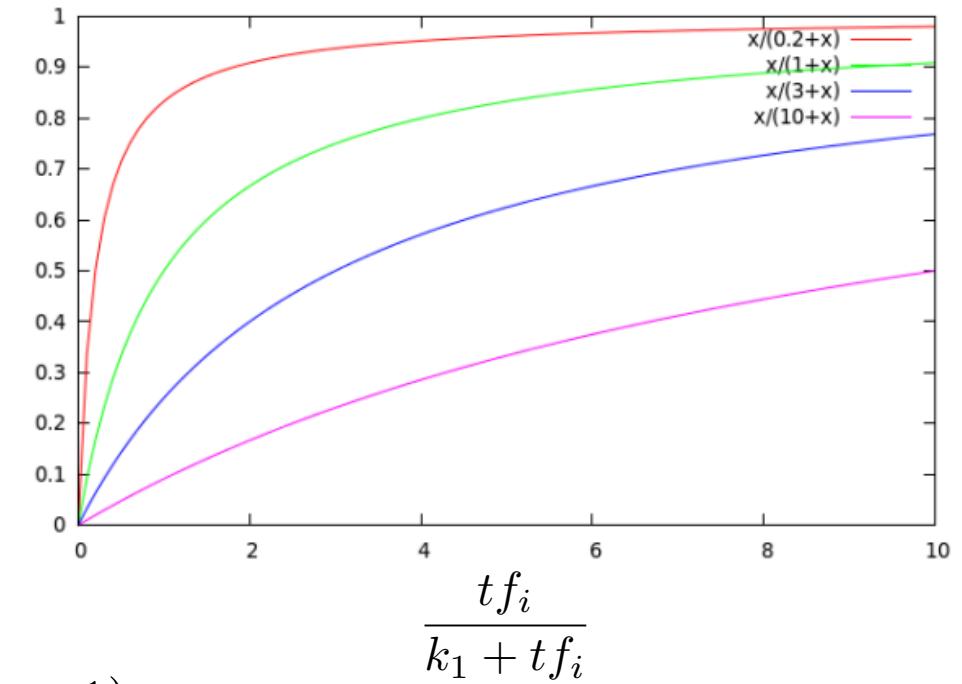
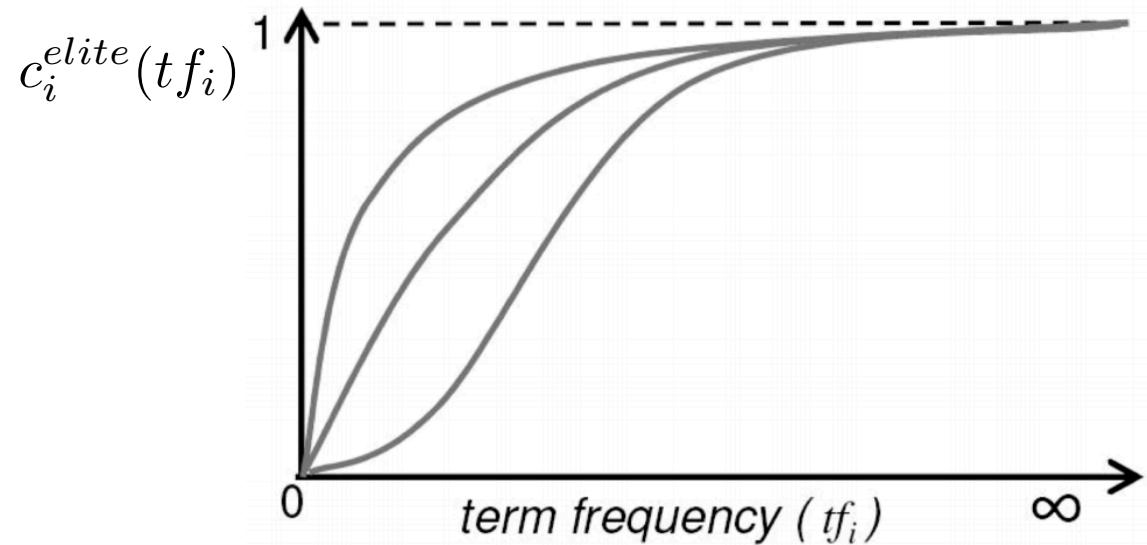


Okapi/BM25

$$p(w_i = tf_i | q, rel = 1) = \boxed{\pi \frac{\lambda^{tf_i}}{tf_i!} e^{-\lambda} + (1 - \pi) \frac{\mu^{tf_i}}{tf_i!} e^{-\mu}}$$

- We do not know λ, μ, π
- MLE for λ, μ, π ? Difficulty to estimate
- Designing a parameter-free model such that it simulates $p(w_i = tf_i | q, rel = 1)$

