

CS347

Lecture 1
April 4, 2001
CPrahakar Rayhanan

Query

- Which plays of Shakespeare contain the words *Brutus AND Caesar* but *NOT Calpurnia*?

Term-document incidence

	<i>Antony and Cleopatra</i>	<i>Julius Caesar</i>	<i>The Tempest</i>	<i>Hamlet</i>	<i>Brutus</i>	<i>Marked</i>
<i>Antony</i>	1	0	0	0	0	1
<i>Brutus</i>	1	1	0	1	0	0
<i>Caesar</i>	1	1	0	1	1	1
<i>Calpurnia</i>	0	1	0	0	0	0
<i>Cleopatra</i>	1	0	0	0	0	0
<i>tempest</i>	1	0	1	1	1	1
<i>marked</i>	1	0	1	1	1	0

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) \rightarrow bitwise *AND*.
- $110100 \text{ AND } 110111 \text{ AND } 101111 = 100100$.

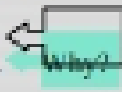
Answers to query

- Antony and Cleopatra, Act III, Scene ii
 - Agrippa (Aide to DOMITIUS ENDOBARSUS): Why, Embarbatus,
 - When Antony found Julius **Caesar** dead,
 - He died almost to weeping; and he wept
 - When at Philipp he found **Brutus** slain.
- Hamlet, Act III, Scene ii
 - Lord Polonius: I did enact Julius **Caesar** I was killed 't the
 - Capitol, **Brutus** killed me.

Bigger corpora

- Consider $n = 1M$ documents, each with about 1K terms.
- Avg 6 bytes/term incl spaces/punctuation - 6GB of data.
- 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?

Inverted index

- Documents are parsed to extract words and these are saved with the Document ID.

Doc 1
I did enact Julius
Caesar I was killed
't the Capitol;
Brutus killed me.

Doc 2
So let it be with
Caesar. The noble
Brutus hath told you
Caesar was ambitious

Word	Doc ID
I	1
did	1
enact	1
Julius	1
Caesar	1
I	1
was	1
killed	1
't	1
the	1
Capitol	1
;	1
Brutus	1
killed	1
me	1
.	1
So	2
let	2
it	2
be	2
with	2
Caesar	2
.	2
The	2
noble	2
Brutus	2
hath	2
told	2
you	2
Caesar	2
was	2
ambitious	2
.	2

- After all documents have been parsed the inverted file is sorted by terms

Term	DocID	Term	DocID
the	1	the	2
the	1	the	3
the	1	the	4
the	1	the	5
the	1	the	6
the	1	the	7
the	1	the	8
the	1	the	9
the	1	the	10
the	1	the	11
the	1	the	12
the	1	the	13
the	1	the	14
the	1	the	15
the	1	the	16
the	1	the	17
the	1	the	18
the	1	the	19
the	1	the	20
the	1	the	21
the	1	the	22
the	1	the	23
the	1	the	24
the	1	the	25
the	1	the	26
the	1	the	27
the	1	the	28
the	1	the	29
the	1	the	30
the	1	the	31
the	1	the	32
the	1	the	33
the	1	the	34
the	1	the	35
the	1	the	36
the	1	the	37
the	1	the	38
the	1	the	39
the	1	the	40
the	1	the	41
the	1	the	42
the	1	the	43
the	1	the	44
the	1	the	45
the	1	the	46
the	1	the	47
the	1	the	48
the	1	the	49
the	1	the	50
the	1	the	51
the	1	the	52
the	1	the	53
the	1	the	54
the	1	the	55
the	1	the	56
the	1	the	57
the	1	the	58
the	1	the	59
the	1	the	60
the	1	the	61
the	1	the	62
the	1	the	63
the	1	the	64
the	1	the	65
the	1	the	66
the	1	the	67
the	1	the	68
the	1	the	69
the	1	the	70
the	1	the	71
the	1	the	72
the	1	the	73
the	1	the	74
the	1	the	75
the	1	the	76
the	1	the	77
the	1	the	78
the	1	the	79
the	1	the	80
the	1	the	81
the	1	the	82
the	1	the	83
the	1	the	84
the	1	the	85
the	1	the	86
the	1	the	87
the	1	the	88
the	1	the	89
the	1	the	90
the	1	the	91
the	1	the	92
the	1	the	93
the	1	the	94
the	1	the	95
the	1	the	96
the	1	the	97
the	1	the	98
the	1	the	99
the	1	the	100

- Multiple term entries in a single document are merged and frequency information added

Term	DocID	Term	DocID	Count
the	1	the	2	1
the	1	the	3	1
the	1	the	4	1
the	1	the	5	1
the	1	the	6	1
the	1	the	7	1
the	1	the	8	1
the	1	the	9	1
the	1	the	10	1
the	1	the	11	1
the	1	the	12	1
the	1	the	13	1
the	1	the	14	1
the	1	the	15	1
the	1	the	16	1
the	1	the	17	1
the	1	the	18	1
the	1	the	19	1
the	1	the	20	1
the	1	the	21	1
the	1	the	22	1
the	1	the	23	1
the	1	the	24	1
the	1	the	25	1
the	1	the	26	1
the	1	the	27	1
the	1	the	28	1
the	1	the	29	1
the	1	the	30	1
the	1	the	31	1
the	1	the	32	1
the	1	the	33	1
the	1	the	34	1
the	1	the	35	1
the	1	the	36	1
the	1	the	37	1
the	1	the	38	1
the	1	the	39	1
the	1	the	40	1
the	1	the	41	1
the	1	the	42	1
the	1	the	43	1
the	1	the	44	1
the	1	the	45	1
the	1	the	46	1
the	1	the	47	1
the	1	the	48	1
the	1	the	49	1
the	1	the	50	1
the	1	the	51	1
the	1	the	52	1
the	1	the	53	1
the	1	the	54	1
the	1	the	55	1
the	1	the	56	1
the	1	the	57	1
the	1	the	58	1
the	1	the	59	1
the	1	the	60	1
the	1	the	61	1
the	1	the	62	1
the	1	the	63	1
the	1	the	64	1
the	1	the	65	1
the	1	the	66	1
the	1	the	67	1
the	1	the	68	1
the	1	the	69	1
the	1	the	70	1
the	1	the	71	1
the	1	the	72	1
the	1	the	73	1
the	1	the	74	1
the	1	the	75	1
the	1	the	76	1
the	1	the	77	1
the	1	the	78	1
the	1	the	79	1
the	1	the	80	1
the	1	the	81	1
the	1	the	82	1
the	1	the	83	1
the	1	the	84	1
the	1	the	85	1
the	1	the	86	1
the	1	the	87	1
the	1	the	88	1
the	1	the	89	1
the	1	the	90	1
the	1	the	91	1
the	1	the	92	1
the	1	the	93	1
the	1	the	94	1
the	1	the	95	1
the	1	the	96	1
the	1	the	97	1
the	1	the	98	1
the	1	the	99	1
the	1	the	100	1

- The file is commonly split into a *Dictionary* and a *Postings* file



- Where do we pay in storage?

