C4M: Medical Document Retrieval Worksheet

Sample documents:

fever
rash
fever

a.html

fatigue
dizzy
nauseated
fatigue

b.html

rash
red
itchy
weak

c.html

pain
swelling
rash
fever
rash

d.html

Term frequency / inverse document frequency:

$$TF(k,d) = \begin{cases} 0 & \text{if } k \text{ does not appear in } d \\ 1 & \text{otherwise} \end{cases}$$

$$\text{IDF}(k,D) = \log \frac{\text{No. of docs in } D}{\text{No. of docs in } D \text{ that contain } k}$$

$$\text{TF-IDF}(k,d,D) = TF(t,d) * IDF(k,D)$$

k represents a keyword, d represents a document, and D represents the set of all documents.

1. Query: nauseated
   (a) $\text{IDF}(k,D) =$

   (b) Complete the `doc_to_score` dictionary, where each key represents a document and each value represents that document's TF-IDF score.

   ```
   doc_to_score = {'a':              ,
                   'b':              ,
                   'c':              ,
                   'd':               }
   ```

   (c) Best match:

2. Query: fever rash
   (a) $\text{IDF}(k_1,D) =$

   $\text{IDF}(k_2,D) =$

   (b) Complete the `doc_to_score` dictionary, where each key represents a document and each value represents that document's TF-IDF score.

   ```
   doc_to_score = {'a':              ,
                   'b':              ,
                   'c':              ,
                   'd':               }
   ```

   (c) Best match: