**NLP – Cosine Similarity Workshop**

**Objective**

To practice using Cosine Similarity as a tool for performing searching for relevant documents based on a given query string

**Exercise**

You are given a text file called “quotes.txt” that contains 51 quotes. Given a query string, list out the most relevant quotes.

First, treat each quote as a separate document and compute TFIDF values for each word in the 51 documents (remember to treat each line of quote as a separate document; your corpus is the 51 documents).

Next, treat the query string as a separate document and compute its TFIDF values as well.

Then, perform Cosine Similarity between the query string (treat it as a document) and the 51 documents.

A cosine similarity **greater than 0** is regarded as a **relevant document** for the query string. **Rank the documents** (i.e. lines of quote) by their cosine similarity scores (descending order - higher cosine similarity first)

For pre-processing, remember to **remove stop-words** from your corpus and **stem each word** in the text.

Use the query string – “**life wise choices**” to test your results.

Given the above query string, your results should be:

Text, letter

Description automatically generated