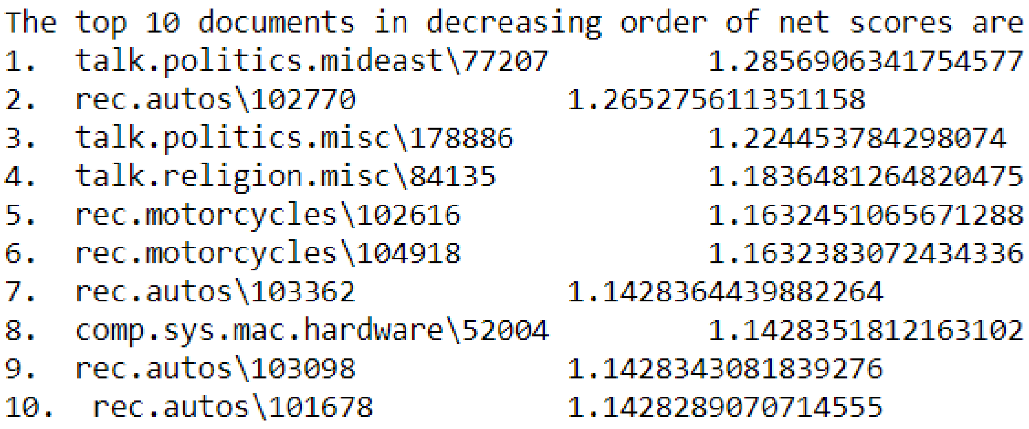
# IR ASSIGNMENT 3 ANALYSIS FILE

Q1.) Taking r = 13

Here r is found by taking the median of the document frequency of all the terms in the vocabulary. While finding the median the doc frequency of 1, 2 , 3 ,4 are not considered because there are many term with these values of doc frequency and finding the median along with these would result in a skewed median.

Query = BMW with the rear wheel spinning wildly and someone groping for the kill switch

K = 10

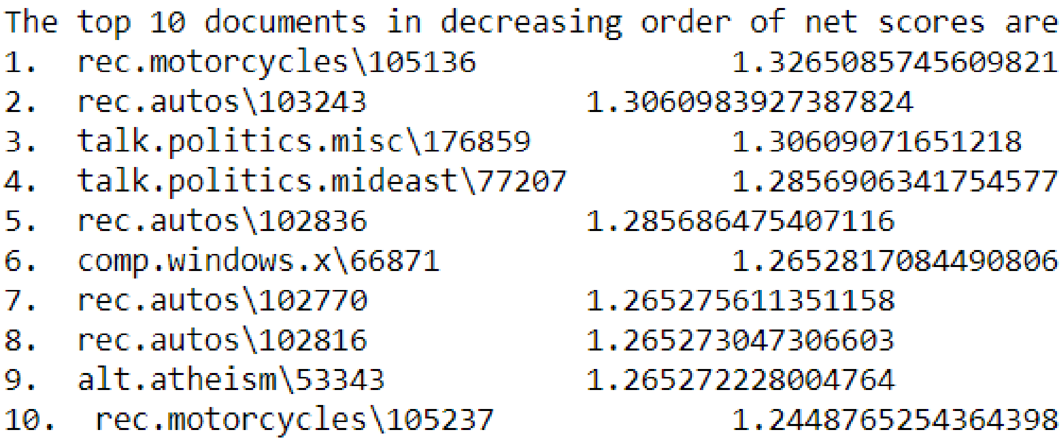


```
The top 10 documents in decreasing order of net scores are
1.  talk.politics.mideast\77207         1.2856906341754577
2.  rec.autos\102770            1.265275611351158
3.  talk.politics.misc\178886           1.224453784298074
4.  talk.religion.misc\84135            1.1836481264820475
5.  rec.motorcycles\102616              1.1632451065671288
6.  rec.motorcycles\104918              1.1632383072434336
7.  rec.autos\103362            1.1428364439882264
8.  comp.sys.mac.hardware\52004         1.1428351812163102
9.  rec.autos\103098            1.1428343081839276
10.  rec.autos\101678           1.1428289070714555
```

