1) I developed a system that detects tweets about about different candidates in a general election. I have classified a gold-standard set of 200 tweets, 100 of which I identified to be about the election and 100 I classed as being about non-election things.

When I vary the similarity threshold of the system from 1-50 I get different numbers of correct and incorrect answers, that is different numbers of True Positives (TP), False Negative (FN), False Positives (FP) and True Negatives (TN) tweets. For example, when my system correctly identifies the tweets as being about the election and it was indeed about the election, it’s a True Positive. When my system says that the tweet is about the election and it is not, then I have got a False Positive.

Taking this data, can you compute the Precision and Recall for the system at each threshold and identify the threshold values at which it does best, according to the F1 measure?

Precision is the fraction of output items that are correct versus incorrect, using GT. Its value equals TP/(FP+TP) which can also be seen as the fraction of relevant instances among the retrieved instances, Precision is a good measure to determine, when the costs of False Positive is high.

Recall is Fraction output items correct in GT. Its value equals TP/(TP+FN) which could also be seen as the number of correct results divided by the number of results that should have been returned. It shall be the model metric we use to select our best model when there is a high cost associated with False Negative.

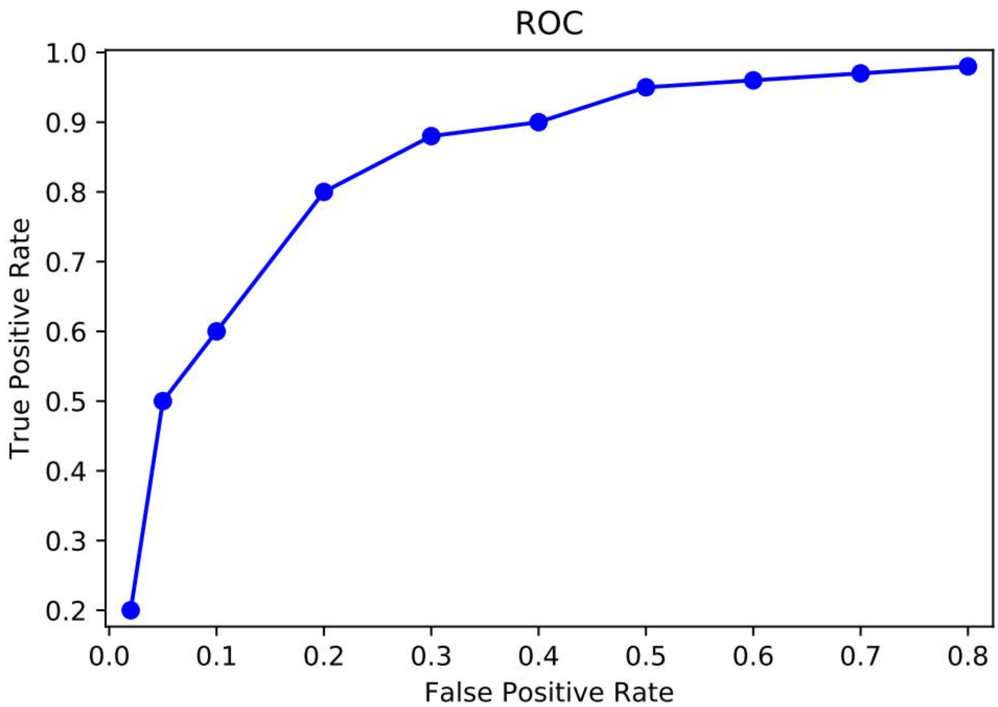
F1 Score is the harmonic mean of precision and recall. It gives equal weighting to both. Its value equals 2*Recall*Precision/(Recall+ Precision). It might be a better measure to use if we need to seek a balance between Precision and Recall and there is an uneven class distribution.

As shown in the following figure, I compute the Precision and Recall for the system at each threshold. according to the F1 measure, threshold 20 does best.

Threshold	TP	FN	FP	TN	Precision	Recall(TPR)	FPR	F1 Score
1	20	80	2	98	0.909090909	0.2	0.02	0.32786885
5	50	50	5	95	0.909090909	0.5	0.05	0.64516129
10	60	40	10	90	0.857142857	0.6	0.1	0.70588235
15	80	20	20	80	0.8	0.8	0.2	0.8
20	88	12	30	70	0.745762712	0.88	0.3	0.80733945
25	90	10	40	60	0.692307692	0.9	0.4	0.7826087
30	95	5	50	50	0.655172414	0.95	0.5	0.7755102
35	96	4	60	40	0.615384615	0.96	0.6	0.75
40	97	3	70	30	0.580838323	0.97	0.7	0.72659176
50	98	2	80	20	0.550561798	0.98	0.8	0.70503597

2) Now, can you plot the ROC for this data?

