The project has two parts:

First part :

1. Read 10 files (.txt)
2. Apply tokenization
3. Apply stemming

Second part :

1. Build positional index & display each term as the following :

<***term***, number of docs containing ***term***; ***doc1***:

position1, position2 … ;

doc2: position1, position2 … ;

etc.>

2. Compute term frequency for each term in each document & display it.

Term	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
Term1										
Term2										
Term3										

3. Compute IDF for each term & display it.
4. Compute TF.IDF matrix for each term & display it.
5. Allow users to write phrase query on positional index and system returns the matched documents for the query.
6. Compute similarity between the query and matched documents.
7. Rank documents based on similarity score.