Here majority of docs are related to autos and motorcycles folders

Taking r = 42

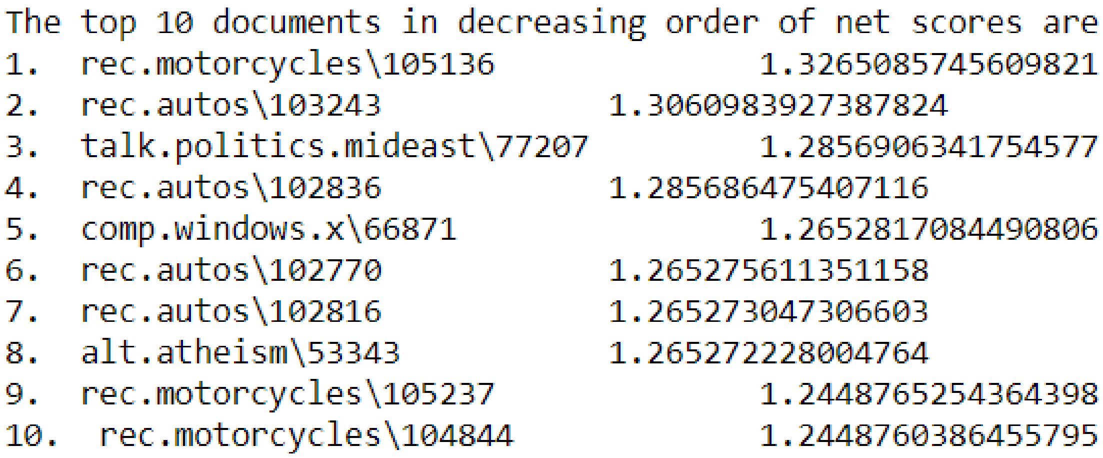
This value of r is the mean of median doc frequency of each of the 20 folders



```
The top 10 documents in decreasing order of net scores are
1.  rec.motorcycles\105136              1.3265085745609821
2.  rec.autos\103243            1.3060983927387824
3.  talk.politics.misc\176859           1.30609071651218
4.  talk.politics.mideast\77207         1.2856906341754577
5.  rec.autos\102836            1.285686475407116
6.  comp.windows.x\66871                1.2652817084490806
7.  rec.autos\102770            1.265275611351158
8.  rec.autos\102816            1.265273047306603
9.  alt.atheism\53343           1.265272228004764
10.  rec.motorcycles\105237             1.2448765254364398
```

Taking r = 36

This value of r is the median of the 20 values corresponding to each folder. These 20 values are the median doc frequency of each of the folders.



```
The top 10 documents in decreasing order of net scores are
1.   rec.motorcycles\105136               1.3265085745609821
2.   rec.autos\103243             1.3060983927387824
3.   talk.politics.mideast\77207          1.2856906341754577
4.   rec.autos\102836             1.285686475407116
5.   comp.windows.x\66871                 1.2652817084490806
6.   rec.autos\102770             1.265275611351158
7.   rec.autos\102816             1.265273047306603
8.   alt.atheism\53343            1.265272228004764
9.   rec.motorcycles\105237               1.2448765254364398
10.   rec.motorcycles\104844              1.2448760386455795
```

Here r = 36 is a good value of r as it has the maximum no. of docs in the top ranks which can be rated as relevant.