ROC curve is the abbreviation of receiver operating characteristic curve. It is used to to see how any predictive model can distinguish between the true positives and negatives. To plot ROC curve for this data, we need the value of TPR and FPR for each threshold. The value of TPR is same as Recall, it reflects the ratio of correct found in all correct. The value of FPR reflects ratio of incorrect found in all incorrect. Their values have been shown in the table above.



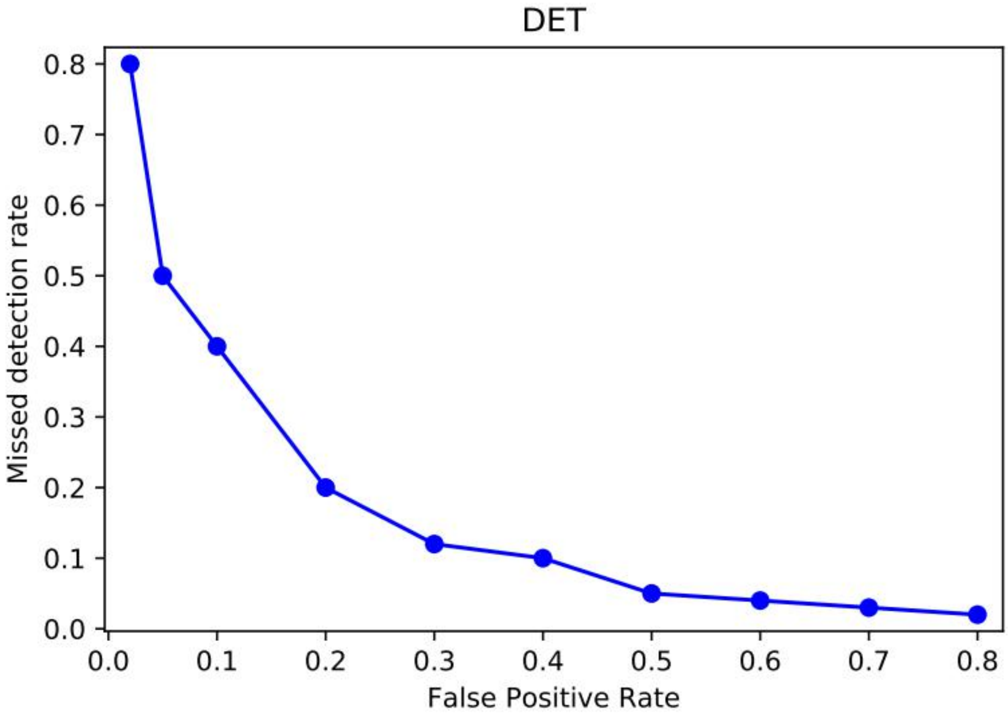
3) Now, can you plot the DET curve for the same data?

DET curves focus on errors more and "zoom in" to key parts of ROC curve by using log axes. To plot DET curve for this data, we need the value of FPR and missed detection rate for each threshold.

miss rate = false negative rate = FN/(TP + FN) = 1 - TPR

The value of FPR and miss rate of each threshold has been shown in the following table.

Threshold	TP	FN	FP	TN	Precision	Recall(TPR)	FPR	F1 Score	Miss Rate
1	20	80	2	98	0.909090909	0.2	0.02	0.32786885	0.8
5	50	50	5	95	0.909090909	0.5	0.05	0.64516129	0.5
10	60	40	10	90	0.857142857	0.6	0.1	0.70588235	0.4
15	80	20	20	80	0.8	0.8	0.2	0.8	0.2
20	88	12	30	70	0.745762712	0.88	0.3	0.80733945	0.12
25	90	10	40	60	0.692307692	0.9	0.4	0.7826087	0.1
30	95	5	50	50	0.655172414	0.95	0.5	0.7755102	0.05
35	96	4	60	40	0.615384615	0.96	0.6	0.75	0.04
40	97	3	70	30	0.580838323	0.97	0.7	0.72659176	0.03
50	98	2	80	20	0.550561798	0.98	0.8	0.70503597	0.02


DET
Missed detection rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8