# CS347

Lecture 1

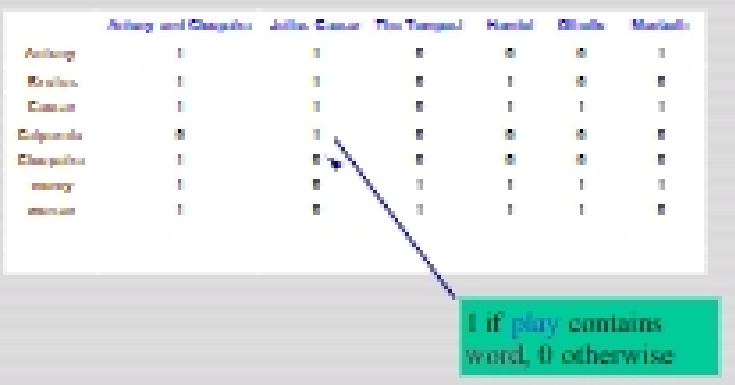
April 4, 2001



## Query

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

## Term-document incidence




## 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 *AND* 101111 = 100100.

## 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.

## 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.

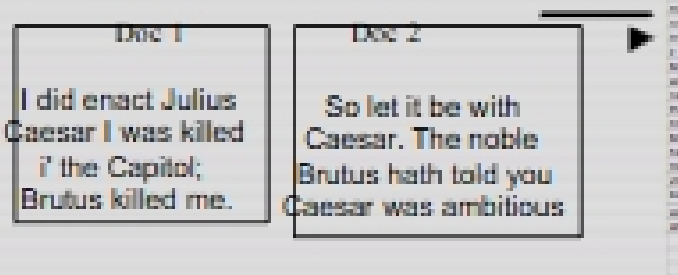
## Can't build the matrix

- 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?

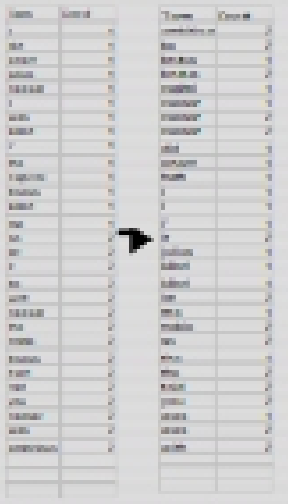
Why?

## Inverted index

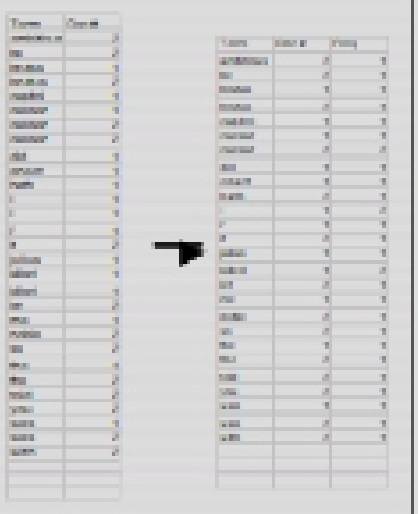
- Documents are parsed to extract words and these are saved with the Document ID.



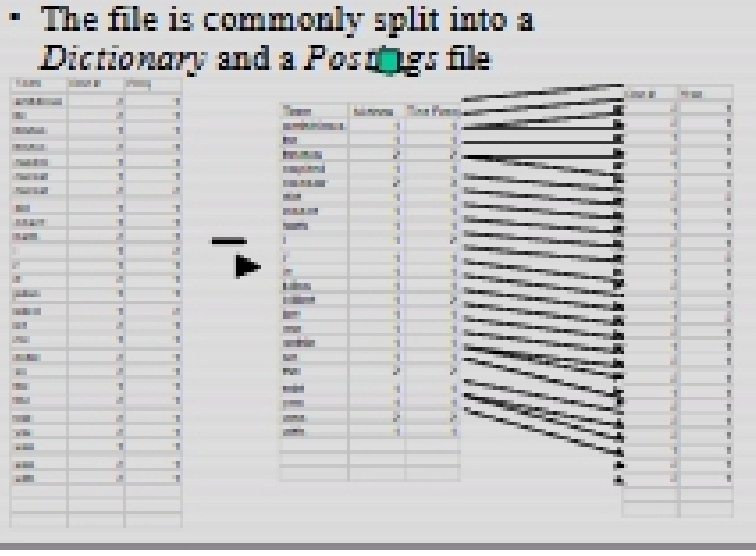
- After all documents have been parsed the inverted file is sorted by terms



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



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



- Where do we pay in storage?