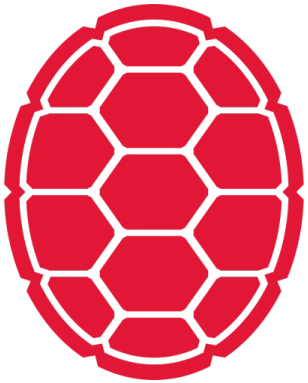


Applications

Computational Linguistics: Jordan Boyd-Graber
University of Maryland
INFORMATION RETRIEVAL

Slides adapted from Jimmy Lin

IR is worth a lot of money . . .

Prerequisites

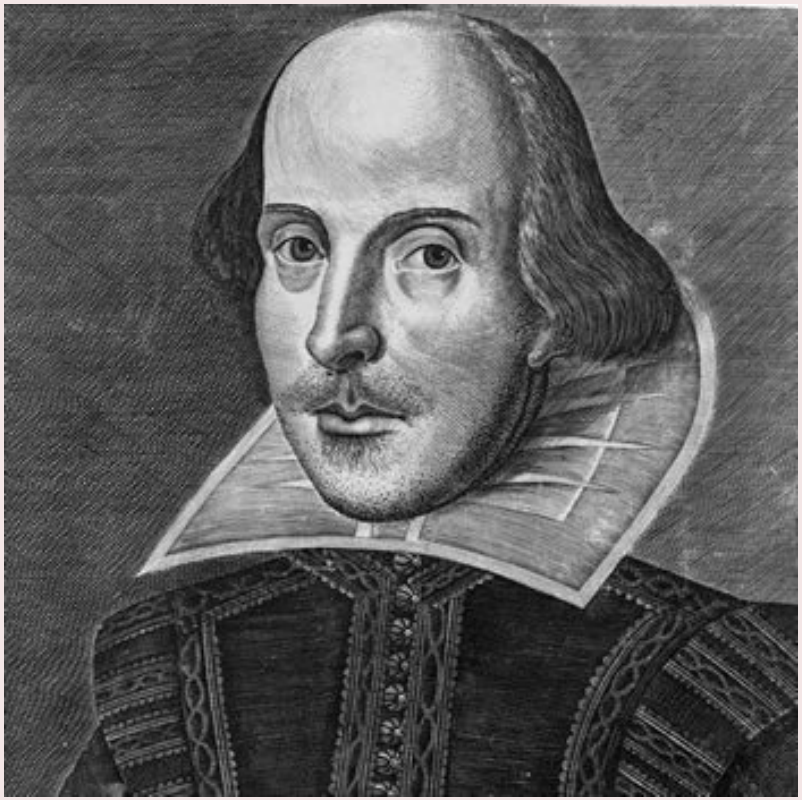
- Search a “collection” of **documents**
- Each document contains **terms** (words)
- Users create queries

General Problem

User

“tragic romance”

Author



“star-crossed lovers”

What’s a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

وقال مارك ريجيف - الناطق باسم
الخارجية الإسرائيلية - إن شارون قبل
الدعوة وسيقوم للمرة الأولى بزيارة
تونس، التي كانت لفترة طويلة المقر
الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام 2 .

Выступая в Мещанском суде Москвы экс-глава ЮКОСа
заявил не совершал ничего противозаконного, в чем
обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ़ीसदी विकास
दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処…アーミテージ前副長官提言

조재영 기자= 서울시는 25일 이명박 시장이 `행정중심복합도시" 건설안
에 대해 `군대라도 동원해 막고싶은 심정"이라고 말했다는 일부 언론의
보도를 부인했다.

We’ll talk more about this later . . .

What’s a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

وقال مارك ريجيف - الناطق باسم
الخارجية الإسرائيلية - إن شارون قبل
الدعوة وسيقوم للمرة الأولى بزيارة
تونس، التي كانت لفترة طويلة المقر
الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام 2 .

Выступая в Мещанском суде Москвы экс-глава ЮКОСа
заявил не совершал ничего противозаконного, в чем
обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ़ीसदी विकास
दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処…アーミテージ前副長官提言

조재영 기자= 서울시는 25일 이명박 시장이 `행정중심복합도시" 건설안
에 대해 `군대라도 동원해 막고싶은 심정"이라고 말했다는 일부 언론의
보도를 부인했다.

