Precession Recall & F-Measure Calculation:

	Predicted	
Actual	Positive	Negative
Positive	47	4
Negative	0	649

We have built the Confusion Matrix by taking Random sample of 700 by splitting 100 each among each group member to calculate the true actual Positive and Negative values. Total Positives found was 51 and remaining was 649 was negatives.

Precision = True Positive / (True Positive + False Positive)

= 47/47 = 100%

Recall = True Positive / (True Positive + False Positive)

= 47/51 = 92.1%

F-Measure = 2 * [(Precesion*Recall)/(precesion+Recall)]

= 95.9%

Title Extraction Approach:

1. We have copied the 60 GB dataset into the Hadoop file system
2. We have examined XML format of the data by printing it to the console using the Head command
3. Used Grep command and cleaned the data
4. Used Grep command to extract the titles from the Data and written the output of the Grep into a new file ()

Names Extraction Approach:

Approach 1:

We tried building a parser (XML SAX Parser) in java to integrate with Hadoop, but it was taking longer time to build the parser matching the XML data structure

Approach 2:

1. The extracted text file is taken into the python (about 1.2 M titles were extracted)
2. Read the data using python with open option, which is an efficient process to read a larger dataset without loading the entire data into the python memory.
3. Used the Probable people package to check if the title is name. Even through this approach we had lot of noise. So, we have taken the package identified names and put one more filter by

directly searching on the Wikipedia page and checked few parameters to conform if it is a name. But this approach takes little longer time as HTTP request parameter is included into the equation. But we can find 1450 names out of 1 Lakh titles.

Enhancements Possible:

We can use Multi programming approach to the above solution to reduce the running time.

Name Domino Approach:

We used python package Matplot lib and we used some of the references available in git hub and plotted rectangle boxes.