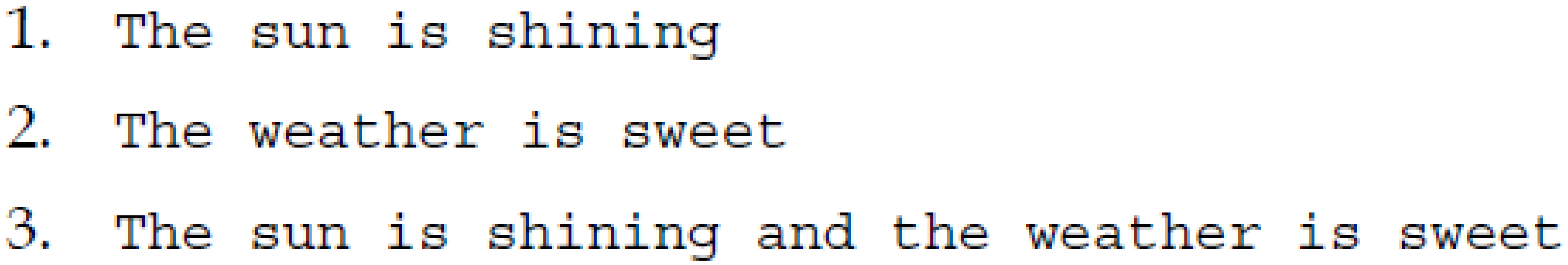
Unit 7 – Inverted Index

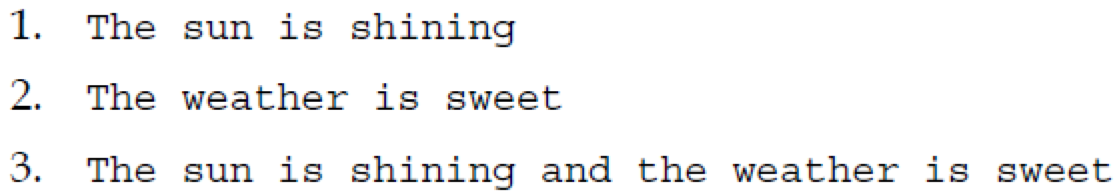
CS 201 - Data Structures II
Spring 2022
Habib University

