14CS333 / UE14CS333 /– Natural Language Processing (6th Sem Elective) **7th Feb 2017**

Assignment 1 : Entropy estimation and Author Identification

Guidelines:

a) Code can be developed in any of C/C++/Java/Python (Open source compiler IDEs)

b) ***Assignment will have to be carried out by teams of size equal to TWO***
(Co Faculty Prof Arpitha Madam's decision will be final, for exceptions in team size)

c) Submission will have to be done, with a demo, on or before deadline.
Summary report (couple of pages max) , alongwith the demo will have to be handed over in ***hard copy to the Co Faculty***

d) Approx 4-weeks of time will be available before submission. Actual dates will be broadcast. Hence look out !

e) Follow fair code of ethics and , **develop your team's version** of code.

f) Your team will be called upon to demo the assignment, to match with submission data you have provided in the Hard Copy.

Problem Definition, Data Generation, Testing and Logging Stats

Problem: **Entropy Estimation and Author Identification**

Data Source : **https://www.gutenberg.org subset of files available to nltk.corpus**

Steps:

1) Pick any three authors with txt file size in the range of 150K-200K words
(for e.g)

```
>>> emma = nltk.corpus.gutenberg.words('austen-emma.txt')
>>> len(emma)
192427
```

2) Refer to the three txt files of Step(1) As C1, C2, C3

3) Build unigram , bigram & trigram models for C1, C2 and C3

4) Compute the Cross Entropy values for C1, C2 and C3, w.r.t each of the 3 models

5) Given a set of 5 - 25 continuous sentences, randomly picked , from C1 or C2 or C3 as anonymous text , assign the probability of the true authorship to the text from the three models for C1, C2 and C3.

6) Use the models for C1, C2 and C3 to produce 3 sets of generated text, of 100-125 words each. Comment on the similarity of the generated text with the original corpora namely C1, C2 and C3.

Your demo should display Model Entropy values, , Author Identification Stats, Probabilities for any given unigram, bigram or trigram, Number of Types and Tokens (for Uni, Bi and Tri grams) (include punctuations etc) in the command line interface. A sample of the same can be given as hard copy submission.

– Your Observations/Learning outcomes about this Assignment-1 (can be added to the hard copy submission)