Problem- Definition





















.....

The control of t

Table Mark

Indexing





and constitution











rested ries



























Operation, femal 3-flat layurus en control stude.

Ranked Retrieval



See of the entire colors on the set of the entire colors of the ent







But and the second of the seco



least on the meanth's problem in man and man seri.

The control of the problem is the state of the control of the state o



Community of the control of the









Photo: Server-bosometric (2 p. c) is deduced in a reason reconstruction for period (declared in a reason reconstruction).

 Server a reason reconstruction reconstruction are reconstruction reconstruction.

 Server and reconstruction reconstruction.















Vector Space Model

























Problem- Definition





















Indexing



tions of control or proceeding to the particular of the control of



















heatof rike encuetas



Processor with the interpretation of the con-Comment and Comment and Commen

TO SECTION ASSESSMENT OF THE PARTY AND THE P







....

32: 20d

-













- HH









Ranked Retrieval











Feast or families not a grob em in maked repletal



























to the first matter as a contract of the state of the sta







Vector Space Model

is the officers wife scars.

Section du = $\sum_{n \in \mathbb{N}} \operatorname{ritief}_{\mathcal{F}}$









The particular behavior and a challed a service of the particular behavior and a challed a service of the particular and t

















Problem- Definition





Berry - [1 2 4 1 5 5 4 10 19

Dictionary - Desired DS

inverted Index

We need which we easily to be
 Could be retained on of some a count of some and to be retained to the count of some and to be retained to the count of some and to be retained to

The merce

1 2 2 5 6 10 21 5 6 CM

If the list length care a end y, the neage takes but via operations, smooth parange somethe docus.

Real Challenges

Indexing



Indexer stees: Fort









Longer phroves can be processed by breefing them down analyyado aktorar to amber into entry AVO antinually pelo AAD cale

Without the does, we cannot verify that the does metaring the above Bookern query do contain the phrase. Colon Encoded

Incidence vectors

So we have a Windon for each time.

Tour own except too the excise in Enric
George and Odjanes Jumples and #
Tour of All

descriptions dictionary 2 Postings



Issues for biword indexes Faire positives, as noted before
 Index blowup due to bigger distinctly
 modified a manufactulously, also as for

Answers to query

Assumy and Chappeline, Act III, Boune is Assumed to TARGET MATERIA. So, the articles was now for Date and processing a surprised of the Act and A

Indoxenstess: Sigtionary & Postings

Solution 2: Positional indexes in the pastings, case, for each term the position(a) in which tobers of it appear:

streng contact of the containing france dech maker I posted ... dech maker I posted ...

Bigger collections

Consider // - turifier documents, each with also in 1000 words.

sport with words.

• Any Style-shand including sports of purchaster

• Style-shad shadow made.

• Souther are \$4.10000 abstract remails among thans.

Where do we pay in storage?

Positional index example

*2:00.42; 6:7.18.20,72.90.20; 2:7.180; 6:17.20,20,20,40; 7:20.20;

But we now need to deal with more than jum equality

Processing a phrase query Force of inverted index confer for each distinct form (o. let, or, est. Marge their starposition from the counterate oil postupos with interior and to let.

Save gaveral method for products searched

Inverted index

test mentional contents.

Ideality and much so dealth, independent sold in males.

Seeker were well force to the arrays be madely for a

The index we just built

Can't build the matrix

Stoke (Winseline has half a trillion of born) Po.

But it has an enge that one billion its.

Query processing: AND Contains process as the purity
Green ARM Green

- So reveloped to a Contains

- So reveloped to

- So revelo

<u>ধিও হথার কার্</u>

What appears the world farmer resident to continue 1997

Proximity queries UMIT /3 STATUTE /3 FOREIAL /2 TORT = 5, 6 if or, A more 25 if a few dolf.

The second of th

Positional index size

 A positional infer expands portings Variage substantially ten haghintes anternament Neverth and a positional little is now standardly used because of the power and usefulness of shapes and promotive power and usefulness of shapes and promotive promotive and reaching rectional outside.

Inverted index construction



Intersection two postings lists

Positional index size

Seed as entry for each occurrence, not just once per document. react to depend on average detainents - Activity entry getting \$100 mms \$10 Mag, both, even some groupers and in \$10,000 mms

Find the special control of the special contr

Rules of thumb

Initial stages of text processing

Tobal standon

- Self-transfer requestors in second foreign

- contract County arrest year an expectation

Tourish Superforming the conserver
 Tourish Superforming the sections
 Tourish Superforming the sections
 Tourish Superforming
 Tourish Superforming
 Tourish
 Tourish

Phrase queries

Positional index size 95–50% of volume of

Countries this mass is a Tright for College upon

Indexer steps: Taken sequence



A first attempt: Biword indexes

indic every consecutive pain of terms in the sec-ecutions: For example the test "Friends, Formers, Countryment would governor the biscoris — pressurement country consistency

Each of these theoretis news of the survivors Two word piness query a occosing biness investigation.

Combination schemes

These two approaches our be professly combined. For particle graphs (Whomas Section), Walters Special first self-constraints promoting positions profit graphs American a William to the point in project of the production of the result of the time the "The Silve".

All the results and the value of the production of the project of the production of the

Information Retrieval (IR) - The Problem



Collection: A set of documents Assume it is a static collection for the moment

Goal: Retrieve documents with information that is relevant to the user's information need and helps the user complete a task

Searching is a Solved Problem!

Hash Table: O(1)

Binary Search Tree: O(log(n))

AVL Tree: O(log(n))

.

Why Do we need a different setup?



Size of the Data!

- 10 billion web pages
- Average size of webpage = 20KB
- 10 billion * 20KB = 200 TB
- Disk read bandwidth = 50 MB/sec
- Time to read = 4 million seconds = 46+ days
- Say there are M = 500K distinct terms among these.
- Even longer to do something useful with the data

Dictionary - Desired DS

A naïve dictionary

An array of struct:

term	document	pointer to
	frequency	postings list
а	656,265	\longrightarrow
aachen	65	\longrightarrow
zulu	221	\longrightarrow

```
char[20] int Postings *
20 bytes 4/8 bytes 4/8 bytes
```

- How do we store a dictionary in memory efficiently?
- How do we quickly look up elements at query time?