We’ll talk more about this later . . . assume we know the answer.

Bag of Words

McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

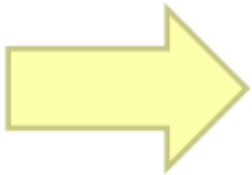
NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

...



"Bag of Words"

14 × McDonalds

12 × fat

11 × fries

8 × new

7 × french

6 × company, said, nutrition

5 × food, oil, percent, reduce, taste, Tuesday

...

Boolean Retrieval

- Users express queries as a Boolean expression
  - AND, OR, NOT
  - Can be arbitrarily nested
  - Retrieval is based on the notion of sets
- Any given query divides the collection into two sets: retrieved, not-retrieved
  - Pure Boolean systems do not define an ordering of the results

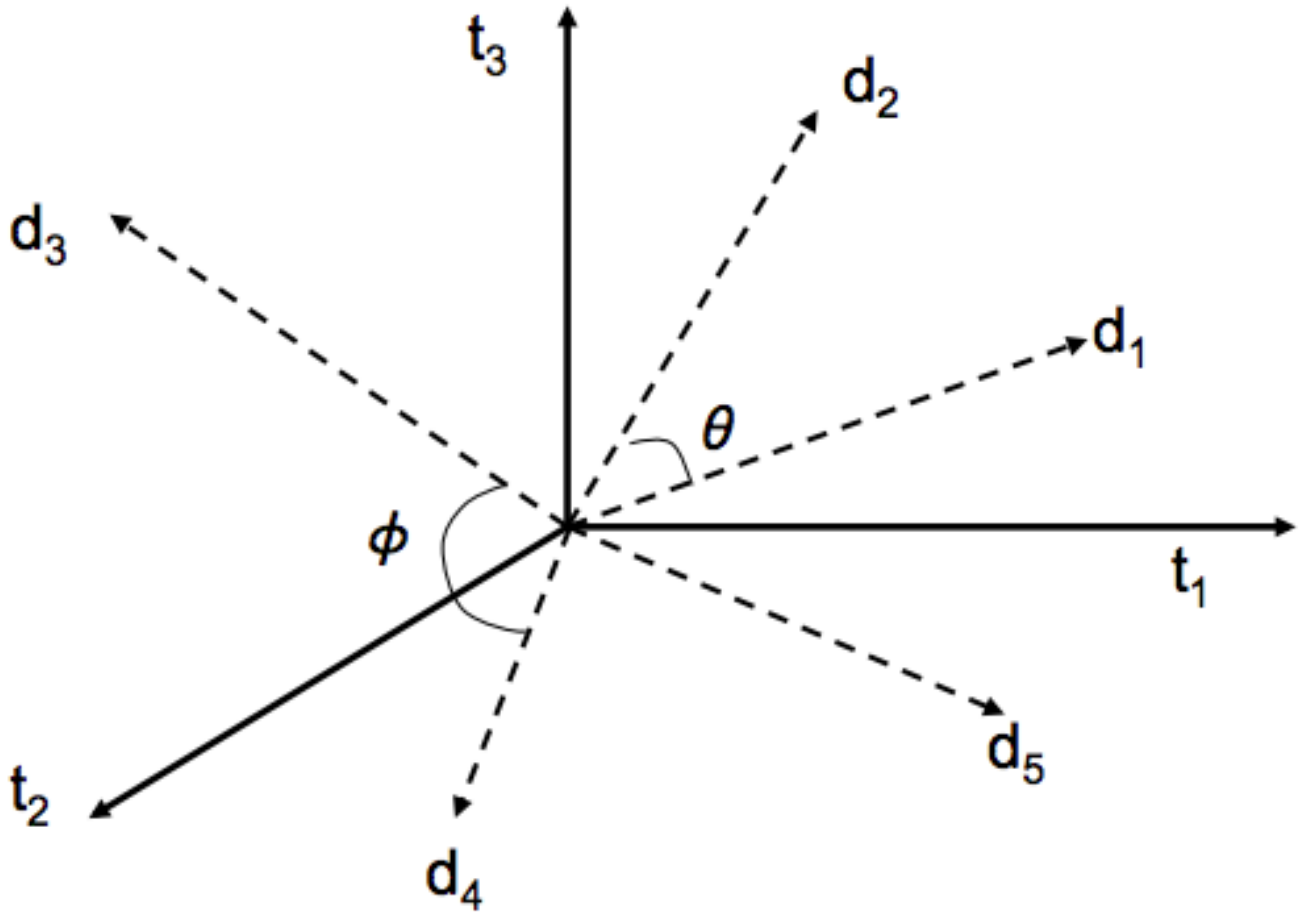
Strengths and Weaknesses

- Strengths
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you're looking for
  - Implementations are fast and efficient
- Weaknesses
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many hits or none
  - When do you stop reading? All documents in the result set are considered "equally good"
  - What about partial matches? Documents that "don't quite match" the query may be useful also

