HMM-For-Seq-Tagging

March 25, 2024

[1]:
```
#Implementation of the Hidden Markov Model in Python
```

[2]:
```
#Exploring Treebank Tagged Corpus
```

[3]:
```
#Importing libraries
import nltk, re, pprint
import numpy as np
import pandas as pd
import requests
import matplotlib.pyplot as plt
import seaborn as sns
import pprint, time
import random
from sklearn.model_selection import train_test_split
from nltk.tokenize import word_tokenize
# reading the Treebank tagged sentences
wsj = list(nltk.corpus.treebank.tagged_sents())
# first few tagged sentences
print(wsj[:40])
```

```
[[('Pierre', 'NNP'), ('Vinken', 'NNP'), (',', ','), ('61', 'CD'), ('years',
'NNS'), ('old', 'JJ'), (',', ','), ('will', 'MD'), ('join', 'VB'), ('the',
'DT'), ('board', 'NN'), ('as', 'IN'), ('a', 'DT'), ('nonexecutive', 'JJ'),
('director', 'NN'), ('Nov.', 'NNP'), ('29', 'CD'), ('.', '.')], [('Mr.', 'NNP'),
('Vinken', 'NNP'), ('is', 'VBZ'), ('chairman', 'NN'), ('of', 'IN'), ('Elsevier',
'NNP'), ('N.V.', 'NNP'), (',', ','), ('the', 'DT'), ('Dutch', 'NNP'),
('publishing', 'VBG'), ('group', 'NN'), ('.', '.')], [('Rudolph', 'NNP'),
('Agnew', 'NNP'), (',', ','), ('55', 'CD'), ('years', 'NNS'), ('old', 'JJ'),
('and', 'CC'), ('former', 'JJ'), ('chairman', 'NN'), ('of', 'IN'),
('Consolidated', 'NNP'), ('Gold', 'NNP'), ('Fields', 'NNP'), ('PLC', 'NNP'),
(',', ','), ('was', 'VBD'), ('named', 'VBN'), ('*-1', '-NONE-'), ('a', 'DT'),
('nonexecutive', 'JJ'), ('director', 'NN'), ('of', 'IN'), ('this', 'DT'),
('British', 'JJ'), ('industrial', 'JJ'), ('conglomerate', 'NN'), ('.', '.')],
[('A', 'DT'), ('form', 'NN'), ('of', 'IN'), ('asbestos', 'NN'), ('once', 'RB'),
('used', 'VBN'), ('*', '-NONE-'), ('*', '-NONE-'), ('to', 'TO'), ('make', 'VB'),
('Kent', 'NNP'), ('cigarette', 'NN'), ('filters', 'NNS'), ('has', 'VBZ'),
('caused', 'VBN'), ('a', 'DT'), ('high', 'JJ'), ('percentage', 'NN'), ('of',
'IN'), ('cancer', 'NN'), ('deaths', 'NNS'), ('among', 'IN'), ('a', 'DT'),
('group', 'NN'), ('of', 'IN'), ('workers', 'NNS'), ('exposed', 'VBN'), ('*',
'-NONE-'), ('to', 'TO'), ('it', 'PRP'), ('more', 'RBR'), ('than', 'IN'), ('30',
'CD'), ('years', 'NNS'), ('ago', 'IN'), (',', ','), ('researchers', 'NNS'),
('reported', 'VBD'), ('0', '-NONE-'), ('*T*-1', '-NONE-'), ('.', '.')], [('The',
'DT'), ('asbestos', 'NN'), ('fiber', 'NN'), (',', ','), ('crocidolite', 'NN'),
(',', ','), ('is', 'VBZ'), ('unusually', 'RB'), ('resilient', 'JJ'), ('once',
'IN'), ('it', 'PRP'), ('enters', 'VBZ'), ('the', 'DT'), ('lungs', 'NNS'), (',',
