Assignment 2: Parts-of-Speech Tagging (POS)

Welcome to the second assignment of Course 2 in the Natural Language Processing specialization. This assignment will develop skills in part-of-speech (POS) tagging, the process of assigning a part-of-speech tag (Noun, Verb, Adjective...) to each word in an input text. Tagging is difficult because some words can represent more than one part of speech at different times. They are **Ambiguous**. Let's look at the following example:

- The whole team played **well**. [adverb]
- You are doing **well** for yourself. [adjective]
- **Well**, this assignment took me forever to complete. [interjection]
- The **well** is dry. [noun]
- Tears were beginning to **well** in her eyes. [verb]

Distinguishing the parts-of-speech of a word in a sentence will help you better understand the meaning of a sentence. This would be critically important in search queries. Identifying the proper noun, the organization, the stock symbol, or anything similar would greatly improve everything ranging from speech recognition to search. By completing this assignment, you will:

- Learn how parts-of-speech tagging works
- Compute the transition matrix A in a Hidden Markov Model
- Compute the emission matrix B in a Hidden Markov Model
- Compute the Viterbi algorithm
- Compute the accuracy of your own model

Important Note on Submission to the AutoGrader

Before submitting your assignment to the AutoGrader, please make sure you are not doing the following:

1. You have not added any *extra* `print` statement(s) in the assignment.
2. You have not added any *extra* code cell(s) in the assignment.
3. You have not changed any of the function parameters.
4. You are not using any global variables inside your graded exercises. Unless specifically instructed to do so, please refrain from it and use the local variables instead.
5. You are not changing the assignment code where it is not required, like creating *extra* variables.

If you do any of the following, you will get something like, `Grader Error: Grader feedback not found` (or similarly unexpected) error upon submitting your assignment. Before asking for help/debugging the errors in your assignment, check for these first. If this is the case, and you don't remember the changes you have made, you can get a fresh copy of the assignment by following these instructions (https://www.coursera.org/learn/probabilistic-models-in-nlp/supplement/saGQf/how-to-refresh-your-workspace).

Table of Contents

- 0 - Data Sources
- 1 - Parts-of-speech Tagging
  - 1.1 - Training
    - Exercise 1 - create_dictionaries (UNQ_C1)
  - 1.2 - Testing
    - Exercise 2 - predict_pos (UNQ_C2)
- 2 - Hidden Markov Models for POS
  - 2.1 - Generating Matrices
    - Exercise 3 - create_transition_matrix (UNQ_C3)
    - Exercise 4 - create_emission_matrix (UNQ_C4)
- 3 - Viterbi Algorithm and Dynamic Programming
  - 3.1 - Initialization
    - Exercise 5 - initialize (UNQ_C5)
  - 3.2 - Viterbi Forward
    - Exercise 6 - viterbi_forward (UNQ_C6)
  - 3.3 - Viterbi Backward
    - Exercise 7 - viterbi_backward (UNQ_C7)
- 4 - Predicting on a Dataset
  - Exercise 8 - compute_accuracy (UNQ_C8)

In [1]:
```
# Importing packages and loading in the data set
from utils_pos import get_word_tag, preprocess
import pandas as pd
from collections import defaultdict
import math
import numpy as np
import w2_unittest
```

0 - Data Sources