Ranked Retrieval

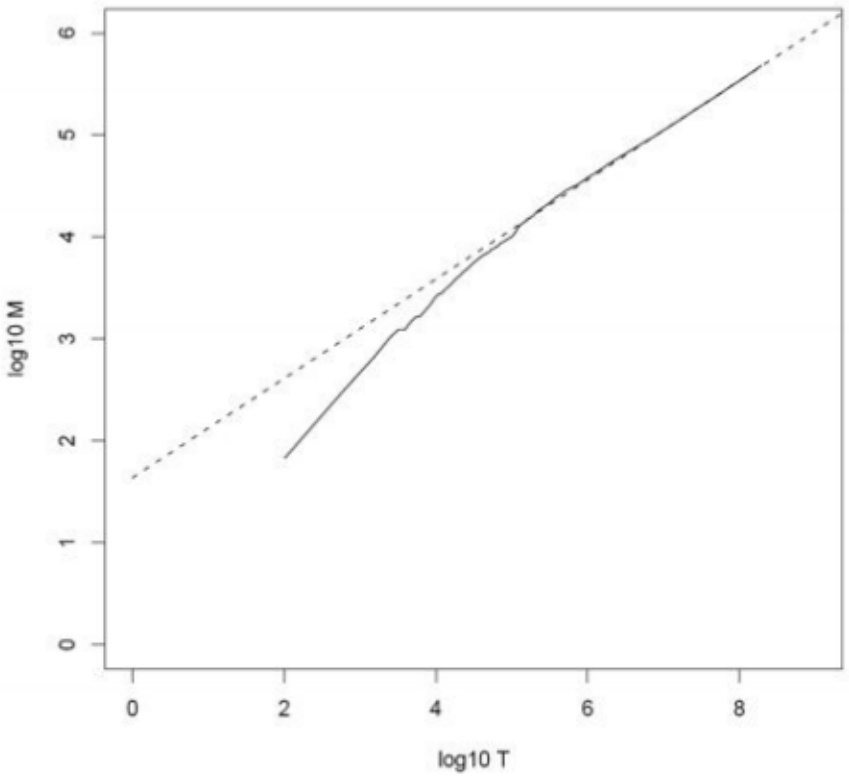
- Order documents by how likely they are to be relevant to the information need
  - Estimate relevance(q, d_i)
  - Sort documents by relevance
  - Display sorted results
- User model
  - Present hits one screen at a time, best results first
  - At any point, users can decide to stop looking
- How do we estimate relevance?
  - Assume document is relevant if it has a lot of query terms
  - Replace relevance with sim(q, d_i)
  - Compute similarity of vector representations

Representing documents



Each document is vector $d_i = \langle w_{i,1}, \ldots w_{i,V} \rangle$ (each word is dimension)

High-Dimensional Space (Heaps' Law)

