|  |  |
| --- | --- |
| NYU Tandon - CS6923 Machine Learning  Name: Jubin Soni (jas1464) | Fall 2016Professor Sandoval |

Homework 3

**Part-1: Coding Part (Using Python for NLP)**

**Section A: Term Frequency with unigrams**

Answer 1:

Most common word: 'the'.

Answer 2:

Three most common words now: 'data', 'with', 'python'.

Answer 3:

Words that don’t add value: 'data', 'with', 'python'.

If we look at 3 most common words we find that there are words like 'with', 'data', 'python' that do not add much value.

However, if we look at 50 most common words we should notice more trivial words: 'with', 'll', '3' that occur more frequently. Hence should also be removed.

Answer 4:

Most common word now is 'time'.

**Section B: Using TF-IDF**

Answer 5:

**Part 2 – Pen and Pencil**

**Problem 2:**

From the given table, True-False Error Condition can be calculated as below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Book** | **r (Actual)** | **y (Predicted)** | **Errors Condition** |
| War and Peace | + | - | FN |
| Pattern Recognition | + | + | TP |
| How to Win Friends and Influence People | - | - | TN |
| The Philosophical Breakfast Club | + | - | FN |
| Harry Potter | - | + | FP |
| Godel Escher Bach | + | + | TP |
| Photoshop for Dummies | - | - | TN |

Now using above calculation, we can total the error conditions for hypothesis to get:

|  |  |
| --- | --- |
| **Error Condition** | **Total** |
| FN | 2 |
| FP | 1 |
| TP | 2 |
| TN | 2 |

And Recall and Precision can be calculated as follows:

Therefore,

**Answer: ; and .**

**Problem 3:**

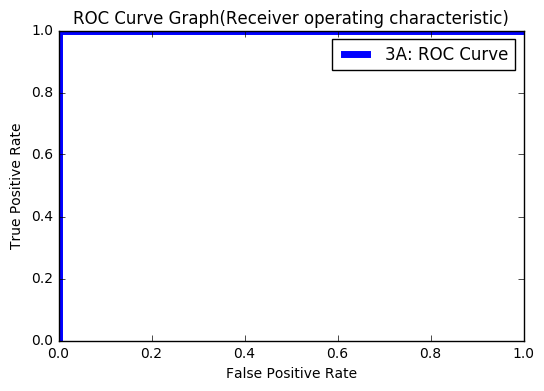
1. Suppose the labels for the samples, in descending order of values assigned by are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

With different , the classifier will predict:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | FN | TN | TP | FP |  |  |
|  |  |  |  |  |  |  | 3 | 3 | 0 | 0 |  |  |
|  |  |  |  |  |  |  | 2 | 3 | 1 | 0 |  |  |
|  |  |  |  |  |  |  | 1 | 3 | 2 | 0 |  |  |
|  |  |  |  |  |  |  | 0 | 3 | 3 | 0 |  |  |
|  |  |  |  |  |  |  | 0 | 2 | 3 | 1 |  |  |
|  |  |  |  |  |  |  | 0 | 1 | 3 | 2 |  |  |
|  |  |  |  |  |  |  | 0 | 0 | 3 | 3 |  |  |

So, we can plot ROC curve as:



Area under curve can be calculated as:

Therefore,

**AUC (Area under curve) is: 1.**

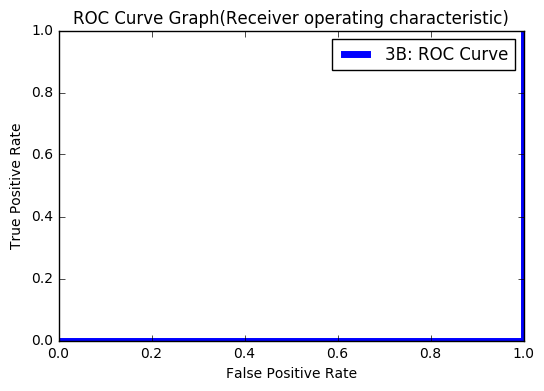
1. Suppose the labels for the samples, in descending order of values assigned by are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

With different , the classifier will predict:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | FN | TN | TP | FP |  |  |
|  |  |  |  |  |  |  | 3 | 3 | 0 | 0 |  |  |
|  |  |  |  |  |  |  | 3 | 2 | 0 | 1 |  |  |
|  |  |  |  |  |  |  | 3 | 1 | 0 | 2 |  |  |
|  |  |  |  |  |  |  | 3 | 0 | 0 | 3 |  |  |
|  |  |  |  |  |  |  | 2 | 0 | 1 | 3 |  |  |
|  |  |  |  |  |  |  | 1 | 0 | 2 | 3 |  |  |
|  |  |  |  |  |  |  | 0 | 0 | 3 | 3 |  |  |

So, we can plot ROC curve as:



**The AUC (Area under curve) is: 0.**

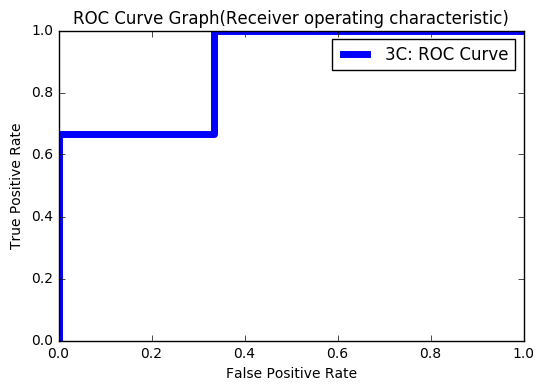
1. Suppose the labels for the samples, in descending order of values assigned by are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

With different , the classifier will predict:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | FN | TN | TP | FP |  |  |
|  |  |  |  |  |  |  | 3 | 3 | 0 | 0 |  |  |
|  |  |  |  |  |  |  | 2 | 3 | 1 | 0 |  |  |
|  |  |  |  |  |  |  | 1 | 3 | 2 | 0 |  |  |
|  |  |  |  |  |  |  | 1 | 2 | 2 | 1 |  |  |
|  |  |  |  |  |  |  | 0 | 2 | 3 | 1 |  |  |
|  |  |  |  |  |  |  | 0 | 1 | 3 | 2 |  |  |
|  |  |  |  |  |  |  | 0 | 0 | 3 | 3 |  |  |

So, we can plot ROC curve as:



Area under curve can be calculated as:

**The AUC (Area under curve) is: 8/9.**