Consider the query:

**Query:** school closed rain

Also, consider the following **stop word** list for Question 1 and 2 only:

[the, to, is, and, in, has, not]

1. (3 points) What are the similarity scores of the query with each document given above using Jaccard coefficient if there are no stop words? What are the similarity scores if we use the stop words?

2. (5 points) The query is converted to a unit vector using tf-idf weighting

\[ w_{t,d} = tf_{t,d} \times \log_{10}(N/df_t) \]

and Euclidean normalization. The documents are converted to unit vectors using raw tf weighting and Euclidean normalization. Now 1) compute and report the cosine similarity of the query with each document if there are no stop words; 2) Report the result when we use the stop words shown above. For this question, briefly show how you arrive at the answer. (Hint: Think about whether you can calculate the answer without computing all the tf-idf values.)

3. (5 points) We will now compute the similarity scores for each document with the query using a type of overlap score measure defined as follows: Each document is represented as a term frequency (tf) vector which is normalized using the maximum tf formula:

\[ 0.25 + \left[ 0.75 \times \frac{tf_{t,d}}{\max_t(tf_{t,d})} \right] \]

The score of a document with respect to a query is then computed as: the sum of the normalized tf values of the query terms in the document vector. Now compute and report the similarity of the query with each document using the stop word list:

[the, to, is, and, in, not]

4. (5 points) Remember that given a query \( q \) and documents \( d_1, d_2, \ldots \), in addition to ranking the documents in order of decreasing cosine similarities, we may also rank the documents \( d_i \) in order of increasing Euclidean distance from \( q \). Euclidean distance between two vectors \( \vec{x} \) and \( \vec{y} \) can be calculated as:

\[ |\vec{x} - \vec{y}| = \sqrt{\sum_i (x_i - y_i)^2} \]

Show that if \( q \) and the \( d_i \) are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities. (Hint: Write out \( |\vec{q} - \vec{d}|^2 \).)