

DMMML, 21 March 2019

Web Search

Text documents → identify subset relevant
to a given requirement

Information Retrieval (IR)

Library cataloguing

Legal documents

Medical literature



Can trust
the author



Not true of
Internet docs

Information need



Query

phrase

Structured
form

Matching
documents

Relevance
ranking

Query is a phrase - collection of words

Preprocess & build a summary of documents

How to model a document

Boolean model : documents are sets
of words

Documents : d_1, d_2, \dots

Vocabulary : t_1, t_2, \dots

↓
terms

Term-Document Matrix

terms

documents

| | 1 | 2 | 3 | 4 | 5 | 6 | ... |
|-----------|----------------------|---------------|-------------|--------|---------|---------|-----|
| | 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 | |
| ... | | | | | | | |

Query: "Antony Cleopatra" implicitly and

Retrieve row for each word

Look for 1 in both rows

Extend to Boolean Queries

"(Antony or Brutus) and not (Cleopatra)"

Very wasteful representation - most entries are 0

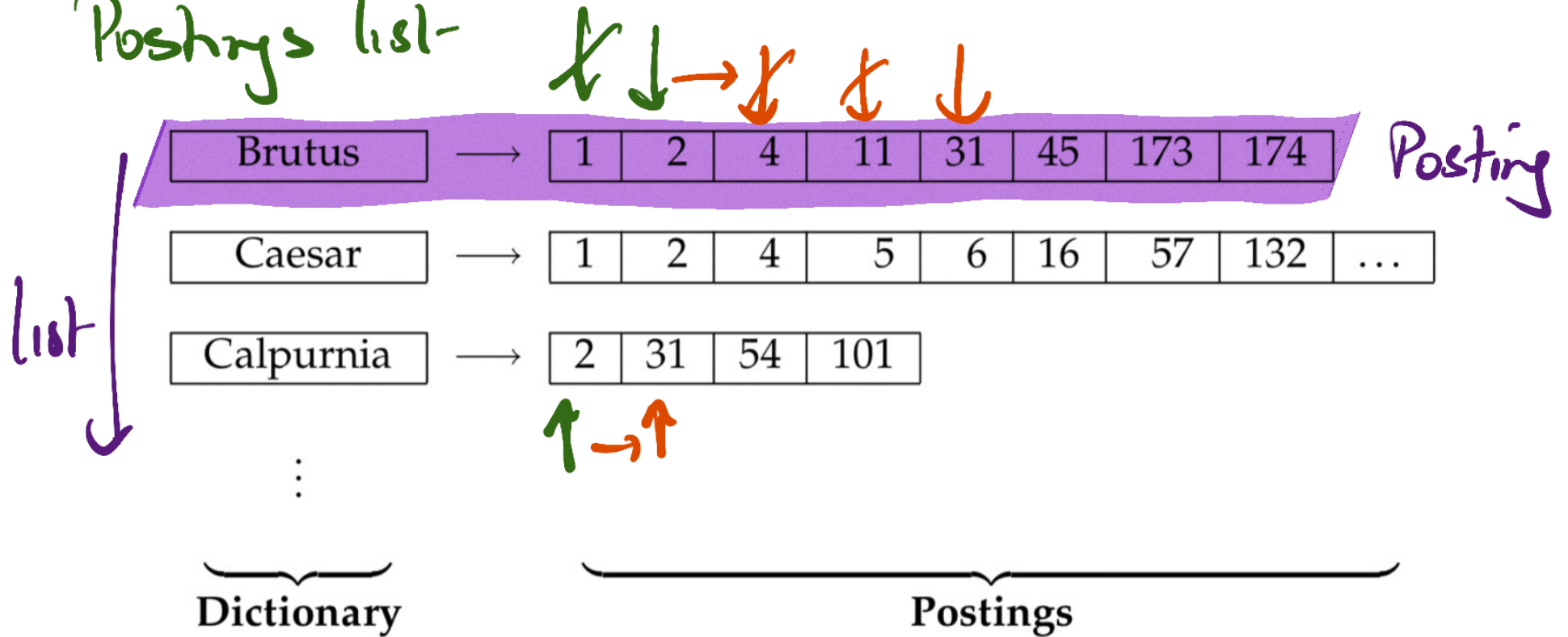
Instead, represent each row as a list

Antony : [1, 2, 6]