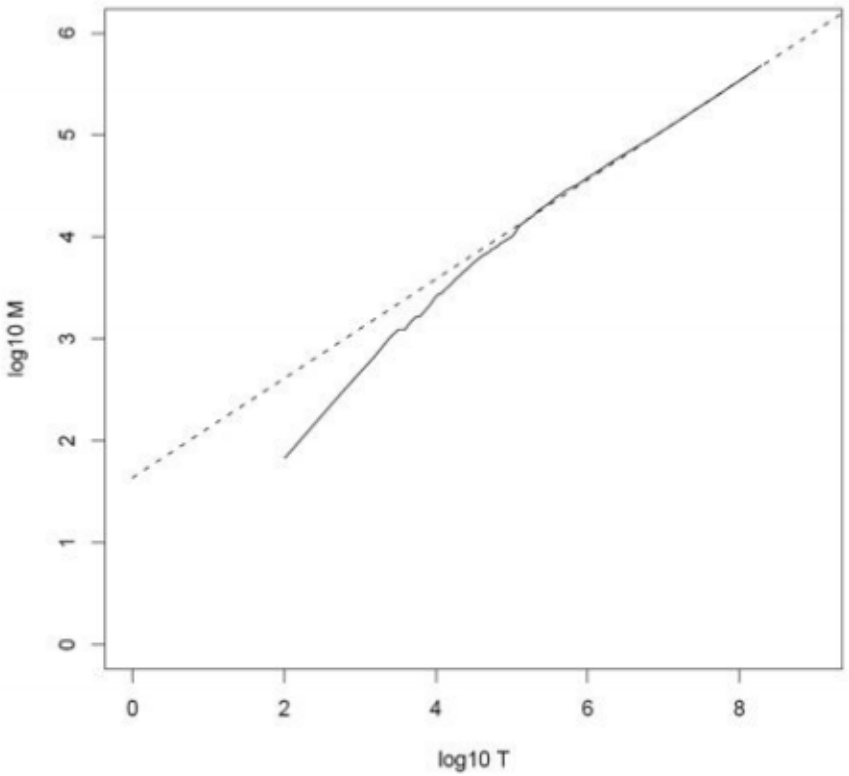


k = 44
b = 0.49

First 1,000,020 terms:
Predicted = 38,323
Actual = 38,365

$$V = kD^b \qquad (1)$$

High-Dimensional Space (Heaps' Law)



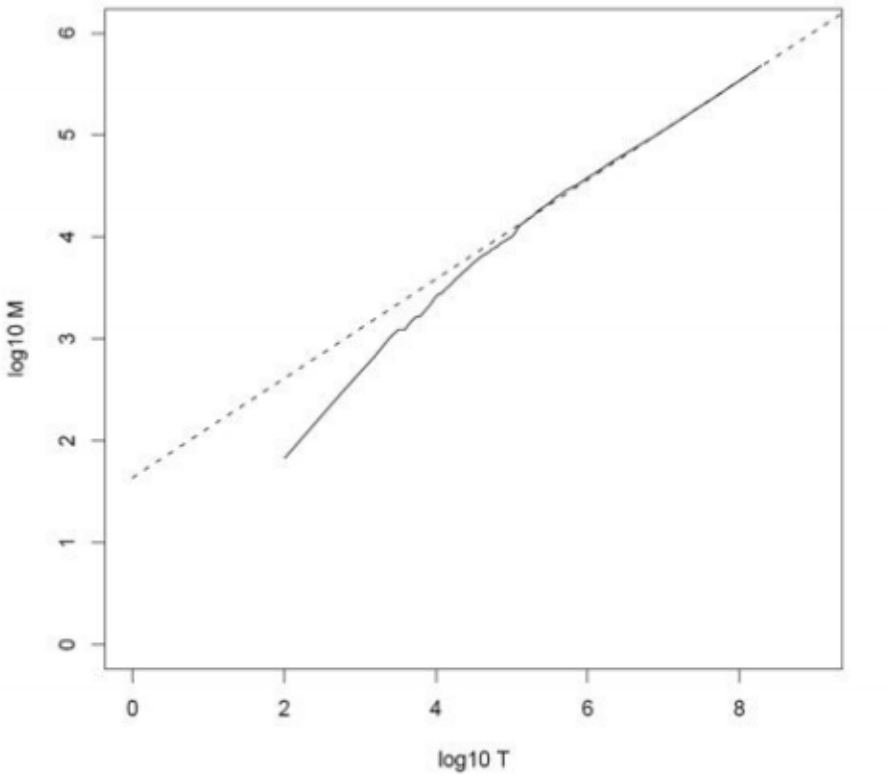
k = 44
b = 0.49

First 1,000,020 terms:
Predicted = 38,323
Actual = 38,365

$$V = kD^b \qquad (1)$$

Vocabulary size

High-Dimensional Space (Heaps' Law)