','), ('with', 'IN'), ('even', 'RB'), ('brief', 'JJ'), ('exposures', 'NNS'),
('to', 'TO'), ('it', 'PRP'), ('causing', 'VBG'), ('symptoms', 'NNS'), ('that',
'WDT'), ('*T*-1', '-NONE-'), ('show', 'VBP'), ('up', 'RP'), ('decades', 'NNS'),
('later', 'JJ'), (',', ','), ('researchers', 'NNS'), ('said', 'VBD'), ('0',
'-NONE-'), ('*T*-2', '-NONE-'), ('.', '.')], [('Lorillard', 'NNP'), ('Inc.',
'NNP'), (',', ','), ('the', 'DT'), ('unit', 'NN'), ('of', 'IN'), ('New', 'JJ'),
('York-based', 'JJ'), ('Loews', 'NNP'), ('Corp.', 'NNP'), ('that', 'WDT'),
('*T*-2', '-NONE-'), ('makes', 'VBZ'), ('Kent', 'NNP'), ('cigarettes', 'NNS'),
(',', ','), ('stopped', 'VBD'), ('using', 'VBG'), ('crocidolite', 'NN'), ('in',
'IN'), ('its', 'PRP$'), ('Micronite', 'NN'), ('cigarette', 'NN'), ('filters',
'NNS'), ('in', 'IN'), ('1956', 'CD'), ('.', '.')], [('Although', 'IN'),
('preliminary', 'JJ'), ('findings', 'NNS'), ('were', 'VBD'), ('reported',
'VBN'), ('*-2', '-NONE-'), ('more', 'RBR'), ('than', 'IN'), ('a', 'DT'),
('year', 'NN'), ('ago', 'IN'), (',', ','), ('the', 'DT'), ('latest', 'JJS'),
('results', 'NNS'), ('appear', 'VBP'), ('in', 'IN'), ('today', 'NN'), ("'s",
'POS'), ('New', 'NNP'), ('England', 'NNP'), ('Journal', 'NNP'), ('of', 'IN'),
('Medicine', 'NNP'), (',', ','), ('a', 'DT'), ('forum', 'NN'), ('likely', 'JJ'),
('*', '-NONE-'), ('to', 'TO'), ('bring', 'VB'), ('new', 'JJ'), ('attention',
'NN'), ('to', 'TO'), ('the', 'DT'), ('problem', 'NN'), ('.', '.')], [('A',
'DT'), ('Lorillard', 'NNP'), ('spokewoman', 'NN'), ('said', 'VBD'), (',', ','),
('``', '``'), ('This', 'DT'), ('is', 'VBZ'), ('an', 'DT'), ('old', 'JJ'),
('story', 'NN'), ('.', '.')], [('We', 'PRP'), ("'re", 'VBP'), ('talking',
'VBG'), ('about', 'IN'), ('years', 'NNS'), ('ago', 'IN'), ('before', 'IN'),
('anyone', 'NN'), ('heard', 'VBD'), ('of', 'IN'), ('asbestos', 'NN'), ('having',
'VBG'), ('any', 'DT'), ('questionable', 'JJ'), ('properties', 'NNS'), ('.',
'.')], [('There', 'EX'), ('is', 'VBZ'), ('no', 'DT'), ('asbestos', 'NN'), ('in',
'IN'), ('our', 'PRP$'), ('products', 'NNS'), ('now', 'RB'), ('.', '.'), ("''",
"''")], [('Neither', 'DT'), ('Lorillard', 'NNP'), ('nor', 'CC'), ('the', 'DT'),
('researchers', 'NNS'), ('who', 'WP'), ('*T*-3', '-NONE-'), ('studied', 'VBD'),
('the', 'DT'), ('workers', 'NNS'), ('were', 'VBD'), ('aware', 'JJ'), ('of',
'IN'), ('any', 'DT'), ('research', 'NN'), ('on', 'IN'), ('smokers', 'NNS'),
('of', 'IN'), ('the', 'DT'), ('Kent', 'NNP'), ('cigarettes', 'NNS'), ('.',
