# Exercise

We now consider the query ***best car insurance*** on a fictitious collection with *N* = 1,000,000 documents where the document frequencies of auto, best, car and insurance are respectively 5000, 50000, 10000 and 1000.

What is the score (cosine similarity) for this query with a document “car insurance auto insurance”? l**ogarithmic term weighting (wf columns) for query and raw term frequency for document**, **idf weighting for the query only, and length normalization for document only**.

# tf-idf example: lnc.ltc

Document: car insurance auto insurance

Query: best car insurance

| Term | Query | | | | | | Document | | | | Prod |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | tf-raw | tf-wt | df | idf | wt | n’lize | tf-raw | tf-wt | wt | n’lize | |
| auto | | | | | | | | | | | |
| best | | | | | | | | | | | |
| car | | | | | | | | | | | |
| insurance | | | | | | | | | | | |

Key to columns:

- tf-raw: raw (unweighted) term frequency,
- tf-wt: logarithmically weighted term frequency,
- df: document frequency,
- idf: inverse document frequency,

# tf-idf example: lnc.ltc

Document: car insurance auto insurance

Query: best car insurance

| Term | Query | | | | | | Document | | | | Prod |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | tf-raw | tf-wt | df | idf | wt | n'lize | tf-raw | tf-wt | wt | n'lize | |
| auto | 0 | | | | | | 1 | | | | |
| best | 1 | | | | | | 0 | | | | |
| car | 1 | | | | | | 1 | | | | |
| insurance | 1 | | | | | | 2 | | | | |

Key to columns:

- tf-raw: raw (unweighted) term frequency,
- tf-wt: logarithmically weighted term frequency,
- df: document frequency,
- idf: inverse document frequency

# tf-idf example: lnc.ltc

Document: car insurance auto insurance
Query: best car insurance

| Term | Query | | | | | | Document | | | | Prod |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | tf-raw | tf-wt | df | idf | wt | n’lize | tf-raw | tf-wt | wt | n’lize | |
| auto | 0 | 0 | 5000 | 2.3 | | | 1 | 1 | | | |
| best | 1 | 1 | 50000 | 1.3 | | | 0 | 0 | | | |
| car | 1 | 1 | 10000 | 2.0 | | | 1 | 1 | | | |
| insurance | 1 | 1 | 1000 | 3.0 | | | 2 | 1.3 | | | |

Key to columns:

- wt: the final weight of the term in the query or document,

# tf-idf example: lnc.ltc

Document: car insurance auto insurance
Query: best car insurance

| Term | Query | | | | | | Document | | | Prod |
|---|---|---|---|---|---|---|---|---|---|---|
| | tf-raw | tf-wt | df | idf | tf-idf | Normalization | tf-raw | tf-wt | Normalization | |
| auto | 0 | 0 | 5000 | 2.3 | 0 | 0 | 1 | 1 | 1/√6 | 0 |
| best | 1 | 1 | 50000 | 1.3 | 1.3 | 0.34 | 0 | 0 | 0 | 0 |
| car | 1 | 1 | 10000 | 2.0 | 2.0 | 0.52 | 1 | 1 | 1/√6 | 2*1/√6 |
| insurance | 1 | 1 | 1000 | 3.0 | 3.0 | 0.78 | 2 | 1.3 | 2/√6 | 3*2/√6 |

Score = 0+0+2*1/√6+3*2/√6 ≈ 3.27