Preprocessing Steps:
Text conversion to lowercase.
Tokenization using nltk.
Removal of stop words using nltk.
Special characters excluding alphanumeric are removed.
All singly occurring characters are removed.
Finally a set of all the words is created.
Assumptions:
Input Query is case insensitive.

PROS And CONS
JACCARD COEFFICIENT
Pro → A and B don't have to be the same size. A and B.
Both continuous and categorical variables may be used.
Cons → Doesn't work efficiently with nominal data.
It doesn't consider term frequency
It doesn't consider the fact that rare terms in a collection are more informative than frequent terms.

COSINE SIMILARITY
Pro → its low-complexity, especially for sparse vectors: only the non-zero dimensions need to be considered
The cosine similarity is beneficial because even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity

Con → When plotted on a multi-dimensional space, the cosine similarity captures the orientation (the angle) of the data objects and not the magnitude.

TF-IDF
Advantages:

- Easy to compute

- basic metric to extract the most descriptive terms in a document

Disadvantages:

- TF-IDF is based on the bag-of-words (BoW) model, therefore it does not capture position in text, semantics, co-occurrences in different documents, etc.

- Cannot capture semantics

Question 3.
2.

The total number of files will be 1!*17!*26!*59!

maxDCG Using: $rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i+1)}$

20.989750804831445

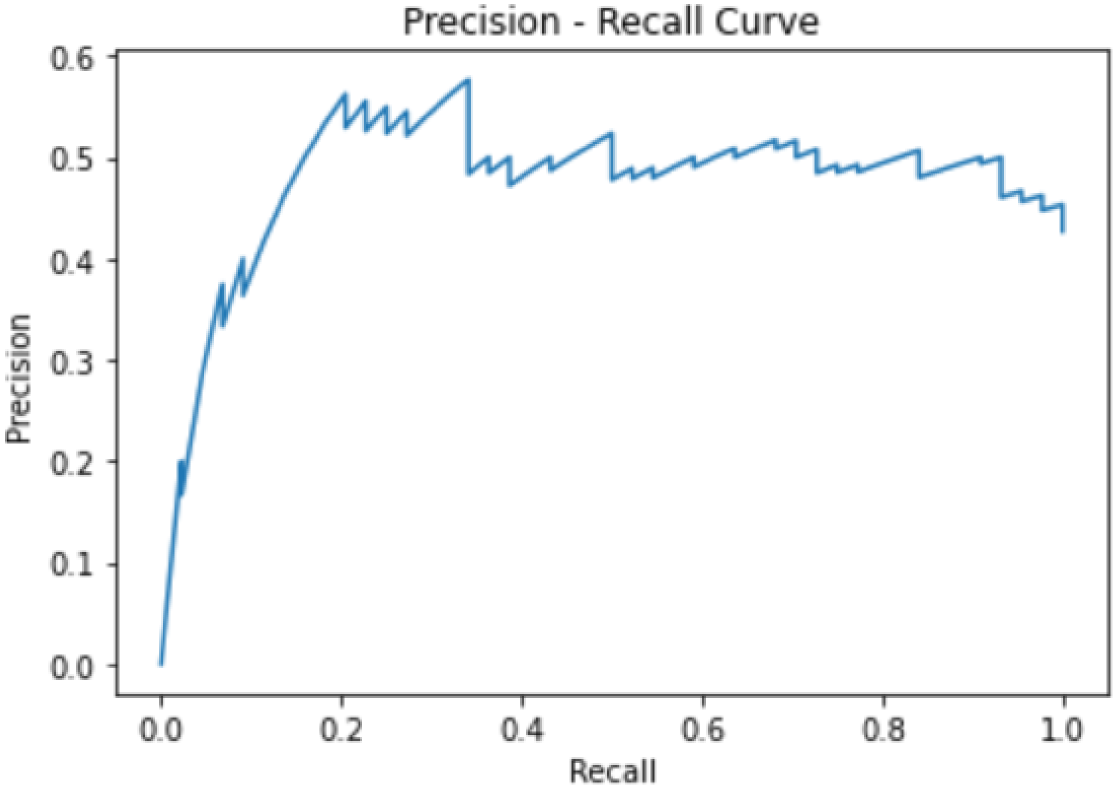
maxDCG Using Alternative Formula : $\sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$

28.98846753873482

3.

nDCG at 50: 0.3521042740324887
nDCG whole Dataset: 0.5979226516897831

4.


Precision - Recall Curve
Precision
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0