Dictionary - Desired DS

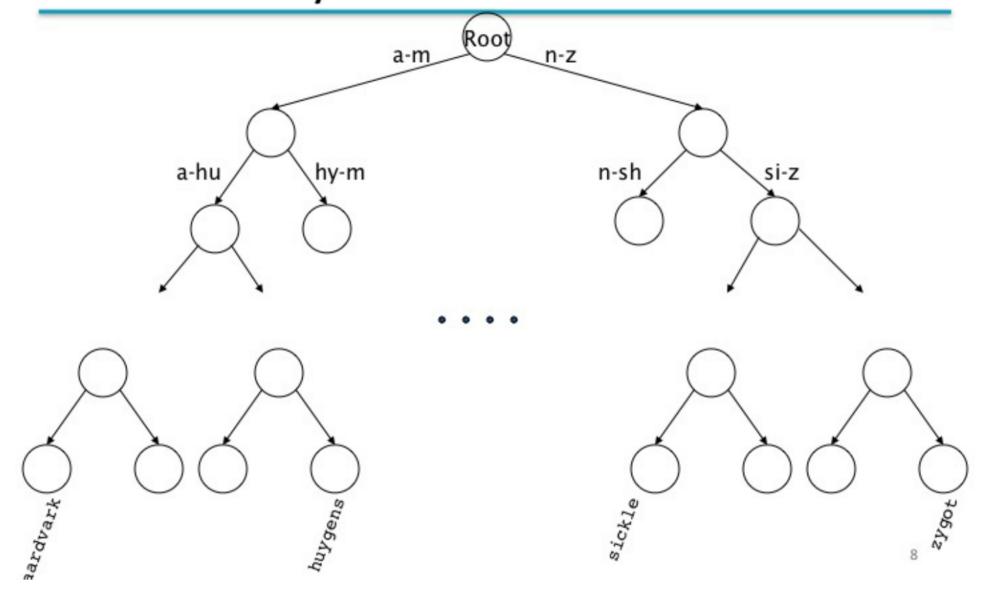
Dictionary data structures

- Two main choices:
 - Hashtables
 - Trees

Hashtables

- Each vocabulary term is hashed to an integer
 - (We assume you've seen hashtables before)
- Pros:
 - Lookup is faster than for a tree: O(1)
- Cons:
 - No easy way to find minor variants:
 - judgment/judgement
 - No prefix search [tolerant retrieval]
 - If vocabulary keeps growing, need to occasionally do the expensive operation of rehashing everything

Tree: binary tree



Trees

- Simplest: binary tree
- More usual: B-trees
- Trees require a standard ordering of characters and hence strings ... but we typically have one
- Pros:
 - Solves the prefix problem (terms starting with hyp)
- Cons:
 - Slower: O(log M) [and this requires balanced tree]
 - Rebalancing binary trees is expensive
 - But B-trees mitigate the rebalancing problem

Real Challenges

Out-of-Vocabulary Words
20-25% Named Entities
Multilingual Queries
Wild Card Queries
Spell Correction
Phrase Search
President of India

Problem- Definition





Incidence vectors

Cone. - 112141 8181818131...



Dictionary data structures









Indexing









La de Maintes









Solution 2: Positional indexes

mes illeway due to bigger dictionary – [stemiole for more than bilwords, placeword

-

Answers to query

Armony and C excepts. Act II Science appropriation is a sense as processed for many contains the sense of the sense o



dreft scottered positions of dreft scottered positions of dreft scottered positions of access

Bigger collections

Avg 6 bytes/word including straces/



2 3.145, 27.88, 275, 2.145, 275, 2.15

Can't build the matrix





Positional index example

- Detruct invested index works for each circlest terms to, by, or, not.

Processing a phrase query

Historica we process a curery? (= | cure.)

- Margerheir on yesittis ilm to enumerate all postant, with "to be or not to be".
- a Serve papers or short for a midwhy resurber

Inverted index

· For each term t, we must store a fail of all should be consisted that modes of dentify each doobly a doolb, a cocument surfail
 a cocument surfail COMP 1 2 4 5 5 1 1 2 2 1 5

what he specialistic world factor. In oblight or decreases (12)

Query processing: AND

Consider processing the overs: edes (MACeres) Long v Bedreit (He-Dictionally) - Retires as cotton Long Coper in the Dictionary

Proximity nueries

Heige, many is more first that executed "Charry, postular and indicator can be used for acting active, between the book common." Districts Adopt the linear manage of posting to handle providing queries. Charges make it works from an other of 20 — Table is little indicator connectly and afficiently when they of 20 or Md.

Inverted Index

we need to right to deep and right full.

- One right construction of coming it committees are con-w-22 lat 45100 South and training residue

Walk through the two postings a motor woods, in time them in the solution months of post respectives. 2 4 5 78 27 H TO MADE 1 - 7 - 2 8 - 2 - 11 - 21 - 14 Green the limiting the constraint of the marger rates October straints. generations. <u>Zn. conf.</u> positing reproved by doction.

Positional index size

A portional index expends porting: storage ent i bugli ir los un le mogresei. — encluded indices and companies.
 Move the ency a positional lind with once some only used because of the power and cooks house of those and provincially desired and other accordance on the provincial section.

Inverted Index construction

Aggree (SSS) Accesses 1

Positional index size

- Mexiconerin for each occurrence, not just come per document.

Index the depend on wavings document of just come per document.

- Out is just to just and continues.

- Out is just and continues.

- DOCK is just and continues of just and just an

Table startion

- Date, much expense this so situlates.

- convex rower, making several severe. recreation or Various and Superior Supe Statement g
 All majorith different forms of a month or maid. Company of the compan

who was the ballets in movempurities and as Proclem Salvester. On pitches The object of the process of the second Salvester is set to extend the second Salvester. On the second se

Initial stages of text processing

Combination schemes

Bules of thomb A positional index is 2—Las large as a non-positional index

- Casalo del Histolo de Paglob de ' Linguijos

These over approaches as a separative y constitute y constitute.

- The area the principal substitute of separative principal substitute of separative substitute of separative substitute of separative substitute substitu strongs first:

- Pres man on the places the Place bile?

- Will are set of \$1500) evaluate a manner soph shocker maner if when no a best of the place of the pla signs attacked mixed indexing adhered:

— tripped with guiley, material will determine in a of the dense that of gloring produced links.

— It is comes abilities in growth and have grapped and members grapped.

indexensteps: Token sequence

A first attempt: Biword indexes

 Index every consecutive pair of terms in the said ar a phrase

For vestriple iterited Triangle florrens,
Countrymer Twosic generate the classific

Countypeer Two a generate the awards — American waters are the common property. Cash of these bloods is now a dictionary term. Two washingtoness proceedings have branching.

2002

Ranked Retrieval

Ranked retrieval

- Publifier bur geer ist heven i beser Jobiern.

 Jegen interferende met werden. Final for a conflict rest of the whole and estimating of the formation of the Confliction.

 The good in age, to have a principle on a newly see an afficient in a poly to decrease an afficient in a poly to distant.

 Her good to their age to distant.

 Her good to their age to distant.

 Her good to their age to distant.

Term document count matrices

I a concept.

Devidence the courseter high recommoder.

Problem with Boolean search:

- itos em que las often recut in either ton few. (40) or too many (2000s) results. proper constitute (1900) (1900). — Olivin I. Taradinal van Aleksia (1904) (1900) (1900) — Olivin III (1900) (1900) (1900) (1900) (1900) — Olivin III
- intribed a lot of dell to converge with a query tradigmoducus a managed ble number of a list CHUs well to line, original to many

Eaglof words model

- Verbaling emobilizations from the throatering of son bits or married. Starts patient than Many and Many Symposium Start.
- This is saided the larged secrets manel.

 In proceed, the proceed bank. The problems linder was able to desiring the false two does not set.

 As all lead to Possoving positional bits market has a

Elementaria se colido una monativi ne aplany socialistici, in mindingram monito, the system manus and camp participatory. Scarment in the contration of the best of special section with the contration of the best of special section. First of the contration of the c

Term frequency tf

- The bear region of the district or over each delected. The make of these techniques and
- Interest. The refer of these betterminest.
 "Consultations of the manufacture are determined as the determinest are determined as the determinest."
 Consultation report upon after the same and account of the determinest. The determinest are determined as the determinest and account of the determinest.
 Consultation is a manufacture.
 Consultation is the determined as the determinest.
 Consultation is the determined as the determinest.