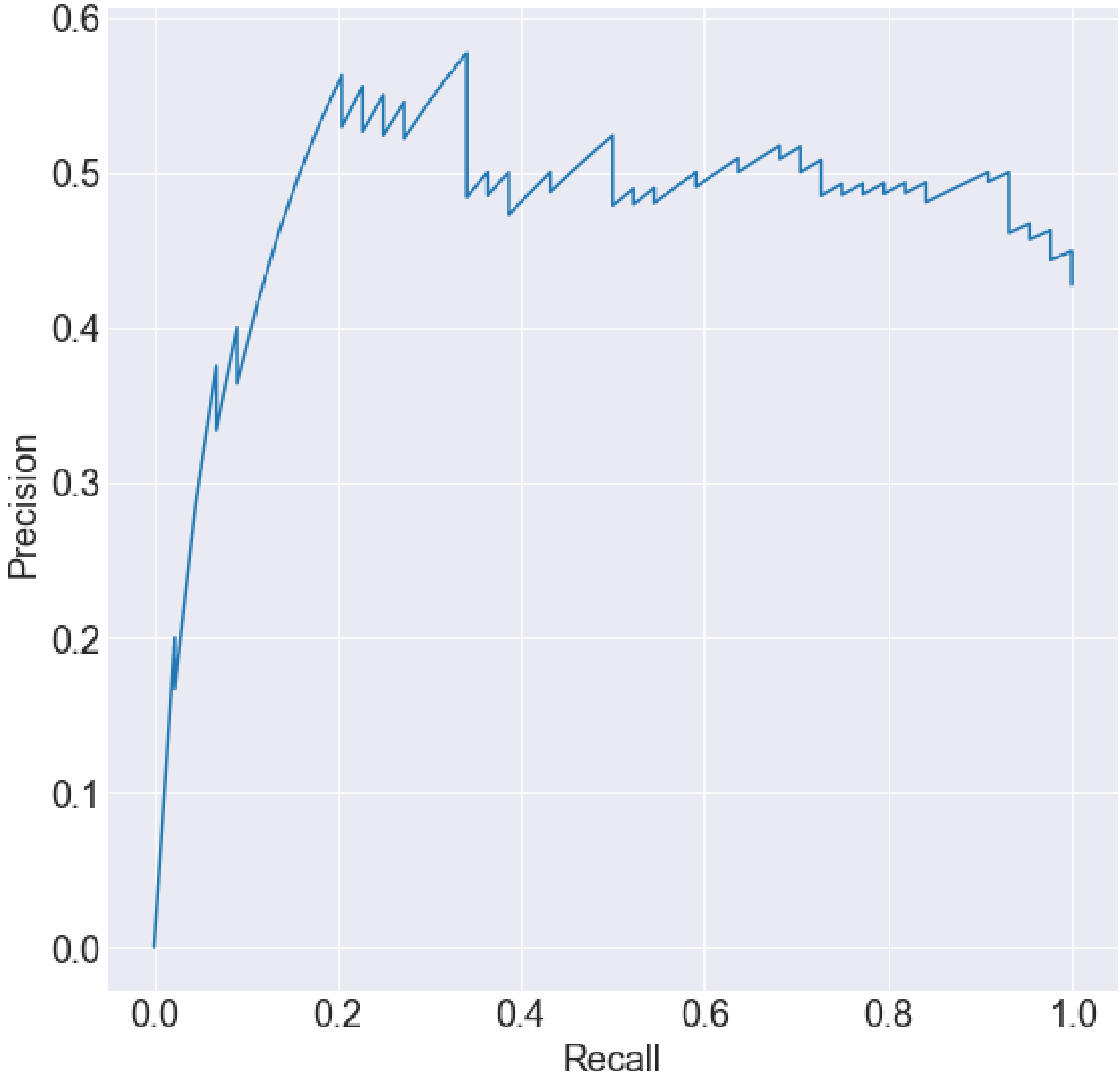
Q2.) Here the no. of files that can be created is:

$$(1! * 17! * 26!) * \sum_{k=0}^{59} \binom{59}{k} k!$$

The nDCG value for qid:4 at 50 is = 0.35612494416255847

The nDCG value for qid:4 for whole dataset is = 0.5784691984582591

The PR curve obtained is:



Q3.)

**1.)** ROC curve takes the False positives on X-axis and True-positive i.e. Recall on Y-axis

PR curve takes Recall on X-axis and Precision values on the Y-axis.

Now for a binary classification problem of classifying samples as positives and negatives. We have a dataset with a fixed number of positives and negatives.

The metrics Precision and recall can be viewed as functions which when applied on the confusion matrix give unique points on the PR curve and the ROC curve.

The different theorems which prove ROC curve and PR curve are very related to each other are:

**1.)** For a given dataset containing a fixed no. of positive and negative samples there exists a one-to-one relationship between the PR curve and the ROC curve. Both the curves plotting the points given by the same confusion matrix provided Recall is not equal to zero.

A point on the ROC curve maps to a unique confusion matrix. Also, while plotting the PR curve the True negatives value is not considered but given the other 3 values and the count of positive and negative samples the true negatives count can be determined uniquely.
This means there exists a one-to-one mapping between the points on PR curve and the confusion matrices. This further implies that there is one-to-one correspondence between PR curve and ROC curve and given one of them the other one can be drawn uniquely.

**2.)** Given a dataset with a fixed number of positive and negative samples one curve dominates the second curve in the ROC space if and only if the first one dominates the second in PR space also.

**3.)** For a set of points in PR space we can make an achievable PR curve which is made using the convex hull of Roc curve. Achievable PR curve dominates all other valid PR curves that could be constructed with these points.

A convex hull for a ROC curve is obtained by following the below criteria:

- Between adjacent points we use linear interpolation
- None of the points on the convex hull should lie above the actual curve
- For any points used to construct the ROC curve the line segment connecting them is equal to or below the convex hull curve.

The values obtained using linear interpolation in PR curve are actually non-achievable. Therefore achievable PR obtained using convex hull in ROC curve is a better way.

