# TF IDF Weighting

## TF-IDF Weighting

- The tf-idf weight of a term $t$ is the product of $t$ 's tf weight and $t$ 's idf weight

$$W_{t,d} = (1 + logtf_{t,d}) \times log_{10}(N/df_t)$$

- **Best Known weighting scheme** in information retrieval
- Increases with **the number of occurrences within a document** → TF
- Increases with **the rarity of the term in the collection** → IDF

## Final Ranking of Documents for a Query

$$Score(q,d) = \sum_{t \in q \cap d} tf.idf_{t,d}$$

## Binary → Count → Weight Matrix

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 5.25 | 3.18 | 0 | 0 | 0 | 0.35 |
| Brutus | 1.21 | 6.1 | 0 | 1 | 0 | 0 |
| Caesar | 8.59 | 2.54 | 0 | 1.51 | 0.25 | 0 |
| Calpurnia | 0 | 1.54 | 0 | 0 | 0 | 0 |
| Cleopatra | 2.85 | 0 | 0 | 0 | 0 | 0 |
| mercy | 1.51 | 0 | 1.9 | 0.12 | 5.25 | 0.88 |
| worser | 1.37 | 0 | 0.11 | 4.15 | 0.25 | 1.95 |

- Each document is now represented by a real-valued vector of **tf-idf weights** $\in R^{|V|}$

## 출처

- stanford IR 강의(https://www.youtube.com/watch?v=4-P3ckZprBk&list=PLaZQkZp6WhWwoDuD6pQCmgVyDbUWl_ZUi&index=11)