Banked retrieval models

- Processor the resulting absorbing we dis-
 - Log frequency weighting • The instrument weight of control dis-

ranked retrieval

- indepet the size of the result see it not up last a

$$\begin{split} & = \sum_{i=1}^{n-1} \left\{ \begin{array}{ll} 1 + \log_{2} L_{i,i} & \text{in } n_{i,i+1} \\ 0 & \text{otherwise} \\ 1 + 0 + O_{i} L + 1, 2 + 1, 2, 12 + 1, 1202 + 1, 902 \\ 0 & \text{over } 1 + 1, n_{i} \leq n_{i} \leq n_{i} \leq n_{i} \leq n_{i} \\ 0 & \text{over } 1 \leq n_{i} \leq n_{i} \leq n_{i} \leq n_{i} \\ 0 & \text{over } 1 \leq n_{i} \leq n_{i} \leq n_{i} \\ 0 & \text{over } 1 \leq n_{i} \leq n_{i} \leq n_{i} \\ 0 & \text{over } 1 \leq n_{i} \leq n_{i} \leq n_{i} \\ 0 & \text{over } 1 \leq n_{i} \\ 0 & \text{over } 1 \leq n_{i} \\ 0 &$$

Feast or famine: not a problem in Soprime as the back of ranked retrieval.

- We wish to return in order the documents most floak to be useful to the searcher
 How can we ranh-order the documents in the callection with respect to a query?
- Assign a score say in [0, 1] to each

This concurred was now well dearment and more Treatment

Document frequency Para terms are more informative than frequent

Adoption on pontaining this term is was likely to be relevant to the query accelerosereric + We want a high weight for sometimes for sometimes.

Query document matching scores

- We need a way of assigning a score to a guery/document pair. Let's dust with a one-form query.
 If the guern term does not occur in the dupon sec, some should see!

Document frequency, continued

- regard lines are handle solved as sections. Or who is a major and of high price in 1 millioning to the common and of high price in 1 millioning to the common and the common section to exact the solution and control to exact the solution and control to the common and control of the common and control of the common and control or the common and control and local lines are solved and local lines and the control lines are of the control lines and the control lines are solved to th

Take 1: Jaccard coefficient

- scholand Bird Circumbac Buch.

 jacond Rick = [A n 2] / [A U 6]

 jacond Rick = [A n 2] / [A U 6]

 jacond Rick = [A n 2] / [A U 6]

 a and Siconbhave to be the same accession as a color parameter accesses the

idf weight

- d(x) the <u>document</u> frequency of tithe number of the automorphisms and the majoral processing of the hand constable ct. at.) endelt sector is l'Environdous mest fors across of the $idt_i = \log_{10}(iWdf_i)$
 - Martine (1994) by the first of the proof of a feet of the Witness Committee (Series) and series

Jaccard coefficient: Scoring example

- Query this of years
- Department 1: conser also by control

idf example, suppose N = 1 million

$idE = log_{ac}(AVdE)$

issues with Jaccard for scoring

Expendit consider a monthly considerable of the state of

Effect of idf on ranking Open the part to open a find an order from the partie, the index of the partie of the parties of the parti

Becal: Binary term-document

Behidecarriery is represented by a binary vector = 10.115

- Collection vs. Document frequency
 - Which word is a feater search term (and should get a higher weight?)

Index Creation

Query Terms



Antony Brutus	ny and Cleopatra 1 1	Julius Caesar 1	The Tempest	Hamlet 0	Othello 0	Macbeth	Docs
Brutus	1	1	0	0	0	1	2000
	1	4			-	0.5	
Cassar			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	

Brutus AND Caesar BUT NOT Calpurnia 1 if play contains word, 0 otherwise

Incidence vectors

- So we have a 0/1 vector for each term.
- To answer query: take the vectors for Brutus,
 Caesar and Calpurnia (complemented) →
 bitwise AND.
 - 110100 AND
 - 110111 *ΑΝΓ*
 - **-** 101111 =
 - -100100

	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

Sec. 1.1

Answers to query

Antony and Cleopatra, Act III, Scene ii

Agrippa [Aside to DOMITIUS ENOBARBUS]: Why, Enobarbus,
When Antony found Julius *Caesar* dead,
He cried almost to roaring; and he wept
When at Philippi he found *Brutus* slain.

Hamlet, Act III, Scene ii

Lord Polonius: I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.



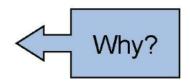
Sec. 1.1

Bigger collections

- Consider N = 1 million documents, each with about 1000 words.
- Avg 6 bytes/word including spaces/ punctuation
 - 6GB of data in the documents.
- Say there are M = 500K distinct terms among these.

Can't build the matrix

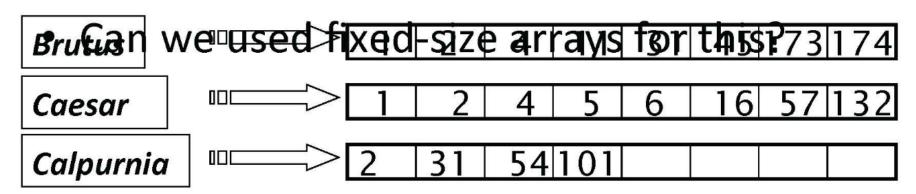
 500K x 1M matrix has half-a-trillion 0's and 1's.



- But it has no more than one billion 1's.
 - matrix is extremely sparse.
- What's a better representation?
 - We only record the 1 positions.

Inverted index

- For each term t, we must store a list of all documents that contain t.
 - Identify each doc by a docID, a document serial number



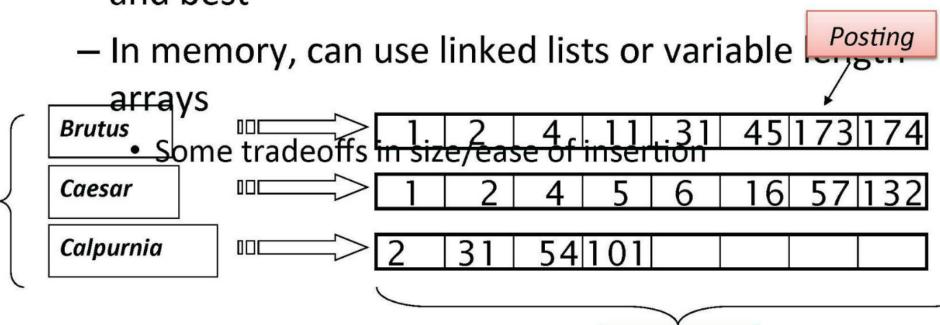
What happens if the word *Caesar* is added to document 14?