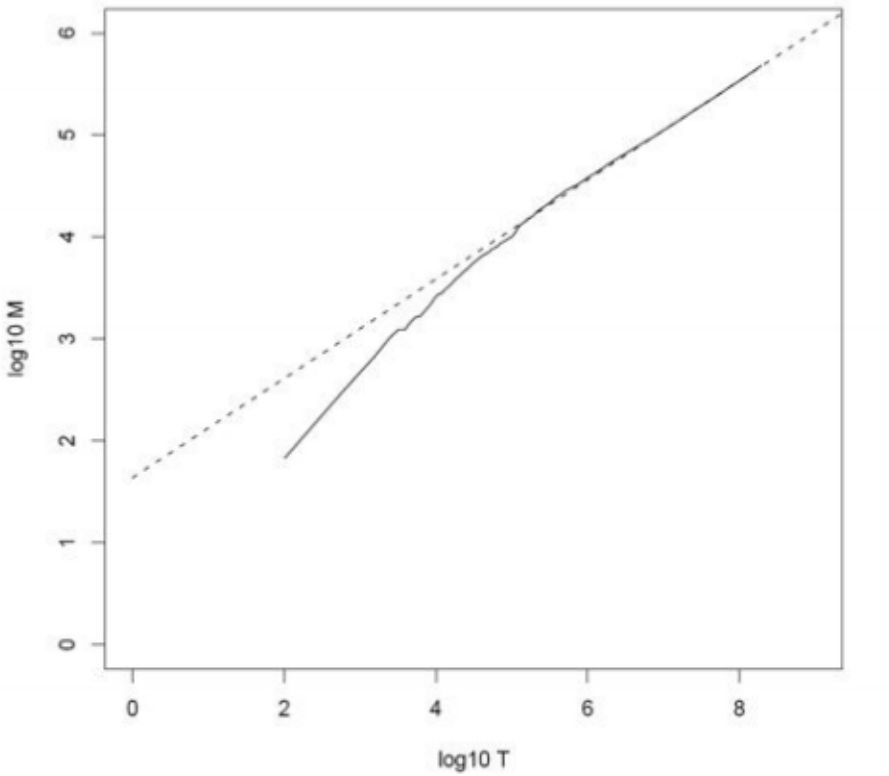
k = 44
b = 0.49

First 1,000,020 terms:
Predicted = 38,323
Actual = 38,365

$$V = kD^b \qquad (1)$$

Number of documents

High-Dimensional Space (Heaps' Law)



k = 44
b = 0.49

First 1,000,020 terms:
Predicted = 38,323
Actual = 38,365

$$V = kD^b \qquad (1)$$

Constants (per-language, type of document)

Intuitions

- Term weights consist of two components
  - Local: how important is the term in this document?
  - Global: how important is the term in the collection?
- Here's the intuition:
  - Terms that appear often in a document should get high weights
  - Terms that appear in many documents should get low weights
- How do we capture this mathematically?
  - Term frequency (local)
  - Inverse document frequency (global)

tf-idf Term Weighting

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \tag{2}$$

- Word i's weight in document j
- Frequency of word i in document j
- Help with interpretation

tf-idf Term Weighting

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \tag{2}$$

- Word i's weight in document j
- Frequency of word i in document j
- Help with interpretation
- Tension: prediction or interpretation

tf-idf Term Weighting

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \tag{2}$$

- Word i's weight in document j
- Frequency of word i in document j
- Help with interpretation
- Tension: prediction or interpretation

Frequency of Terms (Zipf's Law)

The most frequent words ("the") are everywhere but useless for queries.
The most useful words are relatively rare . . . but there are lots of them.

$$f_i = \frac{c}{i} \tag{3}$$

- The frequency of a word i is inversly proportional to
- The rank (in frequency) of word
- Scaled by a constant

Can't just throw out useless words

Frequency of Terms (Zipf's Law)

The most frequent words ("the") are everywhere but useless for queries.
The most useful words are relatively rare . . . but there are lots of them.

$$f_i = \frac{c}{i} \tag{3}$$

- The frequency of a word i is inversly proportional to
- The rank (in frequency) of word
- Scaled by a constant

Can't just throw out useless words

Frequency of Terms (Zipf's Law)

The most frequent words ("the") are everywhere but useless for queries.
The most useful words are relatively rare . . . but there are lots of them.

$$f_i = \frac{c}{i} \tag{3}$$

- The frequency of a word i is inversly proportional to
- The rank (in frequency) of word
- Scaled by a constant

Can't just throw out useless words

Frequency of Terms (Zipf's Law)

