coursera

Lab Help

Parts-of-Speech Tagging - First Steps: Working with text files, Creating a Vocabulary and Handling Unknown Words

In this lecture notebook you will create a vocabulary from a tagged dataset and learn how to deal with words that are not present in this vocabulary when working with other text sources. Aside from this you will also learn how to:

- read text files
- work with defaultdict
- work with string data

In [1]:

```
import string
from collections import defaultdict
```

Read Text Data

A tagged dataset taken from the Wall Street Journal is provided in the file `WSJ_02-21.pos`.

To read this file you can use Python's context manager by using the `with` keyword and specifying the name of the file you wish to read. To actually save the contents of the file into memory you will need to use the `readlines()` method and store its return value in a variable.

Python's context managers are great because you don't need to explicitly close the connection to the file, this is done under the hood:

In [2]:

```
# Read lines from 'WSJ_02-21.pos' file and save them into the 'lines' v
ariable
with open("WSJ_02-21.pos", 'r') as f:
    lines = f.readlines()
```

To check the contents of the dataset you can print the first 5 lines:

In [3]:

```
# Print columns for reference
print("\t\tWord", "\tTag\n")

# Print first five lines of the dataset
for i in range(5):
    print(f'line number {i+1}: {lines[i]}')
```

```
                Word     Tag
```