Analysis(Assignment 3)

(MT19111) Akshaj Patil.

Q1]Query:- gnuplot, etc. make it easy to plot real valued functions of 2 variables

Enter number of docs you wanna retrieve 10

O/P when r=20:-

1481 : and Score= 0.44606037459332903
1002 : and Score= 0.3975736835922318
4533 : and Score= 0.3469396543652538
202 : and Score= 0.30615771225297905
16001 : and Score= 0.2653708650097868
1692 : and Score= 0.2653085884979002
1983 : and Score= 0.26128051993140683
10203 : and Score= 0.22449023117696637
1729 : and Score= 0.2203253181895964
11102 : and Score= 0.18371930589460042

O/P when r=25:-

1002 : and Score= 0.45039709887293766
1481 : and Score= 0.44606037459332903
4533 : and Score= 0.3469396543652538
202 : and Score= 0.30615771225297905
16001 : and Score= 0.2653708650097868
1692 : and Score= 0.2653085884979002
1983 : and Score= 0.26128051993140683
10203 : and Score= 0.22449023117696637
1729 : and Score= 0.2203253181895964
11102 : and Score= 0.18371930589460042
257 : and Score= 0.18309045166422405

O/P when r=13:-

1481 : and Score= 0.44606037459332903
1002 : and Score= 0.3975736835922318
202 : and Score= 0.30615771225297905
16001 : and Score= 0.2653708650097868
1692 : and Score= 0.2653085884979002
1983 : and Score= 0.26128051993140683
10203 : and Score= 0.22449023117696637
1729 : and Score= 0.2203253181895964
257 : and Score= 0.18309045166422405
1246 : and Score= 0.17589059079324287

Here the query is taken from doc 1002 but here g(d) score of doc 1481 is high than g(d) score of doc 1002. Therefore doc 1481 will get more preference over doc 1002.

When r=25 code performs better than r=20 and r=13.

One way to decide value of 'r' is static initialization to 25. Because as per the google heuristic user mostly retrieves 20 documents as there can be false positive so we consider 'r' value little greater than it i.e 25. (25 is chosen because as compare to 15,20 it gives better result.)

Another way is.

Here we have chosen 'r' on basis of 'tf' values of that term in each document. For rare words value of 'r' should be high so that there will be more docs in high-list of rare words. And value of 'r' will be low when word is frequent so that there will be less number of documents in high-list of frequent words. So when query comes then by above logic we are giving more preference rare words so that valid document will be retrieved.

```
2519 : and Score=  0.5718127084504578
16729 : and Score=  0.5308594663033607
18904 : and Score=  0.5114900249472986
19645 : and Score=  0.48997564611985317
19414 : and Score=  0.489942625227737
8852 : and Score=  0.4695636709317276
10455 : and Score=  0.46942620249053457
17175 : and Score=  0.4694194806635896
1240 : and Score=  0.46941879223533234
4802 : and Score=  0.44948519353465477

Using this variable 'r' there are many false positives retrived.
Query is best performed on static 'r' with value 25.
```

Q2:]

o/p:

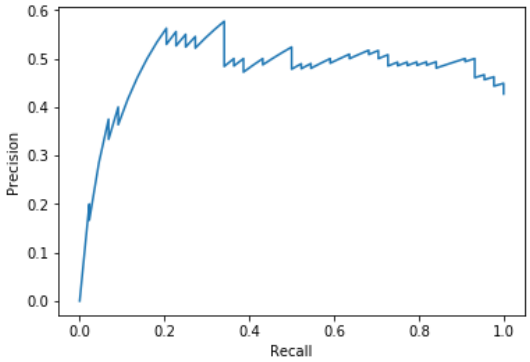
1:] Here file is uploaded in order of url which generate max DCG.

```
Number of combinations from which we get maxDcg are 5.4076132421510985e+121
```

2 :]

```
IDcg = 28.98846753873482
ndcg-@-50 = 0.35612494416255847
ndcg-total = 0.5784691984582591
```

3:] Here we can see that as precision increases recall increases till some point. Then after that as recall increases there is only slight change precision.



[**Q3] Assignment 3:**

1:]Relation Between ROC Curve and PR curve.

Ans:-> With fixed dataset ROC curve determines fixed confusion matrix. As we do not consider True Negative(TN) in PR space so each point can correspond to multiple confusion matrix. But when Recall is not equal to zero then with other three entries we can uniquely determine TN. So here we have one-to-one mapping between confusion matrices and points in PR space. And also we have one-to-one mapping between confusion matrices and points in ROC space. Hence we can translate curve from ROC space to PR space and vice versa.

2:] Prove that curve dominates in ROC space if and only if it dominates in PR Space.

Let us consider two curves curve 1 and curve 2.

Claim 1:If a curve dominates in ROC space then it dominates in PR space

Let us assume Curve 1 dominate in ROC Space and do not dominate in PR space.

Since curve 1 does not dominate in PR curve then there will be one point A on curve 2 whose precision is greater than one point B on curve 1, but both point have same Recall.

i.e PRECISION(A) > PRECISION(B) and RECALL(A) = RECALL(B).

since Recall of both points are equal therefore TPR(A) = TPR(B).

FPR(B) = False Positive(B) / Total Negatives

FPR(A)= Fasle Positive(A)/ Total Negatives

Now here curve 1 dominates curve 2 in ROC therefore FP(A) ≥ FP(B) because Total Negatives are constant.

PRECISION(A) = TP/(FPA + TP)

PRECISION(B) = TP/(FPB + TP)

Now here we can see that PRECISION(A)<=PRECISION(B). but this contradicts our assumption.

Claim 2 : If a curve dominates in PR space then it dominates in ROC space.

Let us assume that curve 1 dominates in PR space but does not dominate in ROC space.

Point A is on curve 2 and point B is on Curve 1. Because curve I dominates in PR space we know that PRECISION(A) ≤ PRECISION(B)->(I) , but RECALL(A)=RECALL(B).

PRECISION(A) = TP /(TP + FP(A))

PRECISION(B) = TP/(TP + FP(B))

By (I) we can say that FP(A)>=FP(B)

Now

FPR(A) = FPA/ Total Negatives

FPR(B) = FPB/ Total Negatives

From above equation we can say that FPR(A) ≥ FPR(B) and this contradicts our original assumption that FPR(A) < FPR(B).

Therefore we can say that if curve dominates in ROC space if and only if it dominates in PR space.

Q3:]It is incorrect to interpolate between points in recall space because as in different levels recall varies but it Precision does not necessarily change. This is because FP is considered in Precision and FN is considered in Recall. In order to tackle this problem, it is better to translate the points to ROC curve(hull in ROC space) and again from ROC curve to PR space. Now, this curve PR space excludes exactly those points which are beneath the convex hull in ROC space.

To interpolate between two points A and B, we must interpolate between their counts TP(A), TP(B), FP(A) and FP(B). For this we find **local skew** defined as:

FP(B)-FP(B)/TP(A)-TP(B)

Now, for all x where 1<=x<=TP(A)-TP(B), we create new points TP(A)+x and calculate corresponding FP by increasing it linearly by local skew. The intermediate points are:

(TP/Total positives, (TP(A)+x)/ ((TP(A)+x+ FP(A)+local skew)

Newly constructed PR space curve is suitable for interpolation.