NAMED ENTITY TAGGING

Introduction: This assignment involves developing a Named Entity Tagger using Stanford Named Entity Tagger. Named-entity recognition (NER) is seeks to locate and classify named entity in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, etc.

Logic:

1) Initially, the Stanford Named Entity Recogniser version 3.9.1 zip file is downloaded from Stanford NLP website.
2) The lines from test file are tokenized using ‘word_tokenize’ function from nltk.
3) The tokens are then passed to StanfordNERTagger and the ouput list of tuple (token, tag) is converted to output_string where if tag is ‘O’, only token is appended to output_string else, token + ‘_’ + tag is appended to output_string.
4) This output_string is written to the ‘output.txt’ file.
5) The above procedure is repeated for all lines in the test file.

Computation of Precision, Recall and F-measure:

1) The program uses scikit learn module to compute the precision, recall and f-score.
2) For this, the program needs ‘expected_tags’ for the given test set. The program takes the already pre-trained text file and computes the expected tag sequence for the lines in the file.
3) The function score(expected_tags, generated_tags) returns precision, recall and f-score for the tag-sequences.

Dataset Used:

Constructed dataset by labelling the corpus to build pre-trained set manually.