Brutus : [1, 2, 4]

Posting - one entry

Postings list-



"Brutus" AND "Calpurnia"

X, 2, X, X, 31

One typical enhancement

Maintain a count with each posting.

| | | | | | | | | | | |
|-----------|----|---|----|----|-----|----|----|-----|-----|-----|
| Brutus | 8 | 1 | 2 | 4 | 11 | 31 | 45 | 173 | 174 | |
| Caesar | 36 | 1 | 2 | 4 | 5 | 6 | 16 | 57 | 132 | ... |
| Calpurnia | 4 | 2 | 31 | 54 | 101 | | | | | |
| ⋮ | | | | | | | | | | |

Dictionary

Postings

²⁸ Brutus \wedge ³⁶ Caesar \wedge ⁴ Calpurnia | Optimiz
 Brutus \wedge Calpurnia \wedge Caesar | evaluat
₂₈ ₄

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

| term | doc. freq. | → | postings lists |
|-----------|------------|---|----------------|
| ambitious | 1 | → | 2 |
| be | 1 | → | 2 |
| brutus | 2 | → | 1 → 2 |
| capitol | 1 | → | 1 |
| caesar | 2 | → | 1 → 2 |
| did | 1 | → | 1 |
| enact | 1 | → | 1 |
| hath | 1 | → | 2 |
| I | 1 | → | 1 |
| i' | 1 | → | 1 |
| it | 1 | → | 2 |
| julius | 1 | → | 1 |
| killed | 1 | → | 1 |
| let | 1 | → | 2 |
| me | 1 | → | 1 |
| noble | 1 | → | 2 |
| so | 1 | → | 2 |
| the | 2 | → | 1 → 2 |
| told | 1 | → | 2 |
| you | 1 | → | 2 |
| was | 2 | → | 1 → 2 |
| with | 1 | → | 2 |

What terms to index?

Stop words - frequently occurring words



like a, the, is, and, or, ...

Identification requires some expertise or extra computation

Usually strategy - drop stop words

Query = Flights from Chennai to Pune

Typically web search does not ignore stop words

Normalization

truck vs trucks

department, departmental?

2 strategies

(a) Use knowledge of language (NLP)

Saw is past form of see

↳ but also a carpenter's tool!

Reduce all forms of "to see" to a

standard representation **Lemmatization**

(b) Syntactic truncation

placing
placement

Rule : $x + \text{ing} \rightarrow x$
 $x + \text{ement} \rightarrow x$

Rewriting rules

$\text{placing} \rightarrow \text{plac}$
 $\text{placement} \rightarrow \text{plac}$

Cement \rightarrow c ?

$x + \text{ement}, |x| > 1 \rightarrow x$

Rule based truncation — Stemming

Porter's Stemming rules for English

Other issues.

Hyphens

Apostrophes

Multword entities — New Delhi

Phrase queries

"President of India"

President ^ of ^ India

Current postings list cannot answer a phrase query meaningfully

Expand terms to sets of words.

Separate posting for "President of India"

Vocabulary V

$|V|$ postings in original postings list

$|V| \times |V|$ pair postings

$|V| \times |V| \times |V|$ tuples

What is the limit on the length of a phrase?

Keep positional information with postings.


Not just a boolean model

$w \rightarrow d_1, d_2, d_3, \dots$

Instead

$w \rightarrow d_1: [p_1, p_2, \dots], d_2: [p'_1, p'_2, \dots]$

to, 993427:  document wise count

original posting  $\langle 1, 6: \langle 7, 18, 33, 72, 86, 231 \rangle; \text{positions in document}$
 $\langle 2, 5: \langle 1, 17, 74, 222, 255 \rangle;$
 $\langle 4, 5: \langle 8, 16, 190, 429, 433 \rangle;$
 $\langle 5, 2: \langle 363, 367 \rangle;$
 $\langle 7, 3: \langle 13, 23, 191 \rangle; \dots \rangle$

Does d_j contain "President of India"?

j is in intersection of 3 postings

President $j: < \underline{k} >$

of $j: < \underline{k+1} >$

India $j: < \quad k+2 >$