'.')], [('``', '``'), ('We', 'PRP'), ('have', 'VBP'), ('no', 'DT'), ('useful',
'JJ'), ('information', 'NN'), ('on', 'IN'), ('whether', 'IN'), ('users', 'NNS'),
('are', 'VBP'), ('at', 'IN'), ('risk', 'NN'), (',', ','), ("''", "''"), ('said',
'VBD'), ('*T*-1', '-NONE-'), ('James', 'NNP'), ('A.', 'NNP'), ('Talcott',
'NNP'), ('of', 'IN'), ('Boston', 'NNP'), ("'s", 'POS'), ('Dana-Farber', 'NNP'),
('Cancer', 'NNP'), ('Institute', 'NNP'), ('.', '.')], [('Dr.', 'NNP'),
('Talcott', 'NNP'), ('led', 'VBD'), ('a', 'DT'), ('team', 'NN'), ('of', 'IN'),
('researchers', 'NNS'), ('from', 'IN'), ('the', 'DT'), ('National', 'NNP'),
('Cancer', 'NNP'), ('Institute', 'NNP'), ('and', 'CC'), ('the', 'DT'),
('medical', 'JJ'), ('schools', 'NNS'), ('of', 'IN'), ('Harvard', 'NNP'),
('University', 'NNP'), ('and', 'CC'), ('Boston', 'NNP'), ('University', 'NNP'),
('.', '.')], [('The', 'DT'), ('Lorillard', 'NNP'), ('spokeswoman', 'NN'),
('said', 'VBD'), ('0', '-NONE-'), ('asbestos', 'NN'), ('was', 'VBD'), ('used',
'VBN'), ('*-1', '-NONE-'), ('in', 'IN'), ('``', '``'), ('very', 'RB'),
('modest', 'JJ'), ('amounts', 'NNS'), ("''", "''"), ('in', 'IN'), ('*',
'-NONE-'), ('making', 'VBG'), ('paper', 'NN'), ('for', 'IN'), ('the', 'DT'),
('filters', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('early', 'JJ'), ('1950s',
'CD'), ('and', 'CC'), ('replaced', 'VBN'), ('*-1', '-NONE-'), ('with', 'IN'),
('a', 'DT'), ('different', 'JJ'), ('type', 'NN'), ('of', 'IN'), ('filter',
'NN'), ('in', 'IN'), ('1956', 'CD'), ('.', '.')], [('From', 'IN'), ('1953',
'CD'), ('to', 'TO'), ('1955', 'CD'), (',', ','), ('9.8', 'CD'), ('billion',
'CD'), ('Kent', 'NNP'), ('cigarettes', 'NNS'), ('with', 'IN'), ('the', 'DT'),
('filters', 'NNS'), ('were', 'VBD'), ('sold', 'VBN'), ('*-3', '-NONE-'), (',',
','), ('the', 'DT'), ('company', 'NN'), ('said', 'VBD'), ('0', '-NONE-'),
('*T*-1', '-NONE-'), ('.', '.')], [('Among', 'IN'), ('33', 'CD'), ('men',
'NNS'), ('who', 'WP'), ('*T*-4', '-NONE-'), ('worked', 'VBD'), ('closely',
'RB'), ('with', 'IN'), ('the', 'DT'), ('substance', 'NN'), (',', ','), ('28',
'CD'), ('*ICH*-1', '-NONE-'), ('have', 'VBP'), ('died', 'VBN'), ('--', ':'),
('more', 'JJ'), ('than', 'IN'), ('three', 'CD'), ('times', 'NNS'), ('the',
'DT'), ('expected', 'VBN'), ('number', 'NN'), ('.', '.')], [('Four', 'CD'),
('of', 'IN'), ('the', 'DT'), ('five', 'CD'), ('surviving', 'VBG'), ('workers',
'NNS'), ('have', 'VBP'), ('asbestos-related', 'JJ'), ('diseases', 'NNS'), (',',
