Fall Semester 2016
Iowa State University

ENGL 520 - Computational Analysis of English

Problem Set 5
Text Classification
(ungraded)

1. Understand what is a "Confusion matrix" in text classification. Read the Wikipedia page as a starting point. (`https://goo.gl/blfYtG`. Understand the notions of: precision, recall, accuracy, F1 score.

2. Look at the confusion matrix shown in the Table below. Calculate the Precision, Recall, and F1 score for each category (A, B, C) and the overall classification accuracy of the classifier.

actual \ predicted	A	B	C
A	496	86	13
B	145	1625	137
C	22	114	340

3. Write a program that can, given a confusion matrix, calculates Precision, Recall, F1 score for each category automatically. There are a couple of issues to consider here. Some of them are:

   - How to receive input confusion matrix- perhaps as a csv file?
   - How to customize your program output based on the number of classes (2 or 4 or 24 instead of 3 classes as in the above example)

4. Two confusion matrices from two classifiers, both trained using the same dataset are shown below. Which one them in your view is a better classifier? Why?

(a) act. \ pred. →	**A**	**B**	**C**
A	496	86	13
B	145	1625	137
C	22	114	340

(b) act. \ pred. →	**A**	**B**	**C**
A	431	158	6
B	66	1767	74
C	11	161	304

Table 1: Confusion matrices comparison

5. Q1 in NLTK Chapter 6. (`http://www.nltk.org/book/ch06.html`)

6. Questions 6–10: Do the 5 questions given at the end of Chapter 5 in Language and Computers textbook.