Term Frequency-Inverse Document Frequency

- generates value for each topic of document
  - documents are [[Vectors as KR]]
    - each topic as vector element
  - term frequency
    - based on how often term occurs in document d_i
    - normaliced with document length
    - $tf_i = \frac{\#term\_occurences}{\#terms}$
  - inverse document frequency
    - based on how many documents contain term
    - $idf_i = log_10(\frac{|D|}{\#documents\_containing\_this\_term})$
  - TFIDF
    - combination of
      - term frequency
      - inverse document frequency
    - element-wise multiplication of tf_i and idf
- example

Document 1: „I love sun!“
Document 2: „I hate sun!“
Document 3: „I love rain!“
Query: „Does someone else love the sun?“

$$dict = \begin{pmatrix} i \\ love \\ hate \\ sun \\ rain \end{pmatrix}$$

Step 2: Inverse document frequency

$$idf = \begin{pmatrix} 0 \\ 0.18 \\ 0.48 \\ 0.18 \\ 0.48 \end{pmatrix}$$

Step 3: normalised term frequency vectors for document and query

$$t_1 = \begin{pmatrix} 0.33 \\ 0.33 \\ 0 \\ 0.33 \\ 0 \end{pmatrix} \quad t_2 = \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \end{pmatrix}$$

$$t_3 = \begin{pmatrix} 0.33 \\ 0.33 \\ 0 \\ 0 \\ 0.33 \end{pmatrix} \quad q_t = \begin{pmatrix} 0 \\ 0.17 \\ 0 \\ 0.17 \\ 0 \end{pmatrix}$$

Step 4: TFIDF vectors for document and query

$$idf = \begin{pmatrix} 0 \\ 0.18 \\ 0.48 \\ 0.18 \\ 0.48 \end{pmatrix}$$

$$tfidf_1 = \begin{pmatrix} 0 \\ 0.0594 \\ 0 \\ 0.0594 \\ 0 \end{pmatrix} \quad tfidf_2 = \begin{pmatrix} 0 \\ 0 \\ 0.1584 \\ 0.0594 \\ 0 \end{pmatrix} \quad tfidf_3 = \begin{pmatrix} 0 \\ 0.0594 \\ 0 \\ 0 \\ 0.1584 \end{pmatrix} \quad q_{tfidf} = \begin{pmatrix} 0 \\ 0.0306 \\ 0 \\ 0.0306 \\ 0 \end{pmatrix}$$