Syeda Saleha Raza

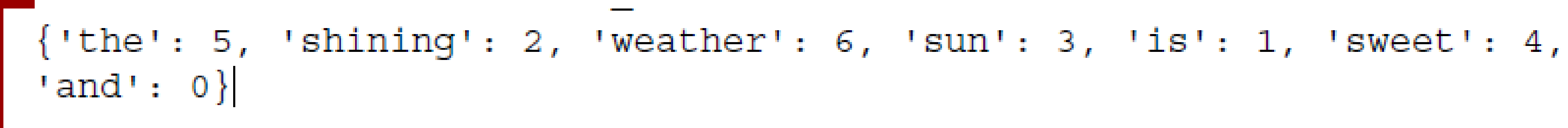
Examples



1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet



1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet



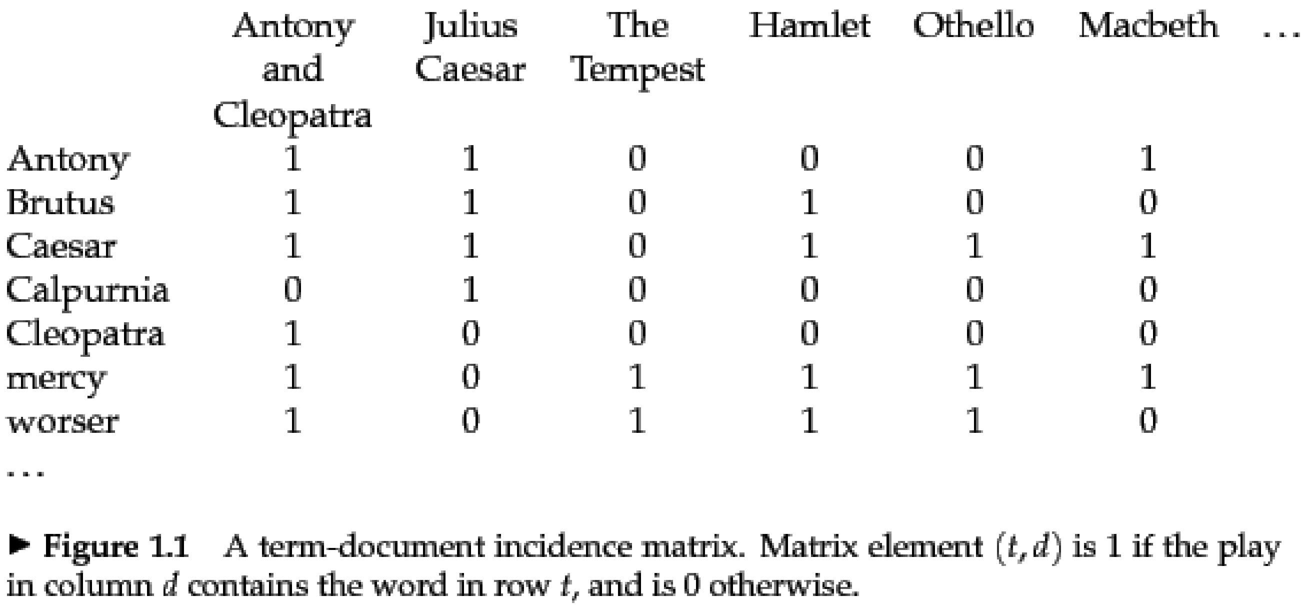
```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

	Doc 1	Doc 2	Doc 3
And	0	0	1
Is	1	1	1
Shining	1	0	1
Sun	1	0	1
Sweet	0	1	1
the	1	1	1
weather	0	1	1

Drawbacks

- Size of matrix
- Too big to fit in memory
- Sparsity

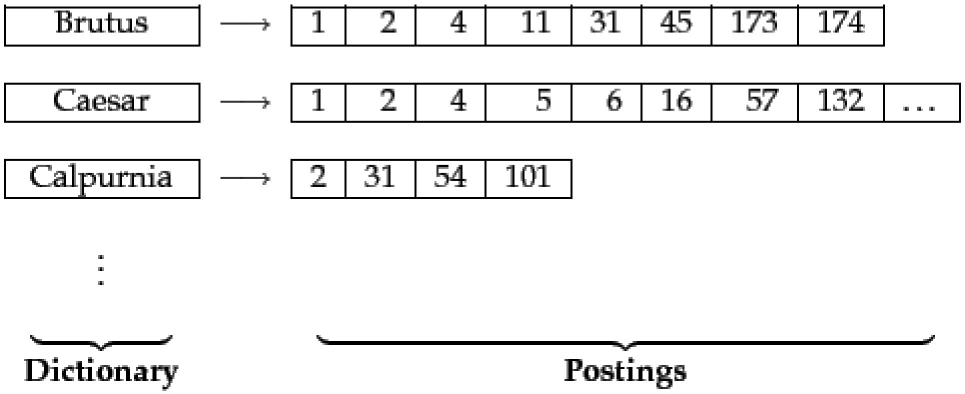
Binary Vectors



	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	
…							

► **Figure 1.1** A term-document incidence matrix. Matrix element (t, d) is 1 if the play in column d contains the word in row t, and is 0 otherwise.

Inverted Index



Brutus → 1 | 2 | 4 | 11 | 31 | 45 | 173 | 174

Caesar → 1 | 2 | 4 | 5 | 6 | 16 | 57 | 132 | …

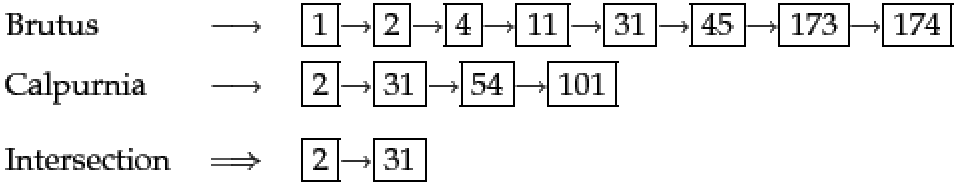
Calpurnia → 2 | 31 | 54 | 101

⋮

► **Figure 1.2** The two parts of an inverted index. The dictionary is commonly kept in memory, with pointers to each postings list, which is stored on disk.

Processing Queries

- Retrieve documents containing ‘Brutus AND Calpurnia’



Brutus ⟶ 1 → 2 → 4 → 11 → 31 → 45 → 173 → 174

Calpurnia ⟶ 2 → 31 → 54 → 101

Intersection ⟹ 2 → 31

Creating an inverted index

- Recall the major steps in inverted index construction:
  1. Collect the documents to be indexed.
  2. Tokenize the text.
  3. Do linguistic preprocessing of tokens.
  4. Index the documents that each term occurs in.

Preprocessing

- Removing stopwords
- Removing hyphens
- Abbreviations
- Maintaining synonyms
- Stemming
- Lemmatization

Text Preprocessing

- *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time
- *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma.*

Exercise

- Given the following set of documents:

Doc 1	new home sales top forecasts
Doc 2	home sales rise in july
Doc 3	increase in home sales in july
Doc 4	july new home sales rise

- Draw term document incidence-matrix for the given set of documents
- Draw the inverted-index representation.

Thanks