XCSE, Xavier University, Bhubaneswar.

Subject: Natural Language Processing
Faculty: Sourav Mandal
Assignment: 1

Course Percentage - 10 marks

Instructions: Prepare neat and clean assignment documents. Use Tables, Graphs, Pictures, etc. for better analysis and descriptions. Mail me complete pdf files. Appreciate handwritten documents. Program files can be shared directly or as document.

1. Writing Regular Expressions [20 points]

Download the some example books from Project Gutenberg that are included Link.

Such as https://www.gutenberg.org/files/1342/1342-h/1342-h.htm

You copy some plain texts and save in a `.txt` files.

This exercise is open ended. Use your favourite programming language, or the Unix `grep` utility, or the following two utility programs, `regexs.py` and `regexcount.py`, to explore this corpus. If you use the python code, use required libraries to write programs such as to search the text file for words over 20 characters long. Note the use of quotes to ensure that the spaces are interpreted as part of the regular expression.

Find some interesting patterns, and write up the regexes that describe those patterns, some example results, and an explanation for why they're interesting. Some examples include: common morphological suffixes, patterns of verbs that introduce dialogue in novels, patterns that indicate proper names, patterns that indicate verbs. Tokenize the text file and mention the counts of top 10 words excluding stop words.

Help: https://medium.com/factory-mind/regex-tutorial-a-simple-cheatsheet-by-examples-649dc1c3f285

http://www.nltk.org/book_1ed/ch03.html

2. Pronouncing Numbers [20 points]

Create a transducer that maps numbers in the range 0 – 999999 represented as digit strings to their English read form, e.g.:

- 1 -> one
- 1 1 -> eleven
- 1 1 1 -> one hundred eleven
- 1 1 1 1 -> one thousand one hundred eleven
- 1 1 1 1 1 -> eleven thousand one hundred eleven

Note: The input symbols can be the same ASCII character set as above, single digits are in ASCII, but the output symbols should be whole English words like "hundred" and "eleven".

Hint: Nondeterminism and epsilon outputs are helpful here, though this is still a one-to-one function.

3. Minimum Edit Distance [20 points]

Find out Minimum Edit Distance between two strings given below and identify any random possible alignment(s) using MED algorithm with Backtrace. Draw Tables wherever necessary.

- explosion → emotion
- elephant → telepathy