Inverted index

We need variable-size postings lists

Dictionary

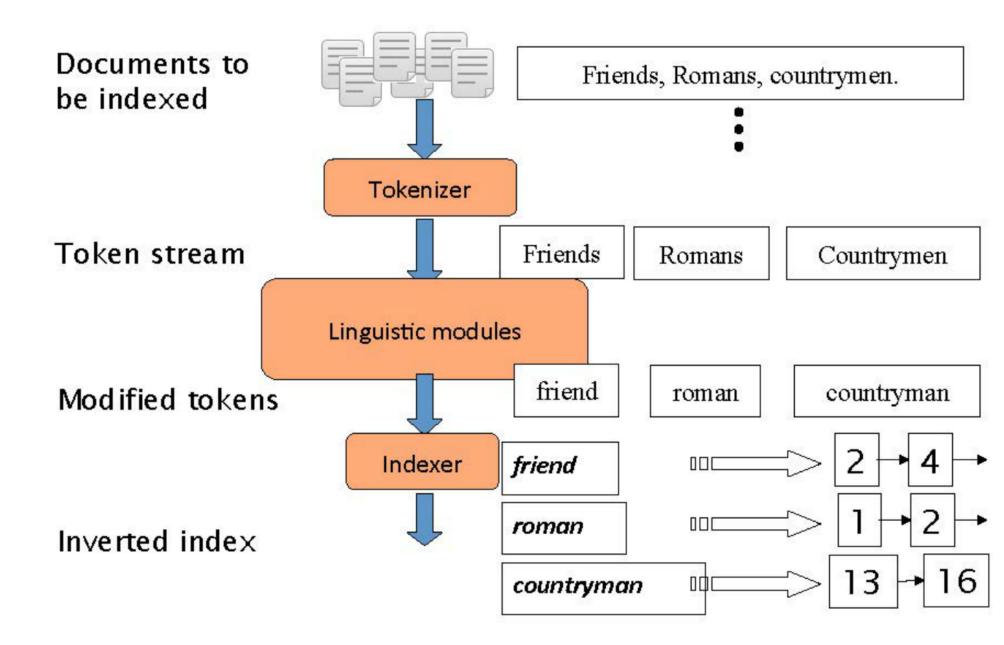
 On disk, a continuous run of postings is normal and best



Sorted by docID (more later on why).

Postings

Inverted index construction



Initial stages of text processing

- Tokenization
 - Cut character sequence into word tokens
 - Deal with "John's", a state-of-the-art solution
- Normalization
 - Map text and query term to same form
 - You want U.S.A. and USA to match
- Stemming
 - We may wish different forms of a root to match
 - authorize, authorization
- Stop words
 - We may omit very common words (or not)
 - · the, a, to, of

Indexer steps: Token sequence

Sequence of (Modified token, Document ID) pairs.

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me. Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	docID
1	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
was	2
ambitious	2

Sec. 1.2

Indexer steps: Sort

- Sort by terms
 - And then docID



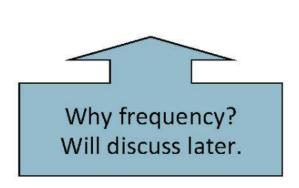
Term	docID
1	1
did	1
enact	1
julius	1
caesar	1
1	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
was	2
ambitious	2

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	2 2 1 2 1 1 2 2 2 1
enact	1
hath	1
1	1
1	1
i	1
it	1 1 1
julius	1
killed	1
killed	1
let	2 1 2 2 1 2 2 2 2 1 2 2 2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

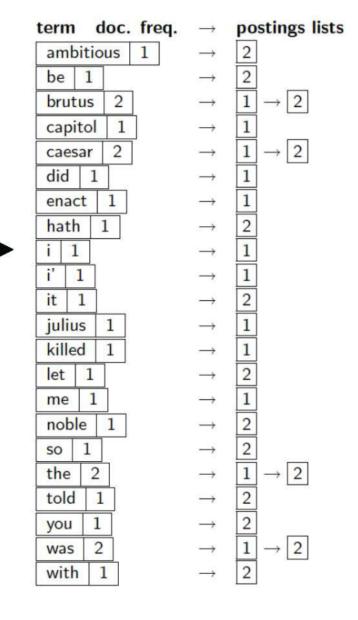
Sec. 1.2

Indexer steps: Dictionary & Postings

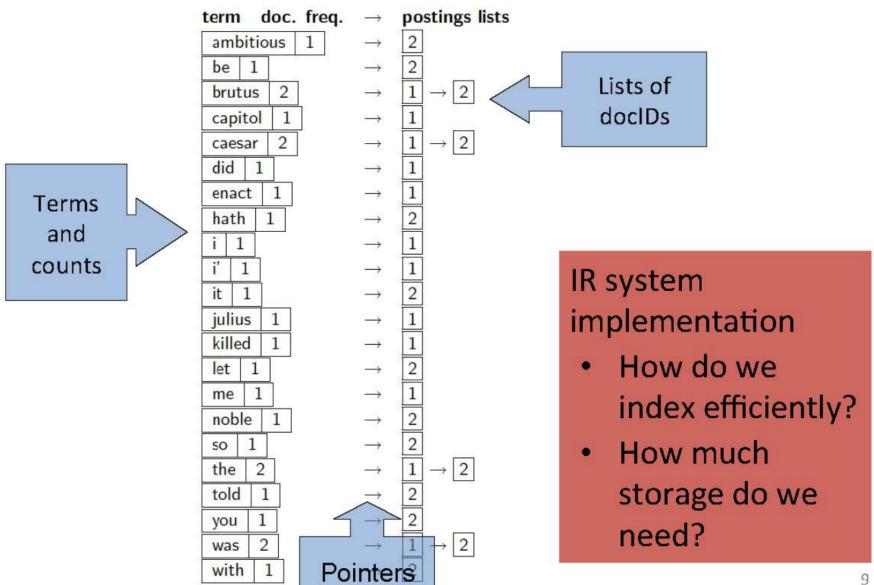
- Multiple term entries in a single document are merged.
- Split into Dictionary and Postings
- Doc. frequency information is added.



Term	docID
ambitious	2
be	2
brutus	2 2 1 2 1
brutus	2
capitol	1
caesar	1 2 2 1
caesar	2
caesar	2
did	
enact	1
hath	1
I I	1
ı	1
î'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2 2 1 2 2 2 1 1 2
	2



Where do we pay in storage?





The index we just built

How do we process a query?



– Later - what kinds of queries can we process?

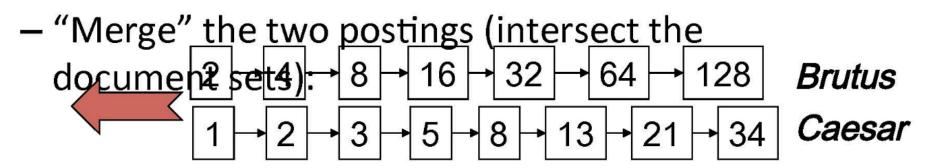
Sec. 1.3

Query processing: AND

Consider processing the query:

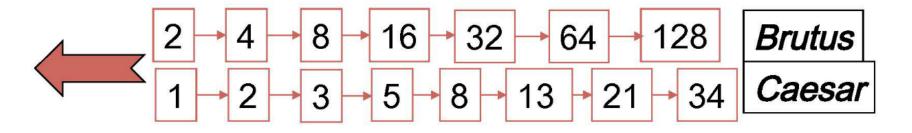
Brutus AND Caesar

- Locate Brutus in the Dictionary;
 - Retrieve its postings.
- Locate Caesar in the Dictionary;
 - · Retrieve its postings.



The merge

 Walk through the two postings simultaneously, in time linear in the total number of postings entries



If the list lengths are x and y, the merge takes O(x+y) operations.

<u>Crucial</u>: postings sorted by docID.