Simulating the 2-Poisson model



$$c_i^{BM25}(tf_i) \approx \log \frac{N}{df_i} \times \frac{tf_i(k_1 + 1)}{k_1(1 - b + b \frac{|dl|}{avgdl}) + tf_i} \quad b = 0.75, k_1 \in [1.2, 2.0]$$

Analysis of BM25 formulation

$$c_i^{BM25}(tf_i) \approx \log \frac{N}{df_i} \times \frac{tf_i(k_1 + 1)}{k_1(1 - b + b \frac{|dl|}{avgdl}) + tf_i}$$

IDF

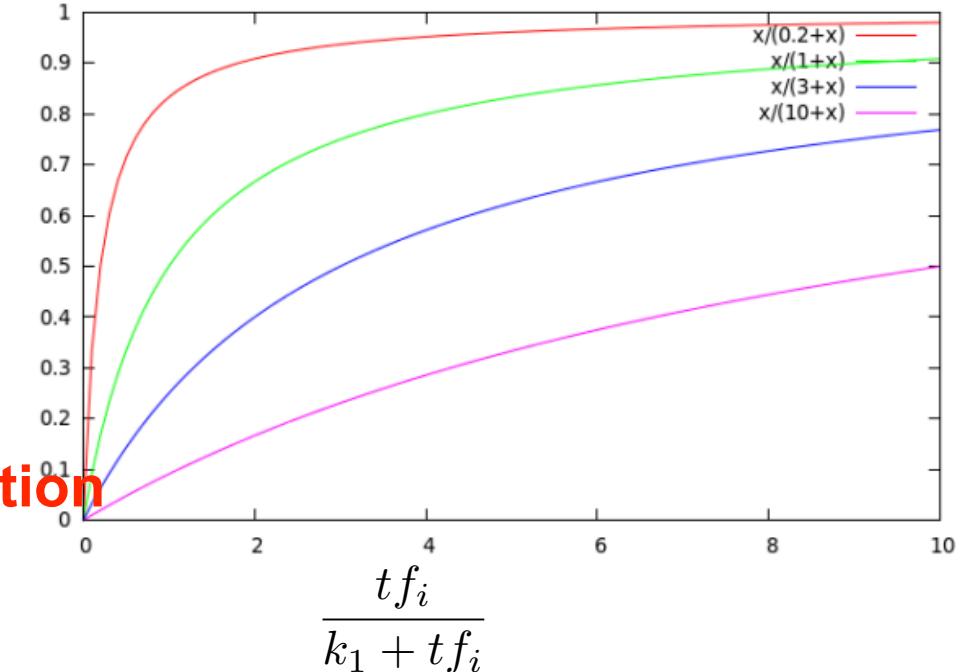
Pivoted document length normalization

$$(1.0 - slope) + slope \times \frac{\text{old normalization}}{\text{average old normalization}}$$



$$1 - b + b \frac{|dl|}{avgdl}$$

$b = 0.75, k_1 \in [1.2, 2.0]$



Multi-field retrieval

How should I cite presentation slides?

Asked 6 years, 10 months ago Active 5 years, 9 months ago Viewed 8k times

A friend has made some nice slides that I could reuse (similar topics). He sent me the slides and commented that if I use them and could cite him that would be nice, I asked him how should I cite the slides but he said that whatever suits better to me he said "Just add my surname in some place where it's not very intrusive".

9

I'm not sure if he doesn't care or he doesn't want to be too picky, but I'd like to cite him, to each one his own.

1

AFAIK, they are related to a paper (but not in the paper) and to his thesis, where they could be as a diagram but definitely not animated. The slides (as such) may be available at some URL, he said they will be but they are not available yet (so I don't have the URL yet). If citing by the URL I guess I could use this: "[How to cite a website URL?](#)"

Should I cite slides? If yes, how?

citations

title

question

BM25F

$$score^{BM25F}(q, d) = \sum_{w_i \in q} \log \frac{N}{df_i} \times \frac{tf_i^F (k_1 + 1)}{k_1 \left(1 - b + b \frac{|dl|^F}{|avgdl|^F} \right) + tf_i^F}$$

