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

Assignment 2 : Continuous text fragment similarity Estimation

Guidelines:

a) Code can be developed in any of 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.

g) This assignment is an option for those, who do not wish to turn in Assignment-1. However, every team is encouraged to implement the assignment .

Problem Definition, Data Generation, Testing and Logging Stats

Problem: **Paragraph Similarity Estimation**

Data/Theory Source :

i) **NLTK Wordnet 3.0**

ii) IEEE transaction paper:

"Sentence Similarity based on Semantic Net and Corpus Statistics ",
Yuhua Li et al., IEEE transactions on Knowledge and Data Engg, Vol 18, No 8, Aug 2006.

Steps:

1. Given a texts T1 and T2 compute sentence similarity s<i, j>, where s[i] is ith sentence from T1 and s[j] is the jth sentence from T2.

2. Implement your s<i , j> algorithm to include Semantic similarity and Order Similarity.

3. Compute the Matrix norm comprising of s<i, j> values, as the measure of similarity between T1 and T2.

4. Your demo can take in , at the command line/txt file, two segments of text and report the similarity measure.

5. Attach a single page hard copy report on the implementation and your observations on the learning outcome.