Phrase queries

- We want to be able to answer queries such as "stanford university" – as a phrase
- Thus the sentence "I went to university at Stanford" is not a match.
 - The concept of phrase queries has proven easily understood by users; one of the few "advanced search" ideas that works
 - Many more queries are implicit phrase queries
- For this, it no longer suffices to store only
 <term : docs> entries

A first attempt: Biword indexes

- Index every consecutive pair of terms in the text as a phrase
- For example the text "Friends, Romans,
 Countrymen" would generate the biwords
 - friends romans
 - romans countrymen
- Each of these biwords is now a dictionary term
- Two-word phrase query-processing is now immediate.

Longer phrase queries

- Longer phrases can be processed by breaking them down
- stanford university palo alto can be broken into the Boolean query on biwords:

stanford university AND university palo AND palo alto

Without the docs, we cannot verify that the docs matching the above Boolean query do contain the phrase.

Can have false positives!

Issues for biword indexes

- False positives, as noted before
- Index blowup due to bigger dictionary
 - Infeasible for more than biwords, big even for them

 Biword indexes are not the standard solution (for all biwords) but can be part of a compound strategy

Solution 2: Positional indexes

 In the postings, store, for each term the position(s) in which tokens of it appear:

```
<term, number of docs containing term; doc1: position1, position2 ...; doc2: position1, position2 ...; etc.>
```

Positional index example

```
<be: 993427;
1: 7, 18, 33, 72, 86, 231;
2: 3, 149;
4: 17, 191, 291, 430, 434;
5: 363, 367, ...>
Which of docs 1,2,4,5
could contain "to be
or not to be"?
```

- For phrase queries, we use a merge algorithm recursively at the document level
- But we now need to deal with more than just equality

Processing a phrase query

- Extract inverted index entries for each distinct term: to, be, or, not.
- Merge their doc:position lists to enumerate all positions with "to be or not to be".
 - to:
 - 2:1,17,74,222,551; 4:8,16,190,429,433; 7:13,23,191; ...
 - *− be:*
 - 1:17,19; 4:17,191,291,430,434; 5:14,19,101; ...
- Same general method for proximity searches

Sec. 2.4.2

Proximity queries

- LIMIT! /3 STATUTE /3 FEDERAL /2 TORT
 - Again, here, /k means "within k words of".
- Clearly, positional indexes can be used for such queries; biword indexes cannot.
- Exercise: Adapt the linear merge of postings to handle proximity queries. Can you make it work for any value of k?
 - This is a little tricky to do correctly and efficiently
 - See Figure 2.12 of IIR

Positional index size

- A positional index expands postings storage substantially
 - Even though indices can be compressed
- Nevertheless, a positional index is now standardly used because of the power and usefulness of phrase and proximity queries ... whether used explicitly or implicitly in a ranking retrieval system.

Sec. 2.4.2

Positional index size

- Need an entry for each occurrence, not just once per document
- Index size depends on average document size
 - Average web page has <1000 terms
 - SEC filings, books, even some epic poems ... easily 100,000 terms

Consider a term with frequency 0.1%

Document size	Postings	Positional postings
1000	1	1
100,000	1	100

Rules of thumb

 A positional index is 2–4 as large as a nonpositional index

Positional index size 35–50% of volume of original text

 Caveat: all of this holds for "English-like" languages

Sec. 2.4.3

Combination schemes

- These two approaches can be profitably combined
 - For particular phrases ("Michael Jackson", "Britney Spears") it is inefficient to keep on merging positional postings lists
 - Even more so for phrases like "The Who"
- Williams et al. (2004) evaluate a more sophisticated mixed indexing scheme
 - A typical web query mixture was executed in ¼ of the time of using just a positional index
 - It required 26% more space than having a positional index alone



Independence Sort







LENGUAND - LENGUAND - LENGUANI - LENGUANI - LENGUANI

Incidence vectors

Independence Temporary & Purchago

- confind unknowly pole also can be unken into the sholes a greey on beautiful storytard anthorally ATO anthorally polo ATO polo etto
- WWW.

Longer phrase queries

- Longer phrases can be processed by breaking their states

Issues for biword indexes

Solution 2: Positional indexes

Antony and Cleopatra, Ad. III. Scene

Answers to query

Proper steps Distribuses & frestings

- Consider # = 1 million decuments, each with about 1000 words. Aug o hytes/word in cading spaces/ - 60% of data in the documents
- Sey there are Min 5000 shring terms empre

Bigger collections

- We only record the Local domain

Where do we pay in storage? The index we just built



Positional index example

- *** 58-55 5-1.18; 57-78; 8-731; 5-1.18; 34; 51; 62; 54; 5-17; 51; 31; 51; 62; 54; 5-28; 35;...;
- Est we now good to deal with more than policyability

Can't build the matrix

SCCR a 2M metals have belt-a-tail for Course

Processing a phrase query

- Extract inverted index entries for each circlest term (a), 66, 65, 466. Merger their decaptorities like to enumerate all positions with "to be arrow to be".
- Same general method for provincits searches
- The good the transport of the state of the

Inverted index

For each term is we must along a flat of all documents's extremium to

An obtacles inthe year bearing second ordered and

Query processing: AND

Proximity queries

- Ajiri, heng, si mai Malikah semilah P. Okari, probasal kabana ana ba postili sada garisat dinastri kabasa sarrasi.
 Povelika Adapatha Pranti mengalat paratrasi te hadi a pranti niya paratrasi te kada pranti niya paratrasi te waki karasy palasi 165 1 na malika te aban daharansat yang atkarati dan Republika Na Maratrasi 165 1 na malika te aban daharansat yang atkarati dan Republika Na Maratrasi 165 dan Republika Na Maratrasi 165 dan Republika Na Maratrasi 165 dan Republika Na Maratrasi Na Maratr

Inverted Index

the stranger of the front bear which in see the design of the control of

The merge



ist fractic are cand vittle mence siner blood

Positional index size

Take 1: Jaccard coefficient

idf weight

denotes the traction of the left matter confit $-4L_{\rm B}N$ Where the of this are determined by $-3L_{\rm B}N$ where $-3L_{\rm B}N$

We use logistic "gine seed of tale", to "compan" the direct of till.

- Apositional ladeo expunsis postings storage substantially
 - Even though indices can be compressed New trades: A partitional lightest new standardy used occurs of the power and such above in process of process (quartics whether used exploting or and as the silv missing retrieval section.

Inverted index construction.



- Intersecting two postings lists (a "merge" algorithm)

Positional index size

Index also depends on average obcurrent digest

Perspecting the SER in the SER SERVICE CONTROL OF S

Phrase queries

Bules of thumb

A positional index is 2—1 as targe as a non-positional Index

Positions index size 35–905 of value of Coversion of this notes for "English-like" in course

Initial stages of text processing

Taken in the
— Collaborates are seen for a result for ex— Collaborates are seen for a result for ex— Collaborate Collaborate Address Seeding
— respective and collaborate collaborates
— in a result Collaborate Collaborates
— result Collaborates Collaborates
— result Collaborates
— result Collaborates
— results — results

North and age
 The arts, but the disease the new of accost course sock
 The arts of the authorized

Stop secret
 Here are confluence we continued
 Here are of

- for example the test "Frients, Borners, Countrymer" would generate the biwords distributions areas consignate Each of these bloods is now a dictionary term
- We want to be do a no arriver cale from an in-tradepolar effection?
 If a principal cale from the arriver is to a size of the arriver.
 See See Time of the arriver is seen as a size of the arri

Combination schemes

These beospares has an importantly continued From which is all news "Although the ", "Burney of the Train of the Tr

Indexensteps: Token sequence

A first attempt: Biword indexes

production of secretary 2 deposit very figuration products and transaction consistency of the figuration of the figurati

Recall: Binary term-documen

Incidence match

Note to proper a proposed as less transport of \$10.00

Collection vs. Document frequency

Ranked Retrieval

Ranked retrieval

- Good for experience with precise a scientist of global medical of the constitute.

