Unstructured Data

Agenda

- Definition
- Target Audience
- Example
- Document and Corpus
- Term Frequency Inverse Document Frequency

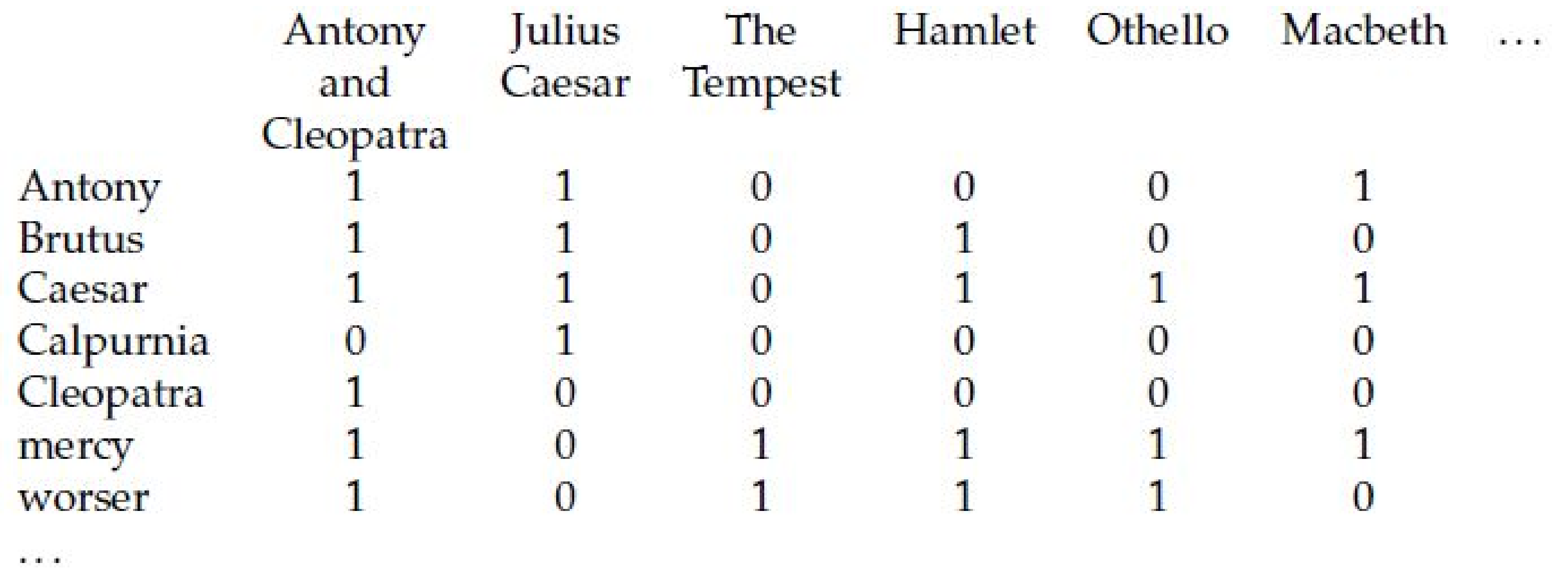
Target Audience

- Search Engines
- Recommendation Engines
- Paralegals
- Librarians

- Works on Structured, unstructured and semi-structured data

Example

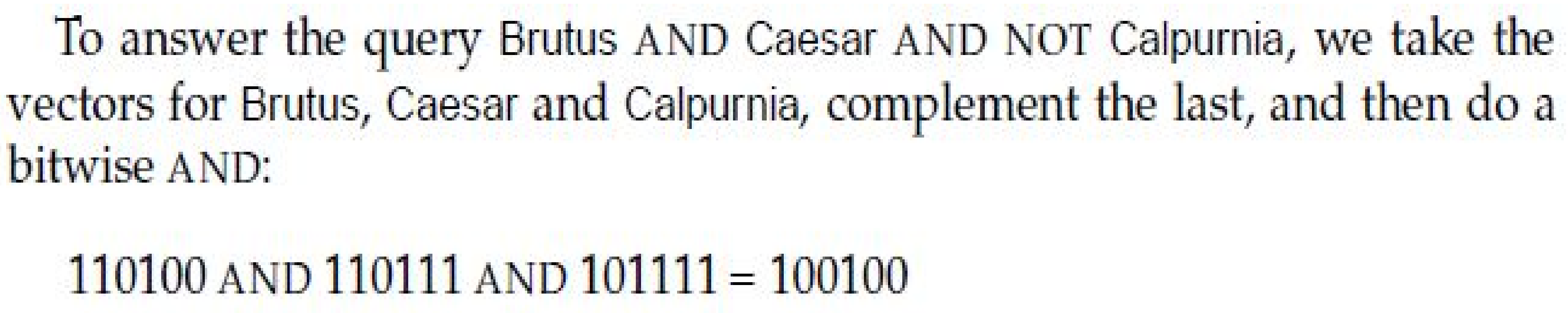
- Search the chapter that contains Brutus and Caesar but not Calpurnia



	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	
...							