- Each variable is estimated as the weighted sum of its field value

$$tf_i = \sum_f \alpha_f \times tf_{i,f} \quad dl = \sum_f \alpha_f \times dl_f \quad avgdl = \sum_f \alpha_f \times avgdl_f$$

alpha_f: hyper parameters

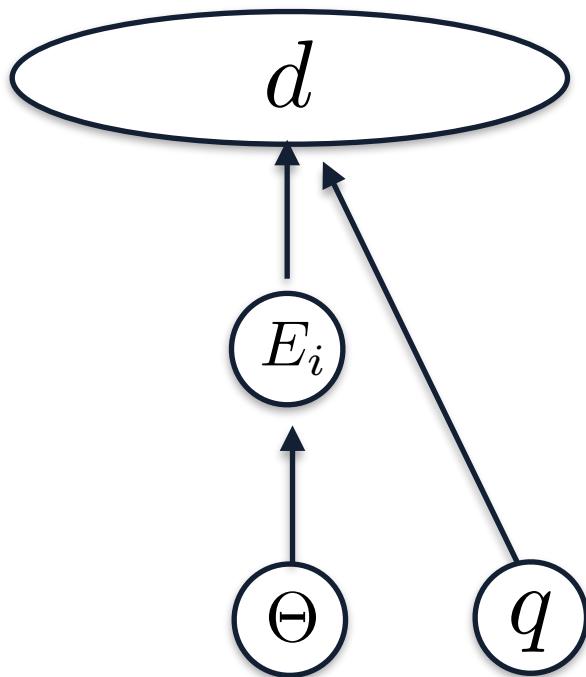
Multi-field retrieval

- HW1: BM25 outperforms TF-IDF in every field & combined

	[1.0, 0.0, 0.0]	[0.0, 1.0, 0.0]	[0.0, 0.0, 1.0]	[1.0, 0.5, 0.5]
Python, bm2, ndcg@10	0.319	0.322	0.293	0.378
Python, tfidf, ndcg@10	0.317	0.274	0.276	0.355
Java, bm2, ndcg@10	0.327	0.287	0.254	0.376
Java, tfidf, ndcg@10	0.315	0.258	0.238	0.349
Javascript, bm2, ndcg@10	0.349	0.330	0.267	0.407
Javascript, tfidf, ndcg@10	0.346	0.289	0.247	0.374

Analysis on the n-Poisson model

- **Advantage:** BM25 is based on the 2-Poisson model



A document-term pair is elite if the document is about the concept denoted by the term

- **Disadvantages:**
 - For unigram term, documents will not fall cleanly into elite/non-elite set
 - For term with multiple tokens, requires a combinatorial explosion of elite set
 - Requires explicit indexing of the ‘elite’ words

Model 3: Language model-based retrieval

- A language model-based retrieval method [Ponte and Croft, 1998]

$$score(q, d) = \log p(q|d) = \prod_{i, w_i \in q} p(w_i = 1|d) \prod_{i, w_i \notin q} (1.0 - p(w_i = 1|d))$$

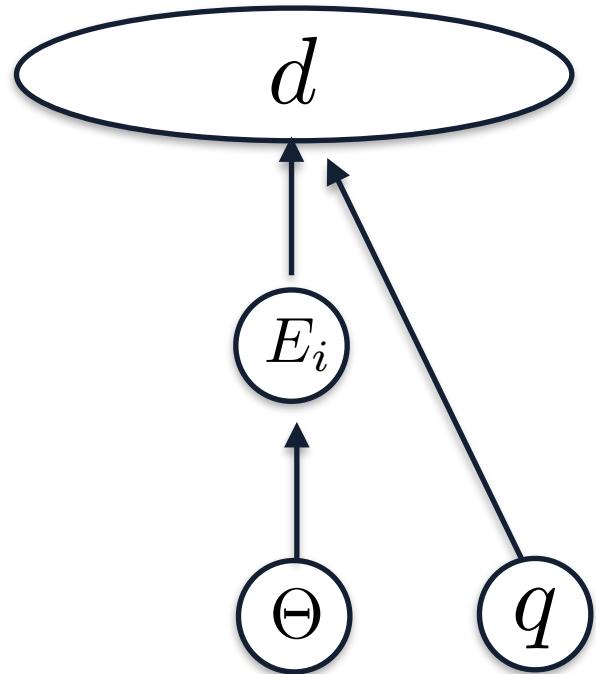
- Bernoulli \rightarrow multinomial

$$score(q, d) = \log p(q|d) = \prod_{w_i=1}^V p(w_i|d)^{c(w_i, q)}$$
$$rank \stackrel{def}{=} \sum_{w_i=1}^V c(w_i, q) \log p(w_i|d)$$
$$p(w_i|d) = \begin{cases} p_{seen}(w_i|d) & \text{if } w_i \text{ is seen in d} \\ \alpha_d p(w_i|C) & \text{o.w.} \end{cases}$$

