Assignment 2

Task Description : The objective of this assignment is to experiment with POS tagging which is a standard sequence labelling task using Conditional Random Field (CRF).

Given a sequence of word tokens, you are required to predict the Part-of-Speech tag of each token.

Eg.

Word	**Tag**
heat	verb
water	noun
in	prep
a	det
large	adj
vessel	noun

CRF is a discriminative model, hence in this assignment, you are required to identify the correct POS tags for each word by introducing your own features. Please follow this link for a detailed description of the working of CRFs.

Steps to be followed to execute the task :

1. Train a CRF model using the available **sklearn crfsuite** wrapper.

2. Corpus to be used: **Universal Dependency Hindi corpus**
   (as provided in Moodle)

   Corpus Format : The corpus is in sentence tokenized form.
   The second column contains the word form and the third column contains its corresponding POS tag.

3. Introduce your own features to achieve the maximum possible F-score on the training set. For feature implementation, please follow the tutorial link provided in Step 1.

   Describe the features you have used in the notebook itself.

   For example: You can use features like : prefix, suffix, case, previous and next words for each of the tokens.

(Hint: Create a new markdown cell and report your justifications there.)

You are also allowed to tune the hyperparameters of the model in addition if you wish to.

4. Report the following on the train & test set :
    a. 10 most common transition features
    b. 10 least common transition features

    Transition features convey, how likely one tag will transition to the next tag.
    For example : ('ADJ', 'NOUN'), 4.114996).
    The numerical value is the transition weight. Higher the weight, more likely is the transition to take place.

5. Report the Precision, Recall, F-score per unique POS tag on the train and the test set. Report the overall accuracy of the task on the train and test set. The test set will be released one day prior to submission deadline.

Deliverables:

• You are required to submit the following in a zip folder **(Format : Assignment_2_YourRollNo.zip**)

- your .ipynb notebook containing the python code (**Assignment_2_YourRollNo.ipynb)**

- Complete the report and submit it in .pdf format. Report format attached in Moodle. (**Assignment_2_YourRollNo.pdf)**