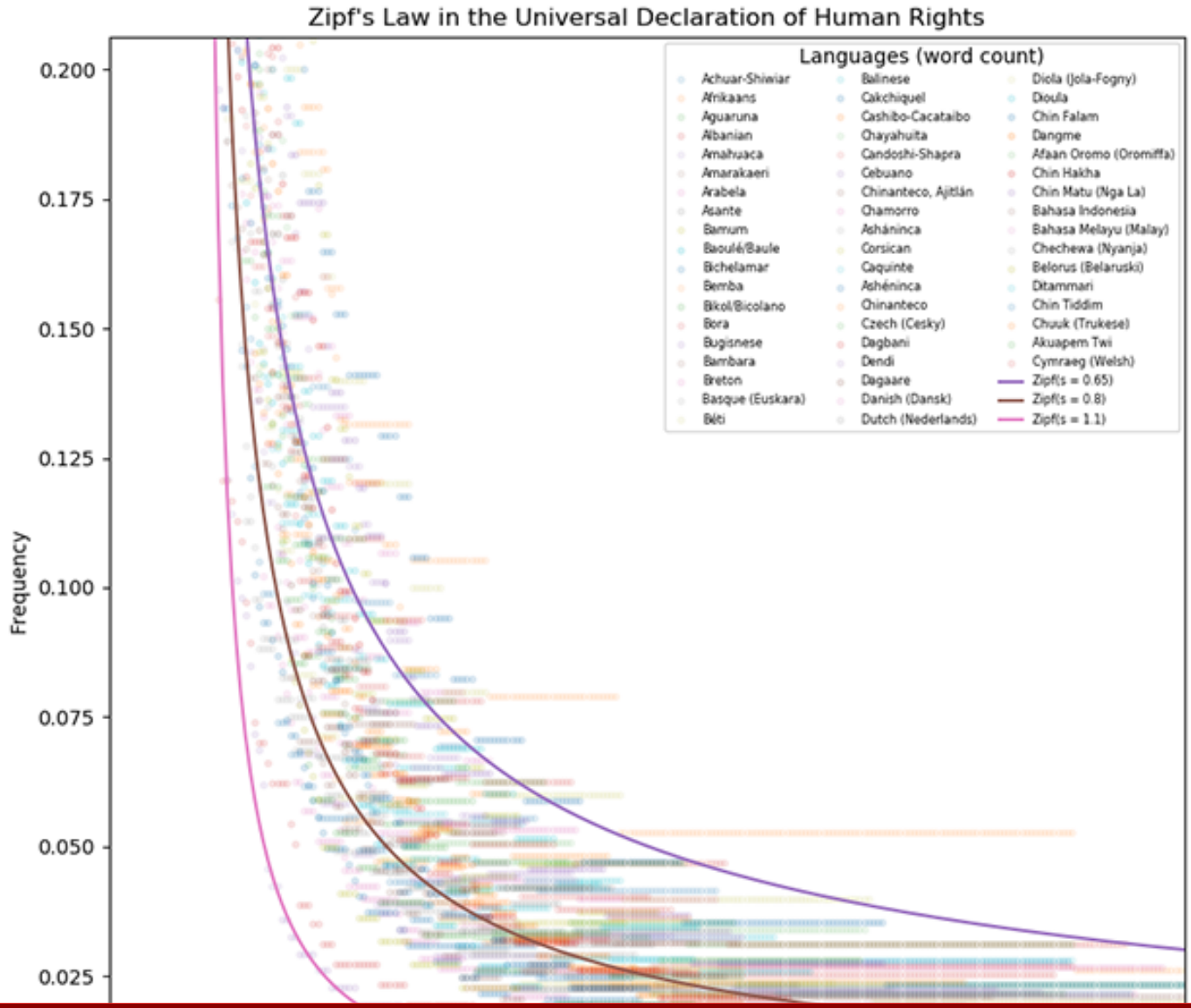
The most frequent words ("the") are everywhere but useless for queries.
The most useful words are relatively rare . . . but there are lots of them.

$$f_i = \frac{c}{i} \tag{3}$$

- The frequency of a word i is inversly proportional to
- The rank (in frequency) of word
- Scaled by a constant