','), ('including', 'VBG'), ('three', 'CD'), ('with', 'IN'), ('recently', 'RB'),
('diagnosed', 'VBN'), ('cancer', 'NN'), ('.', '.')], [('The', 'DT'), ('total',
'NN'), ('of', 'IN'), ('18', 'CD'), ('deaths', 'NNS'), ('from', 'IN'),
('malignant', 'JJ'), ('mesothelioma', 'NN'), (',', ','), ('lung', 'NN'),
('cancer', 'NN'), ('and', 'CC'), ('asbestosis', 'NN'), ('was', 'VBD'), ('far',
'RB'), ('higher', 'JJR'), ('than', 'IN'), ('*', '-NONE-'), ('expected', 'VBN'),
('*?*', '-NONE-'), (',', ','), ('the', 'DT'), ('researchers', 'NNS'), ('said',
'VBD'), ('0', '-NONE-'), ('*T*-1', '-NONE-'), ('.', '.')], [('``', '``'),
('The', 'DT'), ('morbidity', 'NN'), ('rate', 'NN'), ('is', 'VBZ'), ('a', 'DT'),
('striking', 'JJ'), ('finding', 'NN'), ('among', 'IN'), ('those', 'DT'), ('of',
'IN'), ('us', 'PRP'), ('who', 'WP'), ('*T*-5', '-NONE-'), ('study', 'VBP'),
('asbestos-related', 'JJ'), ('diseases', 'NNS'), (',', ','), ("''", "''"),
('said', 'VBD'), ('*T*-1', '-NONE-'), ('Dr.', 'NNP'), ('Talcott', 'NNP'), ('.',
'.')], [('The', 'DT'), ('percentage', 'NN'), ('of', 'IN'), ('lung', 'NN'),
('cancer', 'NN'), ('deaths', 'NNS'), ('among', 'IN'), ('the', 'DT'), ('workers',
'NNS'), ('at', 'IN'), ('the', 'DT'), ('West', 'NNP'), ('Groton', 'NNP'), (',',
','), ('Mass.', 'NNP'), (',', ','), ('paper', 'NN'), ('factory', 'NN'),
('appears', 'VBZ'), ('*-1', '-NONE-'), ('to', 'TO'), ('be', 'VB'), ('the',
'DT'), ('highest', 'JJS'), ('for', 'IN'), ('any', 'DT'), ('asbestos', 'NN'),
('workers', 'NNS'), ('studied', 'VBN'), ('*', '-NONE-'), ('in', 'IN'),
('Western', 'JJ'), ('industrialized', 'VBN'), ('countries', 'NNS'), (',', ','),
('he', 'PRP'), ('said', 'VBD'), ('0', '-NONE-'), ('*T*-2', '-NONE-'), ('.',
'.')], [('The', 'DT'), ('plant', 'NN'), (',', ','), ('which', 'WDT'), ('*T*-1',
'-NONE-'), ('is', 'VBZ'), ('owned', 'VBN'), ('*-4', '-NONE-'), ('by', 'IN'),
('Hollingsworth', 'NNP'), ('&', 'CC'), ('Vose', 'NNP'), ('Co.', 'NNP'), (',',
','), ('was', 'VBD'), ('under', 'IN'), ('contract', 'NN'), ('*ICH*-2',
'-NONE-'), ('with', 'IN'), ('Lorillard', 'NN'), ('*', '-NONE-'), ('to', 'TO'),
('make', 'VB'), ('the', 'DT'), ('cigarette', 'NN'), ('filters', 'NNS'), ('.',
'.')], [('The', 'DT'), ('finding', 'NN'), ('probably', 'RB'), ('will', 'MD'),
('support', 'VB'), ('those', 'DT'), ('who', 'WP'), ('*T*-6', '-NONE-'),
('argue', 'VBP'), ('that', 'IN'), ('the', 'DT'), ('U.S.', 'NNP'), ('should',
'MD'), ('regulate', 'VB'), ('the', 'DT'), ('class', 'NN'), ('of', 'IN'),