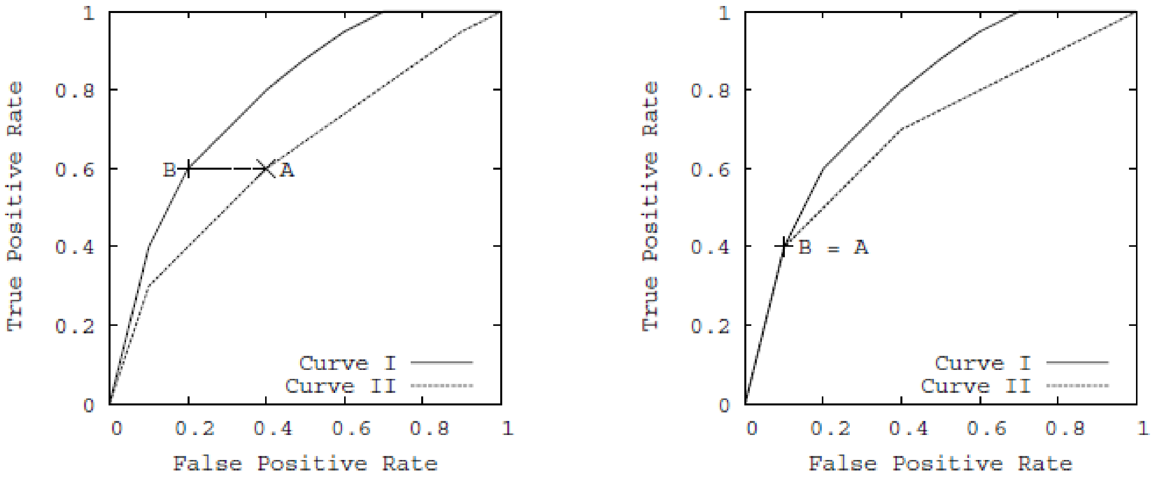
Also, convex hull curve in ROC dominates all other curves that can be obtained by using linear interpolation on those points. Therefore using the theorem of dominance of curves, on converting the points of ROC convex hull to PR space we will obtain a curve that will dominate all other curves obtained using those points in PR space.

Also we can visualize each point on the ROC curve or PR curve as a classifier with a threshold for calling a sample as positive. Now making the convex hull or achievable PR curve is same as making a classifier that picks the best points and thus gives the best values of precision and recall.

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

This prove can be broken into 2 parts:

1.) If a curve dominates in ROC space then it dominates in PR space i.e. (=> relation):
We prove this by contradiction:
Suppose we have curves I and II as shown in below figure:



(a) Case 1: $FPR(A) > FPR(B)$

(b) Case 2: $FPR(A) = FPR(B)$

Here we can see that the curve I dominates in the ROC space. Let us assume that when we translate these curves to the PR space, the curve I no longer dominates curve II. Now this means that there exists points B and A on curves I and II respectively such that Recall(A) = Recall(B) but Precision(A) > Precision(B). Now recall is the same as TPR. As curve I dominates curve II in ROC space FPR(A) >= FPR(B) as can be seen in the figure above. Also Total Positives and Total Negatives are already fixed. TPR(A) = TPR(B):
TPR(A) = TPa / Total Positives
TPR(B) = TPb / Total Positives

This implies: TPa = TPb. Let this value be TP.
Also, FPR(A) >= FPR(B)
FPR(A) = FPa / Total Negatives
FPR(B) = FPb / Total Negatives

This means:
FPa >= FPb

Now,

Precision(A) = TP / (FPa + TP)

Precision(B) = TP / (FPb + TP)

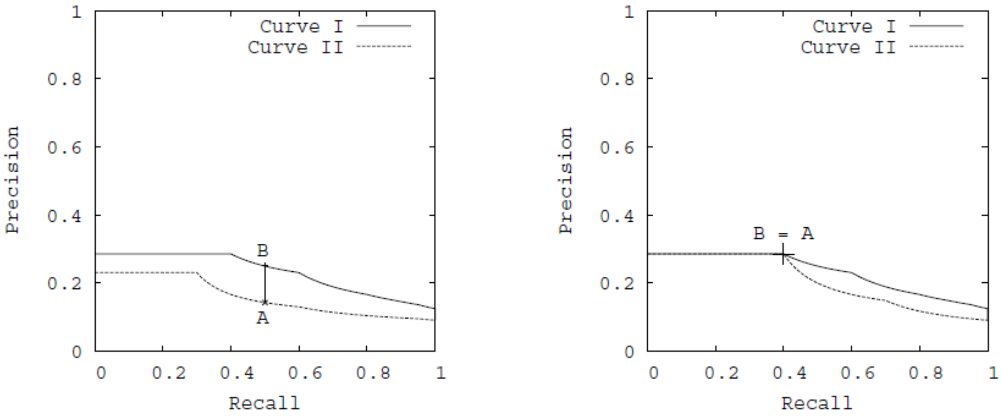
As, FPa >= FPb this implies Precision(A) <= Precision(B)

