Question 1.

Sample query outputs on a subset of the original data given:

```
Enter Query
good probability value
['good', 'probability', 'value']

Enter K
5
Candidates: {'alt.atheism/53076', 'alt.atheism/53117', 'alt.atheism/53789',
'comp.graphics/38976', 'alt.atheism/53800', 'alt.atheism/53564', 'alt.atheism/53100',
'alt.atheism/51137', 'alt.atheism/51316', 'alt.atheism/51122', 'alt.atheism/53639',
'alt.atheism/53237', 'comp.graphics/39052', 'alt.atheism/53246', 'alt.atheism/53813'}

Cosine Similarity based results:
 ['alt.atheism/53246', 'alt.atheism/53100', 'alt.atheism/53117', 'alt.atheism/53237',
'alt.atheism/53813']
```

```
Enter Query
difficult brains
['difficult', 'brain']

Enter K
7
Candidates: {'comp.graphics/38769', 'alt.atheism/54250', 'alt.atheism/53501',
'alt.atheism/54182', 'comp.graphics/39626', 'alt.atheism/53194', 'alt.atheism/53126',
'comp.graphics/37920', 'comp.graphics/38636', 'alt.atheism/53499', 'alt.atheism/53794',
'alt.atheism/53323', 'alt.atheism/53491', 'comp.graphics/37919'}

Cosine Similarity based results:
 ['comp.graphics/37919', 'alt.atheism/53194', 'alt.atheism/53126', 'comp.graphics/
37920', 'alt.atheism/53794', 'alt.atheism/53499', 'alt.atheism/54250']
```

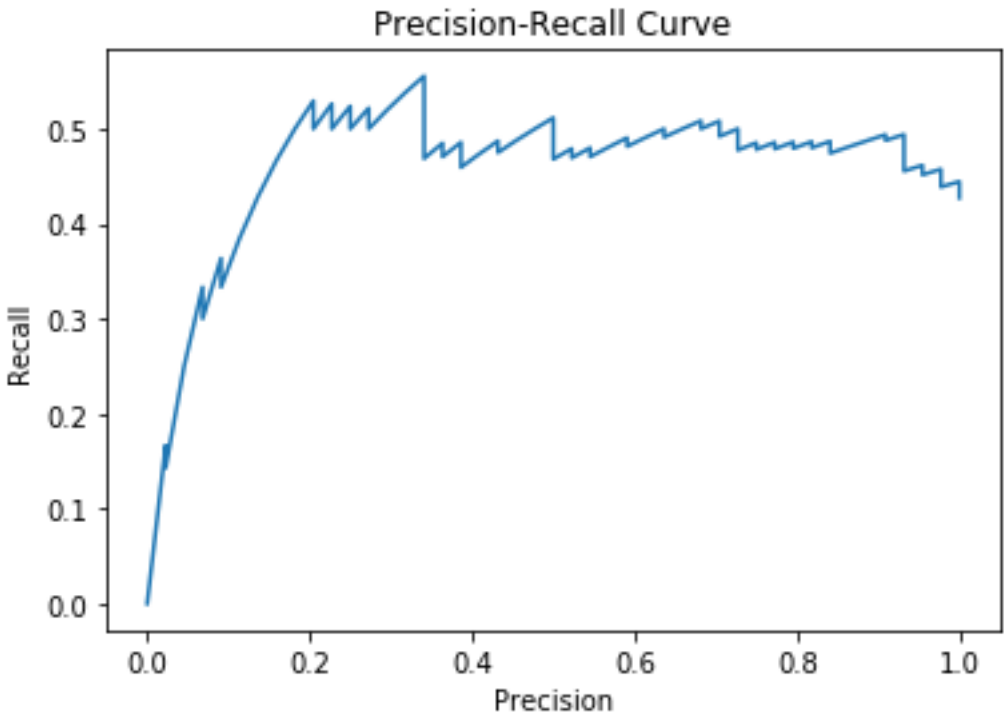
The value of r was checked at intervals of 10, from 10 to 100 and the following observations have been noted:

- If r << k: In this case the results are poor as the champion lists only have very few elements in them and the low lists are used. The meaning of pruning the entire set is lost.
- If r < k: Given the lack of ground truth, exact metrics are not present but on manual comparison of query terms in the files output, the results are not very fitting given presence of better documents in the collection.
- If r > k: This range gave relevant results with good term frequency of query terms in the document along with the relevance score, the champion lists are utilized well. If this value of r is further increased then the point of trimming the entire corpus is lost.

Question 2.

Total URLs with qid:4: **103**

Total max DCG URL files possible: **26078**



P-R Curve Obtained on dim 75

NDCG (50): 0.3521042740324887

NDCG: 0.5979226516897831

Question 3.