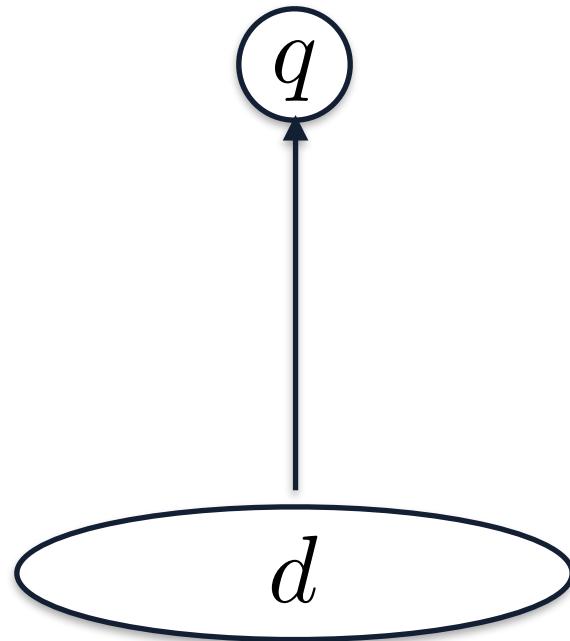
corpus unigram LM



Language model-based retrieval



BM25



Language model

Language model-based retrieval

$$\log p(q|d) = \sum_{w_i}^V c(w_i, q) \log p(w|d)$$

$$= \sum_{w_i, w_i \in d} c(w_i, q) \log p_{seen}(w_i|d) + \sum_{w_i, w_i \notin d} c(w_i, q) \log \alpha_d p(w_i|C)$$

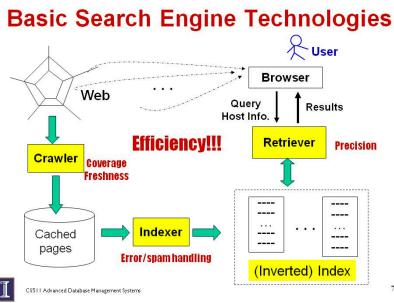
⋮ ⋮ ⋮

$$= \sum_{w_i, w_i \in d} c(w_i, q) \log \frac{p_{seen}(w_i|d)}{\alpha_d p(w_i|C)} + |q| \log \alpha_d + \sum_{w_i=1}^V c(w_i, q) \log p(w_i|C)$$

$$score^{LM}(q, d) \stackrel{rank}{=} \sum_{w_i, w_i \in d} c(w_i, q) \log \frac{p_{seen}(w_i|d)}{\alpha_d p(w_i|C)} + |q| \log \alpha_d \text{ constant}$$

Different senses of ‘model’ [Ponte and Croft, 98]

- First sense (high level): an abstraction of the retrieval task itself

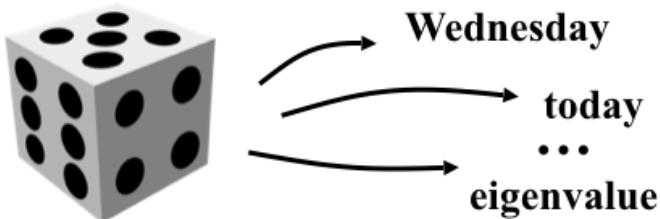


decouple retrieval model + other problems
→ (e.g., indexing)

- Second sense (mid level): modeling the distribution, e.g., 2-Poisson model
 - Thirds sense (low level): which **statistical language model** is used in $p_{seen}(w_i|d)$

Statistical language model

- A probability distribution over word sequences
 - $p(\text{"Today is Wednesday"}) \approx 0.001$
 - $p(\text{"Today Wednesday is"}) \approx 0.000000000001$
 - $p(\text{"The eigenvalue is positive"}) \approx 0.00001$
- Unigram language model
 - Generate text by generating each word INDEPENDENTLY
 - Thus, $p(w_1 w_2 \dots w_n) = p(w_1)p(w_2)\dots p(w_n)$
 - Parameters: $\{p(t_i)\}$ $p(t_1) + \dots + p(t_N) = 1$ (N is voc. size)



$$\begin{aligned} & p(\text{"today is Wed"}) \\ &= p(\text{"today"})p(\text{"is"})p(\text{"Wed"}) \\ &= 0.0002 \times 0.001 \times 0.000015 \end{aligned}$$

Notes on language model-based retrieval

- **Advantages:**
 - Avoided the disadvantages in eliteness
 - Defines a general framework, more accurate $p_{seen}(w_i|d)$ can further improve the model
 - In some cases, has outperformed BM25
- **Disadvantages:**
 - The assumed equivalence between query and document is unrealistic
 - Only studied unigram language model
 - Performance is not always good

Equivalence to KL-divergence retrieval model

$$score^{LM}(q, d) \stackrel{rank}{=} \sum_{w_i, w_i \in d} c(w_i, q) \log \frac{p_{seen}(w_i|d)}{\alpha_d p(w_i|C)} + |q| \log \alpha_d$$