This contradicts the initial assumption that we had take that:

Precision(A) > Precision(B)

So, curve I dominates the curve II in PR space also. Hence proved

2.) Now proving, if a curve dominates in PR space then it dominates in ROC space:
This prove can also be done by contradiction.
Suppose we have curves I and II shown below:



(a) Case 1: $PRECISION(A) < PRECISION(B)$

(b) Case 2: $PRECISION(A) = PRECISION(B)$

Suppose we have curves I and II as shown in the figure above. We can see that curve I dominates curve II in the PR space.
Let us assume that when translated to ROC space, the curve I does not dominate curve II.
This means that for some points B and A on curves I and II respectively in ROC space, TPR(A) = TPR(B) but FPR(A) < FPR(B).
Now we know that recall is same as TPR.
Thus, Recall (A) = Recall(B).
As, curve I dominates curve II in PR space: Precision(A) <= Precision(B)

Recall(A) = TPa / Total Positives
Recall(B) = TPb / Total Positives

This implies, TPa = TPb. Let us denote TPa andTPb by TP.

Also, Precision(A) <= Precision(B) as curve I is dominating in PR space
Now, Precision(A) = TP / (TP + FPa)
Precision(B) = TP / (TP + FPb)

This implies, FPa >= FPb.

Now, we have:
FPR(A) = FPa / Total Negatives
FPR(B) = FPb / Total Negatives

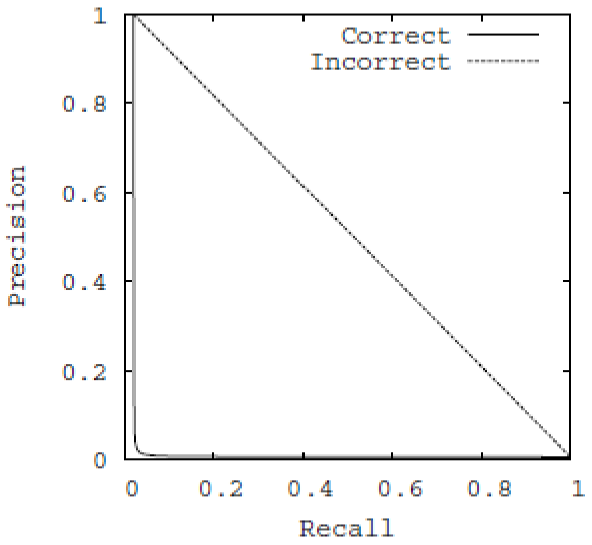
This implies: FPR(A) > FPR(B)
And this is a contradiction of our initial assumption of FPR(A) < FPR(B)
Therefore, curve I dominates in ROC space also

Hence proved for both sides. Therefor proved for if and only if

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

Often area under the curve is used as a simple metric to define how an algorithm performs over the whole space. In PR curve as the level of recall varies, it is not necessary that the precision will also change linearly. The reason being that to get the formula of Precision we replace False negative by false positive in the denominator of formula of recall.
So, if we do linear interpolation here, we get an over-estimate of the performance if we find area under the curve.

Suppose we have the curve below:



Here we have a single point (0.02,1) and it is extended to the endpoints of (0, 1) and (1, 0.008). The dataset contained 433 positives and 56,164 negatives.

Interpolating by using the correct method of using local skew, the AUC-PR is obtained as 0.03, while interpolating using linear connection would give AUC-PR as 0.50 which is an over-estimate.

So, the problem is to find the correct achievable PR curve.

The way to tackle this problem is, is to use local skew to do interpolation, the methos is explained below:

We know that any point A in the PR curve is generated from the underlying TPa and FPa values.

Suppose we have 2 points A and B that are far apart in the PR curve. To find some intermediate points we need to interpolate between their counts TPa, TPb and FPa, FPb. We define a term local skew as the value which denotes the increase in false positives count so as to increase the count of true positives by 1.

Now new points TPa + x can e created for values of x such that : 1<= x <= TPb-TPa

Also corresponding FP can be calculated for each point by incrementing FP by the local skew value.

Local skew = (FPb - FPa) / (TPb - TPa)

Now to get the achievable PR curve, first find the convex hull in ROC curve. Now for each point selected to be included in the hull, the confusion matrix that defines that point is used to find the point in PR curve. Now the correct interpolation method is used between the newly created points to get the achievable PR curve