1. Explain the relationship between ROC curve and PR curve.

In **ROC space**, **FPR** is plotted on the x-axis and **TPR** is plotted on the y-axis. In **PR space**, **Recall** is plotted on the x-axis and **Precision** is plotted on the y-axis. The formulas for the said terms are presented below.

CONFUSION MATRIX:

	Classified Positive	**Classified Negative**
Actual Positive	True Positives (TP)	False Positives (FP)
Actual Negative	False Negatives (FN)	True Negatives (TN)

Recall *= TP / (TP + FN)*

Precision *= TP / (TP + FP)*

True Positive Rate (TPR) *= TP / (TP + FN)*

False Positive Rate (FPR) *= FP / (FP + TN)*

A unique confusion matrix corresponds to a fixed dataset for both points in ROC space and points in PR space provided the fact that recall is not zero.

A one to one mapping can be constructed between confusion matrices and points in PR space as well as for points in ROC space. Hence there also exists a mapping between points of PR space and ROC space. Therefore, a curve in ROC space can be derived from a curve in PR space and vice versa.

2. Prove that a curve dominates in ROC space if and only if it dominates in PR space.

Let us consider two curves **c1** and **c2**. The claim is that if curve *c1* dominates curve *c2* in ROC space then it would dominate the same curve in PR space as well.

Proof by contradiction.

Case 1. If a curve dominates in ROC space then it dominates in PR Space

Assumption: *c1* dominates in ROC space but does not dominate in PR space.

Since *c2* dominates in PR space, there must exist a point A on *c2* with identical recall as point B on *c1* where precision of (c2, A) is greater than precision of (c1, B),
i.e.

$$\text{Recall (c1, A)} = \text{Recall (c2, B)}$$

$$\text{Precision (c2, A)} > \text{Precision (c1, B)}$$

The recall values for both the points on different curves are same, and since we can see from above mentioned formulas that **Recall** and **True Positive Rate (TPR)** are one and the **same**.

Since the curve *c1* dominates *c2* in ROC space, we get:

$$\text{FPR (c1, B)} >= \text{FPR (c2, A)}$$

The value of TPRs is same, so

$$\text{TPR (c1, B)} = \text{TPR (c2, A)}$$

$$\text{TP (c1, B) / Total Positives} = \text{TP (c2, A) / Total Positives}$$

Here denominator is the same as total positives and negatives would stay constant.

Therefore,

$$\text{TP (c1, B)} = \text{TP (c2, A)} = \text{TP}$$

We have the FPR relation as well, which gives:

$$\text{FPR (c1, B)} > \text{FPR (c2, A)}$$

$$\text{FP (c1, B) / Total Negatives} > \text{FP (c2, A) / Total Negatives}$$

$$\text{FP (c1, B)} > \text{FP (c2, A)}$$

We put this relation into the Precision formula, which gives:

TP / FP (c1, B) + TP < TP / FP (c2, A) + TP

Precision (c1, B) < Precision (c2, A)

Which is a direct contradiction to our assumption. Hence this case is not possible.

<u>Case 2.</u> If a curve dominates in PR space then it dominates in ROC space

Assumption: *c1* dominates in PR space but does not dominate in ROC space.

Since *c1* does not dominate in ROC space, there must exist a point A on *c2* with identical TPR as point B on *c1* where FPR of (c2, A) is less than TPR of (c1, B),
i.e.

TPR (c1, B) = TPR (c2, A)

Recall (c1, B) = Recall (c2, A)

FPR (c2, A) < TPR (c1, B)

Since PR space is the area of domination of c1,

Precision (c2, A) <= Precision (c1, B)

Due to similarity of recalls,

TP (c1, B) = TP (c2, A) = TP

We put this relation into the Precision formula, which gives:

TP / FP (c1, B) + TP < TP / FP (c2, A) + TP

We find,

FP (c1, B) <= FP (c2, A)

Now, we have

FP (c1, B) / Total Negatives > FP (c2, A) / Total Negatives

Which implies,

FPR (c1, B) <= FPR (c2, A)

Which is in complete contradiction with our original assumption. Hence this case is not possible as well.

Hence Proved.

3. <u>It is incorrect to interpolate between points in PR space. When and why does this happen? How will you tackle this problem?</u>

The interpolation process is fairly simple in the ROC space and a simple linear equation can suffice to connect two points. However, the same effect cannot be achieved in the PR space due to the fact that Precision and Recall do not linearly depend on one another. A linear interpolation among these variables would be over confident in the results and would always overshoot the true value.

The points in PR space can be achieved by finding a convex hull in the ROC space. We obtain, for each point in the convex hull, an equivalent point in PR space. And thence interpolation can be performed using these newly obtained PR points.

References:

Davis, Jesse & Goadrich, Mark. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, ACM. 06. 10.1145/1143844.1143874.