Assignment 1
POS Tagging

Problem Statement:

Implement a POS tagger in Python using the Hidden Markov Model

Input and output:

- Dataset: **Brown corpus (tagset = "universal")**
- Output: Accuracy (5-fold cross-validation), confusion matrix, per POS accuracy
- Create a document that reports the following:
    1. Draw a confusion matrix report that includes all POS tags
    2. Report per POS accuracy (accuracy for each tag)
    3. Observe the strength and weaknesses of the model with respect to particular POS tags
    4. Perform detailed error analysis with examples
    5. Write a short paragraph on your learning.
- You should also create a simple demo that can take a sentence as input and generate the tags for each word as output.

    NOTE

    1. Use 5-fold cross-validation for reporting all accuracy values
    2. HMM need to be implemented from scratch

Dataset:

- Brown corpus (Available in NLTK library) (http://www.nltk.org/nltk_data/)

Submission Instructions:

- The assignment is to be submitted in groups of 3 (Same group for every assignment and project)
- The submission link will be created on Moodle to submit the assignment
- Only one person from the group with the lowest id is supposed to make the submission
- The name of the folder should be <id1_id2_id3>_Assignment1.zip
    - The uncompressed folder should have the name “<id1_id2_id3>_Assignment1” and shoud contains “code” folder, readme file and a report in pdf format (<id1_id2_id3_Assignment1>.pdf)
        - Example structure:
        - <id1_id2_id3>_Assignment1
            - code/
            - readme.txt
            - <id1_id2_id3_Assignment1>.pdf

    - The readme should contain details about the tools, versions, pre-requisites if any, and how to run the code for both approaches.
    - The report should contain all things mentioned in the problem statement.
        - Accuracies, Per POS accuracies, confusion matrix, error analysis, strengths, and weaknesses of model with respect to particular POS tags, and a short paragraph on your learning.

Deadline

- No-Hard deadline (Continuous Evaluation). The first Evaluation date will be announced soon.

References

- https://www.nltk.org/book/ch05.html
- https://pythonprogramming.net/svm-in-python-machine-learning-tutorial/ (Follow the series, learn, don't copy the code)

We shall check for code copying. Please be aware of neither copying codes from Git or across different teams.