 Moreover, the conductions Applications carress to constitute the Constitute of the consti

Term-document count matrices

Bag of words model

- Tables or dishippid morbineds:
 In a seria, the hardplace of the serious addo one
 wife in deling and there has the more to

 the moduli freezers (from each in an advertible)

 or

- Problem with Boolean search:
- icas; Of Familie

 Foolkiss is delevable analytic other too feet
 (40) or non-mary (1000) medits.

 Comy 1: Medical secretary ACC (1014)

 Comy 2: Medical secretary ACC (1014)

 Comy 3: Medical secretary ACC (1014)

 Comy 3: Medical secretary ACC (1014)

- AHD also see few OE shartes many

Ranked retrieval models

The end grants father than a carry inguigated can under a report on the reconstruction of ad and a manuscript of the section of the section

Term frequency tf

- The description of the descripti

Log frequency weighting

 $w = (1 + \log(t, t) \times \log_{10}(N/dt^2)$

destinative enging stance of demands and - Note that for the days scales in the second

Feast or famine: not a problem in

The sport of the s

Antign a occurs—say in [0, 1]—10 each document

Scoring as the basis of ranked retrieval

- Consider a term in the query that is nare in the

 $Score(q,d) = \sum_{a \in A} tf idf_{ad}$

- Document frequency

to the skill with a larger name in the state with a second restriction of the scalar restriction of the scalar resource about 5 he 0. The more incount the query term in the scalar resource product but scalar resource product but were will be to an an under of alternatives for the scalar resource.

Query-document matching scores

- Document frequency, continued
- Proposition in such as defining the flow with the S Development on the the foreign in the collection to such the proposition.

To such the recover, they is do controlled to the property of the controlled to the

Such document is now represented by time, valued vector of differ weight; $\sim 30^\circ$

Jecombioefficient: Scoring example

- What is the query document match score that the lecture coefficient computes for each of the local documents below?
- Guery Mexic/vectors

 Document I: coescraint is resurb.

idf example, suppose N = 1 million



Issues with Jaccard for scoring

I describ som som krim jungsom gillann som, finnka som som krim stadensmill, finn finnsk som krim stadensk serika finn som krim som krim stadensk serika – americk som krim stadensmille – americk som krim som stadensk som finnsk som krim som krim som krim som finnsk som krim som krim

- Effect of lef on ranking Question Book of have an effect on ranking for one form montes. He

Vector Space Model

Comments as vectors

Quer or se vactors

College of the processor of the college of the coll



3 documents oxymple contd.

Forms bing vector space and drafty



the argie instead of distance

from angles to cost res The following runs actions are applications. - Best communications as a special communication of a communication of the special communication of the special communication of the special communication of the special communications of the special Outer to an endowind place of plant or in the billion (CST 201)





Length committee Experience of the property of the control of t

colleguery/document) A constitution of the second or 17 stepates tighed better of ... Coring for length-normalized vectors Contingent occurrency of the production of amplication does not produce or continued to amplication does not produce or continued to a production of the continued or continued to contin complete-year - ∑_year

Ranked retrieval

- Thus far, our queries have all been Boolean.
 - Documents either match or don't.
- Good for expert users with precise understanding of their needs and the collection.
 - Also good for applications: Applications can easily consume 1000s of results.
- Not good for the majority of users.
 - Most users incapable of writing Boolean queries (or they are, but they think it's too much work).
 - Most users don't want to wade through 1000s of results.
 - This is particularly true of web search.

Problem with Boolean search: feast or famine

- Boolean queries often result in either too few (≈0) or too many (1000s) results.
 - Query 1: "standard user dlink 650" \rightarrow 200,000 hits
 - Query 2: "standard user dlink 650 no card found"
 → 0 hits
- It takes a lot of skill to come up with a query that produces a manageable number of hits.
 - AND gives too few; OR gives too many

Ranked retrieval models

- Rather than a set of documents satisfying a query expression, in ranked retrieval models, the system returns an ordering over the (top) documents in the collection with respect to a query
- Free text queries: Rather than a query language of operators and expressions, the user's query is just one or more words in a human language
- In principle, there are two separate choices here, but in practice, ranked retrieval models have normally been associated with free text queries and vice versa

Feast or famine: not a problem in ranked retrieval

- When a system produces a ranked result set, large result sets are not an issue
 - Indeed, the size of the result set is not an issue
 - We just show the top k (≈ 10) results
 - We don't overwhelm the user

Premise: the ranking algorithm works

Scoring as the basis of ranked retrieval

- We wish to return in order the documents most likely to be useful to the searcher
- How can we rank-order the documents in the collection with respect to a query?
- Assign a score say in [0, 1] to each document
- This score measures how well document and query "match".

Query-document matching scores

- We need a way of assigning a score to a query/document pair
- Let's start with a one-term query
- If the query term does not occur in the document: score should be 0
- The more frequent the query term in the document, the higher the score (should be)
- We will look at a number of alternatives for this

Ch. 6

Take 1: Jaccard coefficient

- A commonly used measure of overlap of two sets A and B is the Jaccard coefficient
- jaccard(A,B) = $|A \cap B| / |A \cup B|$
- jaccard(A,A) = 1
- jaccard(A,B) = 0 if $A \cap B = 0$
- A and B don't have to be the same size.
- Always assigns a number between 0 and 1.

Jaccard coefficient: Scoring example

- What is the query-document match score that the Jaccard coefficient computes for each of the two documents below?
- Query: ides of march
- Document 1: caesar died in march
- Document 2: the long march

Issues with Jaccard for scoring

- It doesn't consider term frequency (how many times a term occurs in a document)
 - Rare terms in a collection are more informative than frequent terms
 - Jaccard doesn't consider this information
- We need a more sophisticated way of normalizing for length
 - Later in this lecture, we'll use $A \cap B | / \sqrt{|A \cup B|}$. . . instead of $|A \cap B| / |A \cup B|$ (Jaccard) for length normalization.

Recall: Binary term-document incidence matrix

	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

Each document is represented by a binary vector $\in \{0,1\}^{|V|}$

Term-document count matrices

- Consider the number of occurrences of a term in a document:
 - Each document is a count vector in N^{|V|}: a column below

	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

Bag of words model

- Vector representation doesn't consider the ordering of words in a document
- John is quicker than Mary and Mary is quicker than John have the same vectors
- This is called the bag of words model.
- In a sense, this is a step back: The positional index was able to distinguish these two documents
 - We will look at "recovering" positional information later on
 - For now: bag of words model

Term frequency tf

- The term frequency tf_{t,d} of term t in document d is defined as the number of times that t occurs in d.
- We want to use tf when computing query-document match scores. But how?
- Raw term frequency is not what we want:
 - A document with 10 occurrences of the term is more relevant than a document with 1 occurrence of the term.
 - But not 10 times more relevant.
- Relevance does not increase proportionally with term frequency.