Can't just throw out useless words

Zipf's Law


Zipf's Law in the Universal Declaration of Human Rights
Frequency
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
Languages (word count)
Achuar-Shiwiar
Afrikaans
Aguaruna
Albanian
Amahuaca
Amarakaeri
Arabela
Asante
Banum
Baoulé/Baule
Bichelamar
Bemba
Bikol/Bicolano
Bora
Bugisnese
Bambara
Breton
Basque (Euskara)
Béti
Balinese
Cakchiquel
Cashibo-Cacataibo
Chayahuita
Candoshi-Shapra
Cebuano
Chinanteco, Ajitlán
Chamorro
Asháninca
Corsican
Caquinte
Ashéninca
Chinanteco
Czech (Cesky)
Dagbani
Dendi
Dagaare
Danish (Dansk)
Dutch (Nederlands)
Diola (Jola-Fogny)
Dioula
Chin Falam
Dangme
Afaan Oromo (Oromiffa)
Chin Hakha
Chin Matu (Nga La)
Bahasa Indonesia
Bahasa Melayu (Malay)
Chechewa (Nyanja)
Belorus (Belaruski)
Ditammari
Chin Tiddim
Chuuk (Trukese)
Akuapem Twi
Cymraeg (Welsh)
Zipf(s = 0.65)
Zipf(s = 0.8)
Zipf(s = 1.1)

Similarity Metric

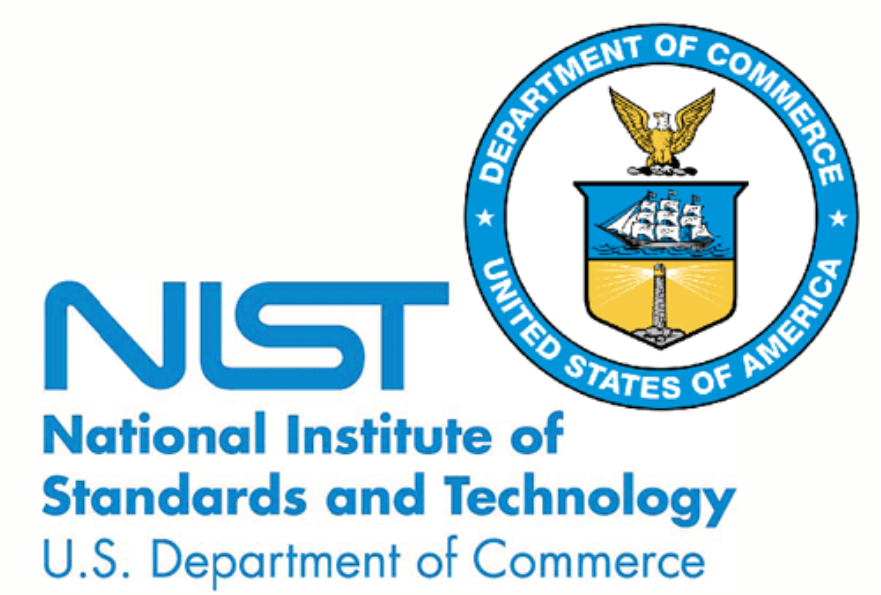
- "Angle" between vectors

$$\cos(\theta) = \frac{\vec{d}_j \cdot \vec{d}_k}{|\vec{d}_j||\vec{d}_k|} \tag{4}$$