('asbestos', 'NN'), ('including', 'VBG'), ('crocidolite', 'NN'), ('more',
'RBR'), ('stringently', 'RB'), ('than', 'IN'), ('the', 'DT'), ('common', 'JJ'),
('kind', 'NN'), ('of', 'IN'), ('asbestos', 'NN'), (',', ','), ('chrysotile',
'NN'), (',', ','), ('found', 'VBN'), ('*', '-NONE-'), ('in', 'IN'), ('most',
'JJS'), ('schools', 'NNS'), ('and', 'CC'), ('other', 'JJ'), ('buildings',
'NNS'), (',', ','), ('Dr.', 'NNP'), ('Talcott', 'NNP'), ('said', 'VBD'), ('0',
'-NONE-'), ('*T*-1', '-NONE-'), ('.', '.')], [('The', 'DT'), ('U.S.', 'NNP'),
('is', 'VBZ'), ('one', 'CD'), ('of', 'IN'), ('the', 'DT'), ('few', 'JJ'),
('industrialized', 'VBN'), ('nations', 'NNS'), ('that', 'WDT'), ('*T*-7',
'-NONE-'), ('does', 'VBZ'), ("n't", 'RB'), ('have', 'VB'), ('a', 'DT'),
('higher', 'JJR'), ('standard', 'NN'), ('of', 'IN'), ('regulation', 'NN'),
('for', 'IN'), ('the', 'DT'), ('smooth', 'JJ'), (',', ','), ('needle-like',
'JJ'), ('fibers', 'NNS'), ('such', 'JJ'), ('as', 'IN'), ('crocidolite', 'NN'),
('that', 'WDT'), ('*T*-1', '-NONE-'), ('are', 'VBP'), ('classified', 'VBN'),
('*-5', '-NONE-'), ('as', 'IN'), ('amphobiles', 'NNS'), (',', ','),
('according', 'VBG'), ('to', 'TO'), ('Brooke', 'NNP'), ('T.', 'NNP'),
('Mossman', 'NNP'), (',', ','), ('a', 'DT'), ('professor', 'NN'), ('of', 'IN'),
('pathlogy', 'NN'), ('at', 'IN'), ('the', 'DT'), ('University', 'NNP'), ('of',
'IN'), ('Vermont', 'NNP'), ('College', 'NNP'), ('of', 'IN'), ('Medicine',
'NNP'), ('.', '.')], [('More', 'RBR'), ('common', 'JJ'), ('chrysotile', 'NN'),
('fibers', 'NNS'), ('are', 'VBP'), ('curly', 'JJ'), ('and', 'CC'), ('are',
'VBP'), ('more', 'RBR'), ('easily', 'RB'), ('rejected', 'VBN'), ('*-1',
'-NONE-'), ('by', 'IN'), ('the', 'DT'), ('body', 'NN'), (',', ','), ('Dr.',
'NNP'), ('Mossman', 'NNP'), ('explained', 'VBD'), ('0', '-NONE-'), ('*T*-2',
'-NONE-'), ('.', '.')], [('In', 'IN'), ('July', 'NNP'), (',', ','), ('the',
'DT'), ('Environmental', 'NNP'), ('Protection', 'NNP'), ('Agency', 'NNP'),
('imposed', 'VBD'), ('a', 'DT'), ('gradual', 'JJ'), ('ban', 'NN'), ('on', 'IN'),
('virtually', 'RB'), ('all', 'DT'), ('uses', 'NNS'), ('of', 'IN'), ('asbestos',
'NN'), ('.', '.')], [('By', 'IN'), ('1997', 'CD'), (',', ','), ('almost', 'RB'),
('all', 'DT'), ('remaining', 'VBG'), ('uses', 'NNS'), ('of', 'IN'), ('cancer-
causing', 'JJ'), ('asbestos', 'NN'), ('will', 'MD'), ('be', 'VB'), ('outlawed',
'VBN'), ('*-6', '-NONE-'), ('.', '.')], [('About', 'IN'), ('160', 'CD'),