- Kullback-Leibler divergence

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$-D(\hat{\theta}_q\|\hat{\theta}_d) = \sum_{w_i=1}^V p(w_i|\hat{\theta}_q) \log p(w_i|\hat{\theta}_d) + \left(- \sum_{w_i=1}^V p(w_i|\hat{\theta}_q) \log p(w_i|\hat{\theta}_q) \right)$$



smoothed

constant

why not the opposite?

$$\stackrel{rank}{=} \sum_{w_i, w_i \in d} p(w_i|\hat{\theta}_q) \log \frac{p_{seen}(w_i|d)}{\alpha_d p(w_i|C)} + \log \alpha_d \quad (\text{Eq. 1})$$

Notes on the KL-divergence retrieval formula and Dirichlet prior smoothing

Estimating $p_{seen}(w_i|d)$

- Estimating $p_{seen}(w_i|d)$ based on the maximum likelihood estimation

$$p_{seen}(w_i|d) = \frac{\text{count}(w_i)}{|dl|}$$

- Disadvantage: if the word is unseen, probability will be 0
- Solution: language model smoothing:

$$\begin{aligned} p_s(w_i|d) &= \frac{c(w_i, d) + \mu p(w_i|C)}{|d| + \mu} & \alpha_d = \frac{\mu}{\mu + |d|} & \quad (\text{plug in Eq. 1}) \\ &= \frac{|d|}{|d| + \mu} p(w_i|d) + \frac{\mu}{|d| + \mu} p(w_i|C) & & \quad \text{Dirichlet smoothing} \end{aligned}$$

corpus unigram LM

Estimating $p_{seen}(w_i|d)$

- Dirichlet smoothing

$$score^{Dir}(q, d) = \sum_{w_i, w_i \in d, p(w_i|\hat{\theta}_q)} p(w_i|\hat{\theta}_q) \log \left(1 + \frac{count(w_i, d)}{\mu p(w_i|C)} \right) + \log \frac{\mu}{\mu + |dl|}$$

- Jelinek-Mercer smoothing

$$p_s(w_i|d) = \lambda p(w_i|d) + (1 - \lambda)p(w_i|C)$$

$$score^{JM}(q, d) = \sum_{w_i, w_i \in d, p(w_i|\hat{\theta}_q)} p(w_i|\hat{\theta}_q) \log \left(1 + \frac{(1 - \lambda)count(w_i, d)}{\lambda p(w_i|C)} \right)$$

Other smoothing methods

- Additive smoothing
- Good-Turing smoothing
- Absolute discounting
- Kneser-ney smoothing

Tuning parameters in smoothing models [Zhai and Lafferty 02]

$$score^{Dir}(q, d) = \sum_{w_i, w_i \in d, p(w_i | \hat{\theta}_q)} p(w_i | \hat{\theta}_q) \log \left(1 + \frac{count(w_i, d)}{\mu p(w_i | C)} \right) + \log \frac{\mu}{\mu + |dl|}$$

- Tuning parameter μ using “leave-one-out” method

$$\hat{\mu} = argmax_{\mu} \sum_{w_i=1}^V \sum_d \log p(w_i | d; w_i \notin d)$$

 **remove** w_i

- Estimating parameter using Newton’s method (2nd derivative)

Tuning parameters in smoothing models [Zhai and Lafferty 02]

$$score^{JM}(q, d) = \sum_{w_i, w_i \in d, p(w_i | \hat{\theta}_q)} p(w_i | \hat{\theta}_q) \log \left(1 + \frac{(1 - \lambda)count(w_i, d)}{\lambda p(w_i | C)} \right)$$

- Tuning parameter λ using MLE for the query probability

$$p(q | \lambda, C) = \sum_d \pi_d \prod_{w_i \in q} ((1 - \lambda)p(w_i | d) + \lambda p(w_i | C))$$

- Expectation-Maximization algorithm:

$$\pi_d^{(k+1)} = \frac{\pi_d^{(k)} \prod_{w_i \in q} ((1 - \lambda^{(k)}) p(w_i | d) + \lambda^{(k)} p(w_i | C))}{\sum_d \pi_d^{(k)} \prod_{w_i \in q} ((1 - \lambda^{(k)}) p(w_i | d) + \lambda^{(k)} p(w_i | C))}$$

$$\lambda^{(k+1)} = \frac{1}{|q|} \sum_d \pi_d^{(k+1)} \sum_{w_i \in q} \frac{\lambda^{(k)} p(w_i | C)}{(1 - \lambda^{(k)}) p(w_i | d) + \lambda^{(k)} p(w_i | C)}$$

Feedback language model [Zhai and Lafferty 01]



[Transportation Security Administration - Official Site](#)
www.tsa.gov ▾ Official site
Charged with providing effective and efficient security for passenger and freight transportation in the United States. Mission, press releases, employment, milestones ...