Term Document Incidence Matrix

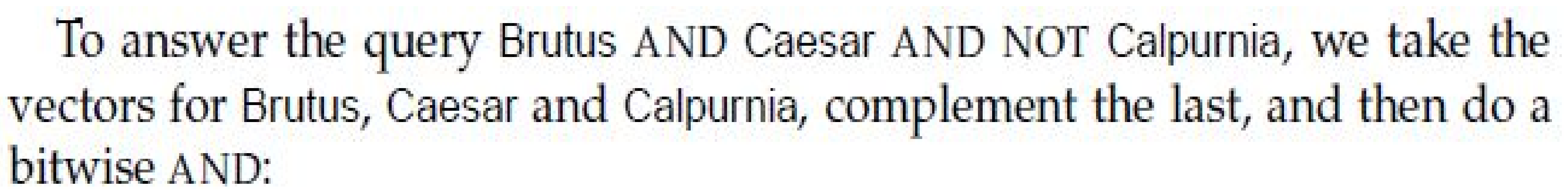
Example Contd.

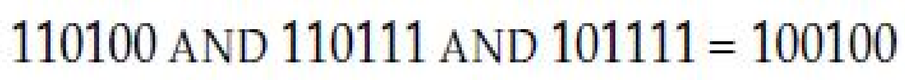


To answer the query Brutus AND Caesar AND NOT Calpurnia, we take the vectors for Brutus, Caesar and Calpurnia, complement the last, and then do a bitwise AND:

110100 AND 110111 AND 101111 = 100100

Example Contd.


To answer the query Brutus AND Caesar AND NOT Calpurnia, we take the vectors for Brutus, Caesar and Calpurnia, complement the last, and then do a bitwise AND:


110100 AND 110111 AND 101111 = 100100

The answers for this query are *Antony and Cleopatra* and *Hamlet*

Document and Corpus

- Corpus is a collection of documents
- Documents can be web page, product review, chapter of a book (book becomes the corpus), a whole book (a collection of books become the corpus), memos etc.

Term Frequency

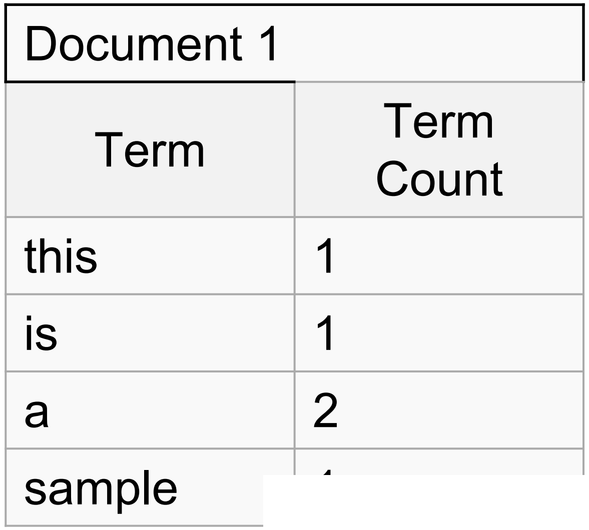
- Word Count → In the previous example we were considering Boolean options only. This time we also considered how many *times* a word occurred in the document

Term Frequency Inverse Document Frequency (TF-IDF)

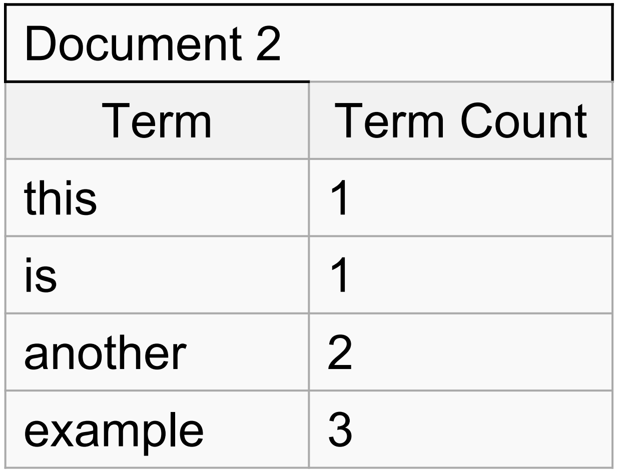
- IDF is a measure of the rareness of a term
- TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

TF-IDF Example

Example:



Document 1	
Term	Term Count
this	1
is	1
a	2
sample	1



Document 2	
Term	Term Count
this	1
is	1
another	2
example	3

$$\mathrm{idf}(\mathrm{this}, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$