Q1) Champion List

Output 1

Value of r is 20

Enter a sentence- I am looking to add voice input capability to a user interface

Enter the number of retrieve file- 5

Most similar documents are-

1) 20_newsgroups/alt.atheism\51124
2) 20_newsgroups/comp.os.ms-windows.misc\10608
3) 20_newsgroups/comp.graphics\38510
4) 20_newsgroups/rec.sport.hockey\53628
5) 20_newsgroups/talk.politics.guns\54819

Output 2

Value of r is 10

Enter a sentence- A program in the archive keymap00.zip on simtel and mirror

Enter number of retrieve file- 5

Most similar documents are-

1) 20_newsgroups/alt.atheism\51124
2) 20_newsgroups/rec.sport.hockey\53628
3) 20_newsgroups/comp.os.ms-windows.misc\9518
4) 20_newsgroups/comp.os.ms-windows.misc\10608
5) 20_newsgroups/comp.graphics\38510

Output 3

Value of r is 30

Enter a sentence- See the keyboard directory of simtel for programs that report

Enter number of retrieve file- 5

Most similar documents are-

1) 20_newsgroups/alt.atheism\51124
2) 20_newsgroups/rec.sport.hockey\53628
3) 20_newsgroups/comp.graphics\38510
4) 20_newsgroups/comp.os.ms-windows.misc\10608
5) 20_newsgroups/comp.os.ms-windows.misc\9518

Output 4

Value of r is 25

Enter a sentence- if you have a compiler to create a new keyboard map

Enter number of retrieve file- 5

Most similar documents are-

1) 20_newsgroups/alt.atheism\51124
2) 20_newsgroups/rec.sport.hockey\53628
3) 20_newsgroups/comp.os.ms-windows.misc\10608
4) 20_newsgroups/comp.graphics\38510
5) 20_newsgroups/comp.os.ms-windows.misc\9942

Output 5

Value of r is 15

Enter a sentence- looks like its time to move that juvenile public

Enter number of retrieve file- 5

Most similar documents are-

1) 20_newsgroups/alt.atheism\51124
2) 20_newsgroups/rec.sport.hockey\53628
3) 20_newsgroups/comp.graphics\38510
4) 20_newsgroups/comp.os.ms-windows.misc\9518
5) 20_newsgroups/comp.os.ms-windows.misc\10608

Q2) DCG

i)

Number of possible sequence with max DCG- 5.407613242151097e+121

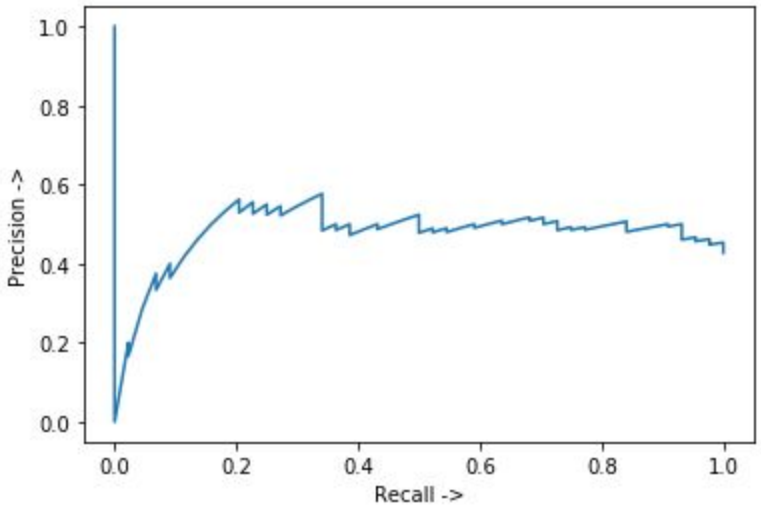
ii)

For query id 4-

nDCG value upto 50- 0.37071213897397365
DCG value upto 50- 7.19450227631398
IDCG value upto 50- 19.407247618668023

nDCG value for whole file- 0.6357153091990775
DCG value for whole file- 12.337484420604602
IDCG value for whole file- 19.407247618668023

iii)