Log-frequency weighting

The log frequency weight of term t in d is

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d}, & \text{if } tf_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

- $0 \to 0, 1 \to 1, 2 \to 1.3, 10 \to 2, 1000 \to 4, \text{ etc.}$
- Score for a document-query pair: sum over terms t in both q and d:
- score

$$= \sum_{t \in q \cap d} (1 + \log t f_{t,d})$$

 The score is 0 if none of the query terms is present in the document.

Sec. 6.2.1

Document frequency

- Rare terms are more informative than frequent terms
 - Recall stop words
- Consider a term in the query that is rare in the collection (e.g., arachnocentric)
- A document containing this term is very likely to be relevant to the query arachnocentric
- → We want a high weight for rare terms like arachnocentric.

Document frequency, continued

- Frequent terms are less informative than rare terms
- Consider a query term that is frequent in the collection (e.g., high, increase, line)
- A document containing such a term is more likely to be relevant than a document that doesn't
- But it's not a sure indicator of relevance.
- → For frequent terms, we want positive weights for words like high, increase, and line
- But lower weights than for rare terms.
- We will use document frequency (df) to capture this.

idf weight

- df_t is the <u>document</u> frequency of t: the number of documents that contain t
 - $-df_t$ is an inverse measure of the informativeness of t
 - $df_t \leq N$
- We define the idf (inverse document frequency)
 of t by $idf_t = log_{10} (N/df_t)$
 - We use log (N/df_t) instead of N/df_t to "dampen" the effect of idf.

Will turn out the base of the log is immaterial.

Sec. 6.2.1

idf example, suppose N = 1 million

term	df_t	idf_t
calpurnia	1	
animal	100	
sunday	1,000	
fly	10,000	
under	100,000	
the	1,000,000	

$$idf_t = \log_{10} \left(N/df_t \right)$$

There is one idf value for each term t in a collection.

Effect of idf on ranking

- Question: Does idf have an effect on ranking for one-term queries, like
 - iPhone
- idf has no effect on ranking one term queries
 - idf affects the ranking of documents for queries with at least two terms
 - For the query capricious person, idf weighting makes occurrences of capricious count for much more in the final document ranking than occurrences of person.

Sec. 6.2.1

Collection vs. Document frequency

 The collection frequency of t is the number of occurrences of t in the collection, counting multiple occurrences.

Example:

Word	Collection frequency	Document frequency
insurance	10440	3997
try	10422	8760

 Which word is a better search term (and should get a higher weight)?

tf-idf weighting

 The tf-idf weight of a term is the product of its tf weight and its idf weight.

$$\mathbf{W}_{t,d} = (1 + \log t \mathbf{f}_{t,d}) \times \log_{10}(N/d\mathbf{f}_t)$$

- · Best known weighting scheme in information retrieval
 - Note: the "-" in tf-idf is a hyphen, not a minus sign!
 - Alternative names: tf.idf, tf x idf
- Increases with the number of occurrences within a document
- Increases with the rarity of the term in the collection



Final ranking of documents for a query

$$Score(q,d) = \sum_{t \in q \cap d} tf.idf_{t,d}$$

Binary \rightarrow count \rightarrow weight matrix

	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

Each document is now represented by a real-valued vector of tf-idf weights $\in \mathbb{R}^{|V|}$

Ranked Retrieval

Ranked retrieval

- The first in a particular world for the common section of the comm
 - Ritabas a lot or sid into come up with inquery that produces or manageasts number of influ--With given see few, OR gives too many.

Term document count matrices

*	a discurrent Den document Seloui					
	with a set frequency	and bearing	No female	-	Mark.	
	190	19				
	100	87				

Problem with Society search:

- Stocker accessorier result in ordination for four forms many (1000c) results.
 Gunnation for four forms and SET 0 200 200 No. Gunnation for the SET 0 200 200 No. Gunnation for four forms and the SET 0 200 200 No. Gunnation for four forms for four forms for four four forms.
- Policy for a cold of decreased and trybip properly commoding for an energy decreased as the resident returns an ordering contract by past observation for a collection of inappetr a coldent.
 Internal property and a manufacture of and a manufacture organization of the contract of and a manufacture organization of the contract of and a manufacture organization.

Ranked retrieval models

Term frequency tf

- Bag of words model Vibratives process and on december about the ordering of words in a coolumns
- The form the price I (yet from the description) of defining up to produce of the support of the other pears the Mayard Mayor pears does the Pearse will be **Trining as the bay of words mode!
 • It is one, this is not place to Tail and the original state as a side to distinguish these has documents.
 • What is indicated the product the matter bits to the side of the state of the side of t
 - define a pre-mode of man state countries.

 Consider the and selection and generalized when gives results at the consideration of the co

Log-frequency weighting

- From the time runking algorithm works

Feast or famine: not a problem in ranked retrieval

When a system products can find resulting large most set can make a large most set can make an issue — In deed, this large most set can always a facility of the most set of the make a large most set of the most set of

- · The being percyyothical term! In a b-. The large magnetic properties of the state of the stat

Scoring as the basis of ranked retrieval

- result segret as a set in the section.

 Have one such and the following of the callector with respect to a graph.

 Adapt section—see in 0, 31—to call decrease.

 His source results the well-decreased and green fraces?

Document frequency Document frequency, continued

Final cacking of documents for a query.

 $\mathsf{Score}(q,d) = \sum\nolimits_{i \subseteq q \cap d} \mathsf{tf.idf}_{i,d}$

Fare terms are more informative than frequent form
 Foreign accounts
 Condition are not in the query than is need to the condition of indian large, excellengements.

Adocument containing this term is very Bally to be relevant to the query internation by

- 3. We want a high weight for sare terms like

ti idi weighting

 The Hild solphists to misting and at all its disolphism (its disolphism). $\mathbf{w}_{t_i} = (1 - \log t f_{t_i}') * \log_{\mathbf{k}'}(N / \mathrm{d}f_i)$

Betchmanning trigodient in about too male in Betchmanning trigodient in about too male in Hall de mail de transport of a manager in A male transport (A) in A in Betchmanning (A) in

- We need a value of antighting a convention a case systemator participating and antighting a convention a case systemator part.

 If they consider the considering system conventions in the convention convention to the convention convention to the convention of the convention to the

The analyse make implicit method that is a server of considering properties of the server in the control of the collection May Depth and the server in the collection May Depth and the server in the first interest that for each other interest in the server in the ser

Binary → count → weight matrix | No. | No.

tech constraint is everywherent to a restorbed vector of their very state of the

- · Jaconijs,et Jaconijs, la U.K.
- accord(4,4) = 1 (accord(A.5) = 0 fra n D = 0
- A and Adop't have to be the same doe.
 Above assigns a number between Band 1.

- We define the of (inverse document frequency of the lidit lag ₁₁ (Midf.)

idf weight

- di, is the document flow ency of to the currier of documents that contain it
 di, an extra menute of the information of the distribution.
- We can be graph by material of April, to "Command the office of the

00	b =	e.	11.
	valor te		_
	4.84	100	
	name.	1.16	
4	1.	90.00	
	200	\$00.000	
	i v	1.000,000	

$idf_i - log_{ig}(Mdf_i)$ Thurs is one left value for each sure nin a collection.

Jaccard coefficient: Scoring example

What is the many observed mater some that the isomediscredificant computer for each of the medicuments wildow?
Game also affected:
Game also affected:
Game also affected in the south

idf example, suppose N=1 million

Recall: Binary term-document incidence matrix Issues with Jaccard for scoring Lidour Loriskie kreinfrag ong Doermang Indoording demonstration mengha-handere in demonstration mengha-handere in demonstration of strandered that the automate of strandered on the strandered of the Wellington extracted behander on your commander for breather of the strandered on the strandered of the

Collection vs. Document frequency



Which word is a better search term (and should get a higher weight)?