[Prohibited Items](#)
The My TSA mobile application provides 24/7 access to helpful ...

[TSA Precheck Ad](#)
Learn about TSA Pre™ expedited screening! No longer remove ...

[Careers](#)
TSA is comprised of nearly 50,000 security officers, inspectors, air ...
See results only from tsa.gov

[Airport security - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Airport_security ▾
Airport security refers to the techniques and methods used in protecting passengers, staff and aircraft which use the airports from accidental/malicious harm, crime ...
Airport enforcement ... · Process and equipment · Notable incidents

[An Overview of Airport Security Rules - About](#)
studenttravel.about.com › Student Transportation Options ▾
Airport security rules are a travel drag: get through airport security and get to the fun part (travel) faster by knowing what the airport security rules are in advance.

[News about Airport Security](#)
bing.com/news
[No need to beef up airport security - govt](#)
YahooNews · 1 minute ago
Airport security doesn't need to be strengthened because 30 to 40 New Zealanders are being monitored over links to terrorist groups, the government says. Prime Minister John Key on Wednesday revealed the existence of...

Feedback documents

[Airport security - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Airport_security ▾
Airport security refers to the techniques and methods used in protecting passengers, staff and aircraft which use the airports from accidental/malicious harm, crime ...
Airport enforcement ... · Process and equipment · Notable incidents

[An Overview of Airport Security Rules - About](#)

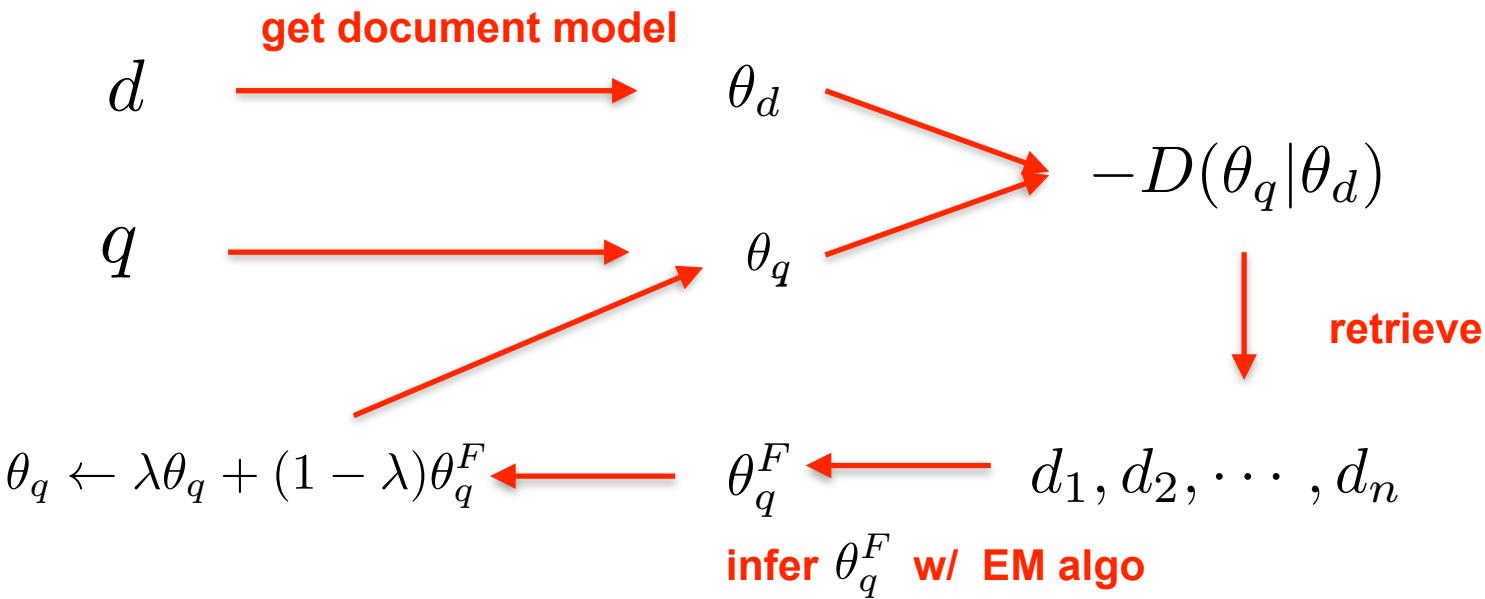
studenttravel.about.com › Student Transportation Options ▾
Airport security rules are a travel drag: get through airport security and get to the fun part (travel) faster by knowing what the airport security rules are in advance.