Q3) ROC vs PR

i) The relationship between the ROC curve and the PR curve-

ROC and PR curves are generated to show the performance of an algorithm on a dataset. Any dataset contains a fixed number of positive and negative examples. In the paper[1], Authors showed a deep relation between the ROC and the PR curve-

I. For any dataset of positive and negative examples, there will a one-to-one resemblance between a curve in ROC space and a curve in PR space, such that the curves contain exactly the same confusion matrices if Recall != 0.
II. For a fixed number of positive and negative examples, one curve dominates the second curve in ROC space if and only if the first dominates the second in PR space.

ii) A curve dominates in ROC space if and only if it dominates in PR space-

i) **<u>If a curve dominates in ROC space then it dominates in PR space</u>**

Proof by contradiction. Suppose we have two curves. Now curve I dominate in ROC space. yet, once we translate these curves in PR space, curve I no longer

dominates. It means there must a point X on curve II and point Y on the curve I with identical Recall has lower Precision. In simple word, PRECISION(X) > PRECISION(Y) yet RECALL(X) = RECALL(Y). Since RECALL(X) = RECALL(Y) and Recall is identical to TPR, we have that TPR(X) = TPR(Y).
As curve, I dominates curve II in ROC space FPR(X) ≥ FPR(Y) and total positives and total negatives are fixed and since TPR(X) = TPR(Y):

TPR(X) = TPX /Total Positives
TPR(Y) = TPY/Total Positives

we now have TPX = TPY and thus denote both as TP and FPR(X) ≥ FPR(Y):

FPR(X) = FPX/Total Negatives
FPR(Y) = FPY/Total Negatives

This implies that FPX ≥ FPY because

PRECISION(X) = TP/(FPX + TP)
PRECISION(Y) = TP/(FPY + TP)

we now have that PRECISION(X) ≤ PRECISION(Y). But this contradicts our original assumption that PRECISION(X) > PRECISION(Y).

ii) **<u>If a curve dominates in PR space then it dominates in ROC space</u>**

Proof by contradiction. Suppose we have two curves. Now curve I dominate in PR space. yet, once we translate these curves in ROC space, curve I no longer dominates. , there exists some point X on curve II such that point Y on curve I with identical TPR yet FPR(X) < TPR(Y). Since RECALL and TPR are the same, we get that RECALL(X) = RECALL(Y). Because curve I dominates in PR space we know that PRECISION(X) ≤ PRECISION(Y) and RECALL(X) = RECALL(Y):

RECALL(X) = TPX / Total Positives
RECALL(Y) = TPY / Total Positives

We know that TPX = TPY, so we will now denote them simply as TP. Because PRECISION(X) ≤ PRECISION(Y):

PRECISION(X) = TP / (TP + FPX)

$$PRECISION(Y) = TP / (TP + FPY)$$

we find that FPX ≥ FPY. Now we have

$$FPR(X) = FPX / Total\ Negatives$$
$$FPR(Y) = FPY / Total\ Negatives$$

This implies that FPR(X) ≥ FPR(Y) and this contradicts our original assumption that FPR(X) < FPR(Y).

iii) Interpolate between points in PR space-

It is straightforward to interpolate between points in ROC space by simply drawing a straight line connecting the two points. However, in the Precision-Recall space, interpolation is more complicated. As the level of Recall varies, the Precision does not necessarily change linearly due to the fact that FP replaces FN in the denominator of the Precision metric. In these cases, linear interpolation is a mistake that yields an overly-optimistic estimate of performance. We can use the method proposed by Goadrich et al. (2004) to approximate the interpolation between two points in PR space.

References:

1. https://www.biostat.wisc.edu/~page/rocpr.pdf
2. Goadrich, M., Oliphant, L., & Shavlik, J. (2004). Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction. Proceedings of the 14th International Conference on Inductive Logic Programming (ILP). Porto, Portugal