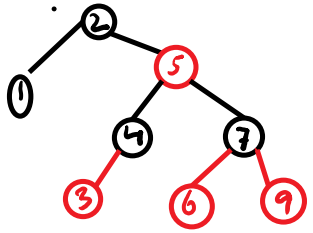


Habib University
shaping futures

# CS 201 Data Structure II (L2 / L5)

# Term frequency and weighting

**Muhammad Qasim Pasta**

qasim.pasta@sse.habib.edu.pk

**Slides are designed to be filled during the lectures. Some details are intentionally mentioned to be discussed in the class. These slides should not be used as reading.**

# term-document incidence matrix


- A matrix in which each value indicates whether a term occur in a document or not

- Document 1: It is going to rain today.
- Document 2: Today I am not going outside.
- Document 3: I am going to watch the season premiere.

# Term Frequency

- Number of occurrences of a term in a document

| | Doc1 | Doc2 | Doc3 |
|---|---|---|---|
| car | 27 | 4 | 24 |
| auto | 3 | 33 | 0 |
| insurance | 0 | 33 | 29 |
| best | 14 | 0 | 17 |

- There could biasness: what if 27 is out of 100 words and 4 is out of 10 words?
- Solution: divide number of occurance by total words

# Inverse document frequency

- measures how often a term appears across all documents in the corpus

$$\text{idf}_t = \log \frac{N}{\text{df}_t}.$$

| | Doc1 | Doc2 | Doc3 |
|---|---|---|---|
| car | 27 | 4 | 24 |
| auto | 3 | 33 | 0 |
| insurance | 0 | 33 | 29 |
| best | 14 | 0 | 17 |

# Tf-idf weighting

- a weighting scheme

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$$

- Score for a query:

$$\text{Score}(q, d) = \sum_{t \in q} \text{tf-idf}_{t,d}.$$