# Natural Language Processing - CS6370
# Spell Check Assignment Report

Aravind Sankar CS11B033
Sriram V CS11B058

September 26, 2014

## 1 Introduction

This assignment involved designing an efficient spell checker. This spell checking assignment was divided into three parts :

- Word checker where spelling corrections are given for a misspelled word not present in the dictionary.
- Phrase checker where spelling corrections are given for misspelled word(s) present in phrases.
- Sentence checker where spelling corrections are given misspelled word(s) present in sentences.

  The Spell Checker was implemented in *Python*.

## 2 Corpora used :

- The Unix dictionary for american english, which has close to 73,000 valid english words was used as the dictionary to identify if a given word has a spelling error.
- Corpus of Contemporary American English (COCA) dataset, which has 1,000,000 most frequent n-grams. This was used in the phrase and sentence spell checkers.
- The brown corpus was also initially used for learning context words, but isn't part of the final spell checker model.

## 3 Word Checker

The basic idea of approach for Word Checker was obtained from the paper by Kernighan et.al which was titled - A Spelling Correction Program Based on a Noisy Channel Model. This paper doesn't take the context in which a misspelt word appears into account, and provides spelling corrections for standalone words.