IR Assignment 2 Report
Group 47

Q1:

Jaccard Coefficient

- Pros: Faster and requires less space as we don't need to make the size of query vector equal to that of document vector
- Cons: Does not take into account important metrics like term frequency and ordering of words

TF-IDF Matrix

- Pros: Takes into account term frequency of document and query. Terms occurring highly in a small number of documents are given more weightage and less weightage to terms occurring in almost all documents.
- Cons: Slower and requires more space as we have to make the size of query equal to length of document. Doesn't take into account position of term in document, semantics and co-occurrences in different documents.

Q3:

Accuracy across given train:test ratios are high. Comparatively for k=10, there isn't a linear pattern in accuracy vs train:test ratios.