- More generally, dot (inner) product . . . normalized vectors

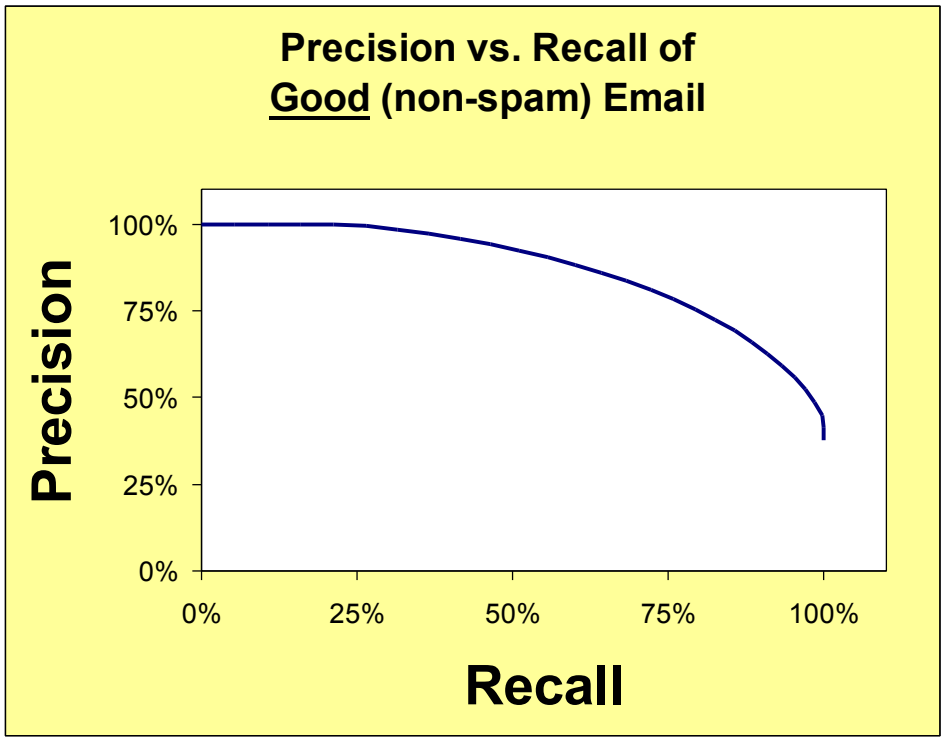
$$\text{sim}(d_j, d_k) = \vec{d}_j \cdot \vec{d}_k = \sum_{i=1}^{n} w_{i,j} w_{i,k} \tag{5}$$

Evaluating Results



- TREC: set of “gold” relevant documents
- How many of the documents found?
- Annual bake-off of IR systems

Precision and Recall



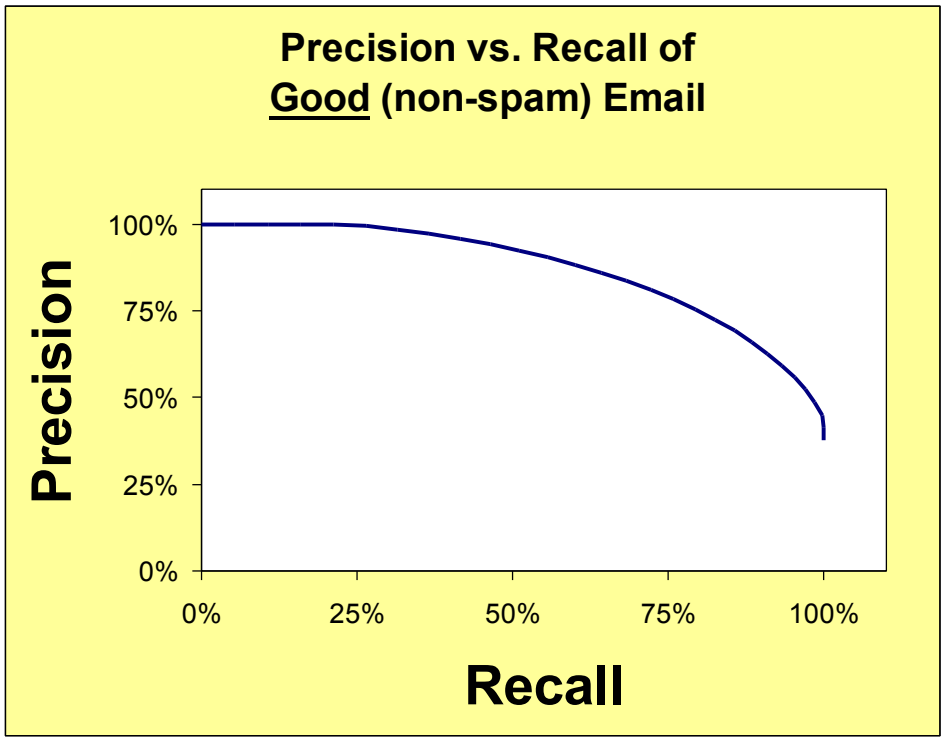
- Precision: Of what you returned, how much was right?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \tag{6}$$

- Recall: Of what could be right, how much did you find?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \tag{7}$$

Precision and Recall



- Precision: Of what you returned, how much was right?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \tag{6}$$

- Recall: Of what could be right, how much did you find?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \tag{7}$$

- F-measure: geometric mean