Assignment 2 in Information Retrieval

In this assignment, you should build a simplified similarity detection system. This goal is to implement **a Jaccard similarity** function that measures the similarity between sentences/documents and provide a similarity score based on how similar the sentences are compared to each other.

Then use these scores to return document IDs to the user sorted by most similar.

Jaccard Similarity is calculated using this rule:

$$J(A, B) = |A \cap B| \ / \ |A \cup B|$$

Example:

- **Query:** idea of March
- **Doc1:** Ceaser died in March
- **Doc2:** the long March
- J(q,Doc1) = 1/6=0.1667
- J(q, Doc2) = 1/5=0.2

Output:

Doc2 0.2

Doc1 0.1667

You can use the intersection function from the previous Assignment. You can use the same data files for testing your function.

General Notes:

1. Students will form a team of **3~4 members** from the same group.
2. Please don't forget the naming convention: **ID1_ID2_ID3_ID4_Group**, for example: 20180010_20180020_20180030_20180040_S10