*protect passengers,
accidental/malicious
harm, crime, rules*

Feedback language model [Zhai and Lafferty 01]

$$score^{JM}(q, d) = \sum_{w_i, w_i \in d, p(w_i | \hat{\theta}_q)} p(w_i | \hat{\theta}_q) \log \left(1 + \frac{(1 - \lambda)count(w_i, d)}{\lambda p(w_i | C)} \right)$$

$$p(w_i | q) = \frac{count(w_i, q)}{|q|} \quad \textcolor{red}{sparsity}$$



Evaluation on smoothing methods [Zhai & Lafferty 02]

Collection	query	Optimal-JM	Optimal-Dir	Auto-2stage
AP88-89	SK	20.3%	23.0%	22.2%*
	LK	36.8%	37.6%	37.4%
	SV	18.8%	20.9%	20.4%
	LV	28.8%	29.8%	29.2%
WSJ87-92	SK	19.4%	22.3%	21.8%*
	LK	34.8%	35.3%	35.8%
	SV	17.2%	19.6%	19.9%
	LV	27.7%	28.2%	28.8%*
ZIFF1-2	SK	17.9%	21.5%	20.0%
	LK	32.6%	32.6%	32.2%
	SV	15.6%	18.5%	18.1%
	LV	26.7%	27.9%	27.9%*

Evaluation on smoothing methods [Zhai & Lafferty 01b]

collection		Simple LM	Mixture	Improv.	Div. Min.	Improv.
AP88-89	AvgPr	0.21	0.296	+41%	0.295	+40%
	InitPr	0.617	0.591	-4%	0.617	+0%
	Recall	3067/4805	3888/4805	+27%	3665/4805	+19%
TREC8	AvgPr	0.256	0.282	+10%	0.269	+5%
	InitPr	0.729	0.707	-3%	0.705	-3%
	Recall	2853/4728	3160/4728	+11%	3129/4728	+10%
WEB	AvgPr	0.281	0.306	+9%	0.312	+11%
	InitPr	0.742	0.732	-1%	0.728	-2%
	Recall	1755/2279	1758/2279	+0%	1798/2279	+2%

Comparison between BM25 and LM [Bennett et al. 2008]

Collection	Method	Parameter	MAP	R-Prec.	Prec@10
Trec8 T	Okapi BM25	Okapi	0.2292	0.2820	0.4380
	JM	$\lambda = 0.7$	0.2310 (p=0.8181)	0.2889 (p=0.3495)	0.4220 (p=0.3824)
	Dir	$\mu = 2,000$	0.2470 (p=0.0757)	0.2911 (p=0.3739)	0.4560 (p=0.3710)
	Dis	$\delta = 0.7$	0.2384 (p=0.0686)	0.2935 (p=0.0776)	0.4440 (p=0.6727)
	Two-Stage	auto	0.2406 (p=0.0650)	0.2953 (p=0.0369)	0.4260 (p=0.4282)
Trec8 TD	Okapi BM25	Okapi	0.2528	0.2908	0.4640
	JM	$\lambda = 0.7$	0.2582 (p=0.5226)	0.3038 (p=0.1886)	0.4600 (p=0.8372)
	Dir	$\mu = 2,000$	0.2621 (p=0.3308)	0.3043 (p=0.1587)	0.4460 (p=0.3034)
	Dis	$\delta = 0.7$	0.2599 (p=0.1737)	0.3105 (p=0.0203)	0.4880 (p=0.1534)
	Two-Stage	auto	0.2445 (p=0.2455)	0.2933 (p=0.7698)	0.4400 (p=0.1351)

However, BM25 outperforms LM in other cases

Summary on Hyperparameter tuning

- RSJ: no parameter
- BM25: Due to the formulation of two-Poisson, parameters are difficult to estimate, so use a parameter free version to replace it
- Language model
 - Leave-one-out
 - EM algorithm

Translation-based language model [Xue et al. 2008]

- The retrieval model can benefit from incorporating knowledge in the formulation

$$p(w_i|d) = \frac{|dl|}{|dl| + \lambda} p_{mix}(w_i|d) + \frac{\lambda}{|dl| + \lambda} p(w_i|C)$$

$$p_{mix}(w_i|d) = (1 - \beta)p(w_i|d) + \beta \sum_{t \in d} p_{tr}(w_i|t)p(t|d)$$

- Translation matrix:

