BDK12 Exercises

KEY

Use the file BDK12\_Exercise.xlsx to complete the following questions. This contains 10 “documents,” each of which has a title, a sentence of text, and some MeSH indexing terms. These documents are from \_[cite]\_\_\_.

1. Complete an inverted table of words. That is, for each of the words in the collection, record its document frequency (DF), collection frequency (CF), inverse document frequency (IDF), the document number(s) in which the term appears, the and the associated term frequency (TF) for each of those documents. Use the table in the “Inv Word File” sheet of the Excel document to.

**Document frequency:** number of documents in which the word appears

**Collection frequency:** number of total times the word appears (in all documents)

**Inverse document frequency:** a measure of how common the word is within the collection. Calculated as

**Term frequency:** number of times the word appears in the given document.

1. In your own words, define the meaning of the product TF\*IDF. For each document, which word(s) has the highest value of TF\*IDF? Which has the lowest? What is your interpretation of these findings? Do you think this TF\*IDF is a good way to weight terms?

TF\*IDF is a measure of how important a word is in a document from a collection of documents.

Highest:

Doc 1: Hypertension (2.44) Doc 6: Term (4.00)

Doc 2: Complication, Failure (3.04) Doc 7: Cardiovascular, Raise (4.00)

Doc 3: Goal, Reduce (4.00) Doc 8: Glaucoma, Intraocular (4.00)

Doc 4: Usually (3.40) Doc 9: Heart (4.56)

Doc 5: Drug (3.04) Doc 10: Heart (6.08)

Lowest:

Doc 1: Blood (1.15) Doc 6: Beta (1.40)

Doc 2: Blood (1.15) Doc 7: Beta, Blocker (1.40)

Doc 3: Blood (1.15) Doc 8: Hypertension, Pressure (1.22)

Doc 4: Age, Including, Life, Pharmacologic (1.70) Doc 9: Blood (1.15)

Doc 5: Blood (1.15) Doc 10: Blood (1.15)

1. Suppose you are querying an IR system and documents 3, 4, 5, 6, and 8 have been deemed relevant. If the system retrieves documents 2, 3, 4, and 7, what is the recall of the search? What is the precision?

Recall = 2/5, precision = 2/4

1. Suppose you searching the database with the query “hypertension AND treatment”. Assuming the IR system is using a simple text search, which documents will be retrieved? What if you searched “hypertension OR treatment”?

Hypertension AND treatment retrieves docs 3, 5, and 8.

Hypertension OR treatment retrieves docs 1, 2, 3, 4, 5, 8, and 10.