Vector Space Model

Representative sections

- Numerous Monerous Indianasis



Color stated yearings followed to

Gueries assectats

Bernstein, D. Constant (Charmeter, editor).
 Les de Constant (Charmeter, editor).



Formalizing vectors are provincy

This can detailed to even a vectorial for more investment and gave not to the candidate of the candidat

3 documenta pramote contri.

"Tomath digits" and firm a document of the organization of the many of the organization." The complete of the organization of Accessed to a reversal began because progress the fact court and last, continued of the last of the selection.

le oberet per contrattement : mai comme informer provides and belong consultations : bed comme information and which information Ander two communically describe places on the rest successifier, Well



continuous listen Amount to be Bergill down in well to de 2 %, see that the performance of the larger in the see the first popular.

cosine) presy recurrenti The first of the second section in materials of the second section of the second section of

Effect of ldf on ranking

Question: Does let have an effect or randing for anothern queries. We

Filter - effect on a large retain quality - effect or self-dent for a large following the grown - effect of the self-dent grown of the se

seinerfor ingrenorial activities To a full-randoul velocity is in that, case, the corporate is exist within $m_1((d) - 2nd - \sum_{i=1}^{M} a_i d_i$

Documents as vectors

- Now we have a |V|-dimensional vector space
- Terms are axes of the space
- Documents are points or vectors in this space
- Very high-dimensional: tens of millions of dimensions when you apply this to a web search engine
- These are very sparse vectors most entries are zero

Queries as vectors

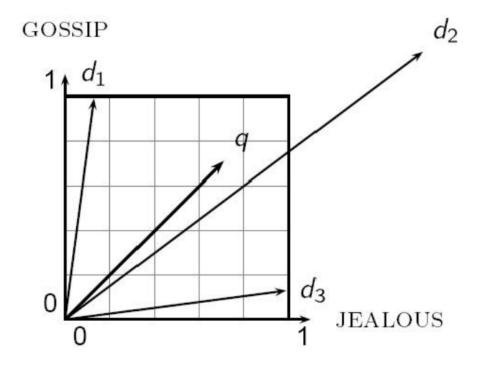
- Key idea 1: Do the same for queries: represent them as vectors in the space
- Key idea 2: Rank documents according to their proximity to the query in this space
- proximity = similarity of vectors
- proximity ≈ inverse of distance
- Recall: We do this because we want to get away from the you're-either-in-or-out Boolean model
- Instead: rank more relevant documents higher than less relevant documents

Formalizing vector space proximity

- First cut: distance between two points
 - (= distance between the end points of the two vectors)
- Euclidean distance?
- Euclidean distance is a bad idea . . .
- ... because Euclidean distance is large for vectors of different lengths.

Why distance is a bad idea

The Euclidean distance between q and d_2 is large even though the distribution of terms in the query q and the distribution of terms in the document d_2 are very similar.



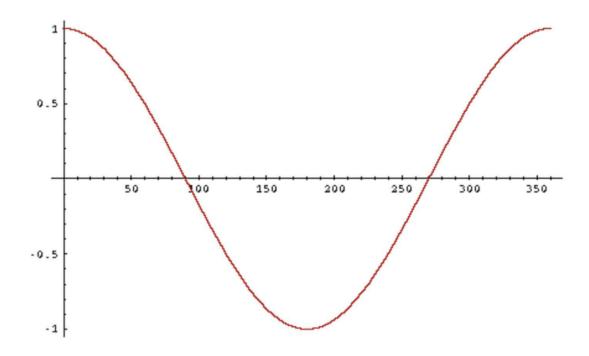
Use angle instead of distance

- Thought experiment: take a document d and append it to itself. Call this document d'.
- "Semantically" d and d' have the same content
- The Euclidean distance between the two documents can be quite large
- The angle between the two documents is 0, corresponding to maximal similarity.
- Key idea: Rank documents according to angle with query.

From angles to cosines

- The following two notions are equivalent.
 - Rank documents in <u>decreasing</u> order of the angle between query and document
 - Rank documents in <u>increasing</u> order of cosine(query,document)
- Cosine is a monotonically decreasing function for the interval [0°, 180°]

From angles to cosines



• But how – and why – should we be computing cosines?

Length normalization

 A vector can be (length-) normalized by dividing each of its components by its length – for this we use the L₂ norm:

 $\left\| \vec{x} \right\|_2 = \sqrt{\sum_i x_i^2}$

- Dividing a vector by its L₂ norm makes it a unit (length) vector (on surface of unit hypersphere)
- Effect on the two documents d and d' (d appended to itself) from earlier slide: they have identical vectors after length-normalization.
 - Long and short documents now have comparable weights

Sec. 6.3

cosine(query,document)

Dot product
$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

 q_i is the tf-idf weight of term i in the query d_i is the tf-idf weight of term i in the document

 $\cos(\vec{q}, \vec{d})$ is the cosine similarity of \vec{d} and \vec{d} ... or, equivalently, the cosine of the angle between \vec{d} and \vec{d} .

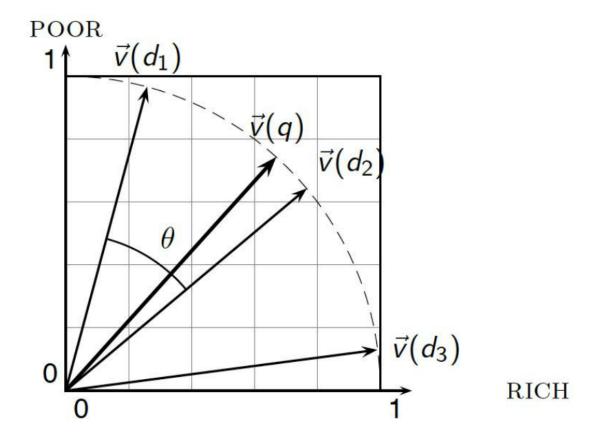
Cosine for length-normalized vectors

 For length-normalized vectors, cosine similarity is simply the dot product (or scalar product):

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

for q, d length-normalized.

Cosine similarity illustrated



Cosine similarity amongst 3 documents

How similar are

the novels

SaS: Sense and

Sensibility

PaP: Pride and

Prejudice, and

WH: Wuthering

Heights?

term	SaS	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

Term frequencies (counts)

Note: To simplify this example, we don't do idf weighting.

3 documents example contd.

Log frequency weighting

After length normalization

term	SaS	PaP	WH
affection	3.06	2.76	2.30
jealous	2.00	1.85	2.04
gossip	1.30	0	1.78
wuthering	0	0	2.58

term	SaS	PaP	WH
affection	0.789	0.832	0.524
jealous	0.515	0.555	0.465
gossip	0.335	0	0.405
wuthering	0	0	0.588

```
cos(SaS,PaP) \approx
0.789 \times 0.832 + 0.515 \times 0.555 + 0.335 \times 0.0 + 0.0 \times 0.0 \approx 0.94
cos(SaS,WH) \approx 0.79
cos(PaP,WH) \approx 0.69
```

Why do we have cos(SaS,PaP) > cos(SAS,WH)?