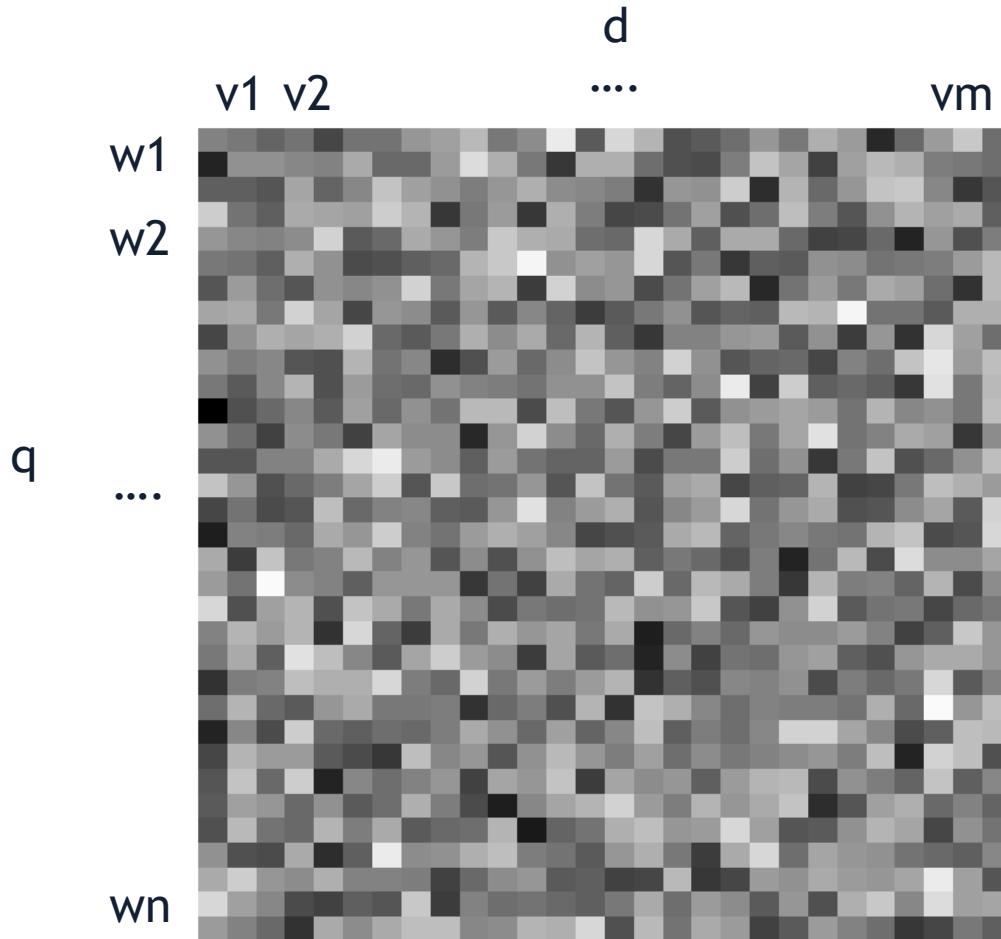
	Le	programme				
And		a	été	mis	en	application
the						
program						
has						
been						
implemented						

The table illustrates a translation matrix between English words (rows) and French words (columns). The English words are: And, the, program, has, been, implemented. The French words are: Le, programme, a, été, mis, en, application. The matrix shows which English words are translated into which French words. For example, 'And' is translated into 'Le', 'the' into 'a', 'program' into 'programme', 'has' into 'été', 'been' into 'mis', and 'implemented' into 'en application'. The matrix is sparse, with most cells being empty.

Discussion on query length

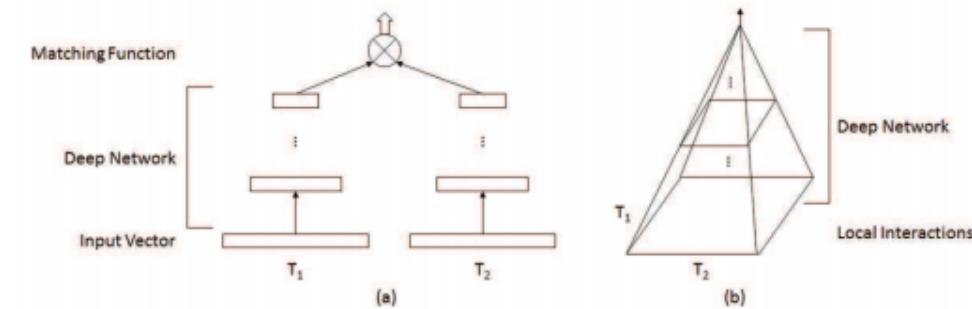
- What if the query is very long?
 - For example, the query is a paragraph or a document
 - The problem of retrieval is turned into a matching problem
 - i.e., semantic matching

Deep semantic matching [Pang et al. 2016]



each cell:

distributed representation of words (word2vec)



Homework 1

- Homework 1 is released in Canvas:
- Implementing TF-IDF and BM25 on the LinkSO dataset:
 - <https://sit.instructure.com/courses/44342/assignments/218604>