('workers', 'NNS'), ('at', 'IN'), ('a', 'DT'), ('factory', 'NN'), ('that',
'WDT'), ('*T*-8', '-NONE-'), ('made', 'VBD'), ('paper', 'NN'), ('for', 'IN'),
('the', 'DT'), ('Kent', 'NNP'), ('filters', 'NNS'), ('were', 'VBD'), ('exposed',
'VBN'), ('*-7', '-NONE-'), ('to', 'TO'), ('asbestos', 'NN'), ('in', 'IN'),
('the', 'DT'), ('1950s', 'CD'), ('.', '.')], [('Areas', 'NNS'), ('of', 'IN'),
('the', 'DT'), ('factory', 'NN'), ('*ICH*-2', '-NONE-'), ('were', 'VBD'),
('particularly', 'RB'), ('dusty', 'JJ'), ('where', 'WRB'), ('the', 'DT'),
('crocidolite', 'NN'), ('was', 'VBD'), ('used', 'VBN'), ('*-8', '-NONE-'),
('*T*-1', '-NONE-'), ('.', '.')], [('Workers', 'NNS'), ('dumped', 'VBD'),
('large', 'JJ'), ('burlap', 'NN'), ('sacks', 'NNS'), ('of', 'IN'), ('the',
'DT'), ('imported', 'VBN'), ('material', 'NN'), ('into', 'IN'), ('a', 'DT'),
('huge', 'JJ'), ('bin', 'NN'), (',', ','), ('poured', 'VBD'), ('in', 'RP'),
('cotton', 'NN'), ('and', 'CC'), ('acetate', 'NN'), ('fibers', 'NNS'), ('and',
'CC'), ('mechanically', 'RB'), ('mixed', 'VBD'), ('the', 'DT'), ('dry', 'JJ'),
('fibers', 'NNS'), ('in', 'IN'), ('a', 'DT'), ('process', 'NN'), ('used',
'VBN'), ('*', '-NONE-'), ('*', '-NONE-'), ('to', 'TO'), ('make', 'VB'),
('filters', 'NNS'), ('.', '.')], [('Workers', 'NNS'), ('described', 'VBD'),
('``', '``'), ('clouds', 'NNS'), ('of', 'IN'), ('blue', 'JJ'), ('dust', 'NN'),
("''", "''"), ('that', 'WDT'), ('*T*-1', '-NONE-'), ('hung', 'VBD'), ('over',
'IN'), ('parts', 'NNS'), ('of', 'IN'), ('the', 'DT'), ('factory', 'NN'), (',',
','), ('even', 'RB'), ('though', 'IN'), ('exhaust', 'NN'), ('fans', 'NNS'),
('ventilated', 'VBD'), ('the', 'DT'), ('area', 'NN'), ('.', '.')], [('``',
'``'), ('There', 'EX'), ("'s", 'VBZ'), ('no', 'DT'), ('question', 'NN'),
('that', 'IN'), ('some', 'DT'), ('of', 'IN'), ('those', 'DT'), ('workers',
'NNS'), ('and', 'CC'), ('managers', 'NNS'), ('contracted', 'VBD'), ('asbestos-
related', 'JJ'), ('diseases', 'NNS'), (',', ','), ("''", "''"), ('said', 'VBD'),
('*T*-1', '-NONE-'), ('Darrell', 'NNP'), ('Phillips', 'NNP'), (',', ','),
('vice', 'NN'), ('president', 'NN'), ('of', 'IN'), ('human', 'JJ'),
('resources', 'NNS'), ('for', 'IN'), ('Hollingsworth', 'NNP'), ('&', 'CC'),
('Vose', 'NNP'), ('.', '.')], [('``', '``'), ('But', 'CC'), ('you', 'PRP'),
('have', 'VBP'), ('*-1', '-NONE-'), ('to', 'TO'), ('recognize', 'VB'), ('that',
'IN'), ('these', 'DT'), ('events', 'NNS'), ('took', 'VBD'), ('place', 'NN'),
('35', 'CD'), ('years', 'NNS'), ('ago', 'IN'), ('.', '.')], [('It', 'PRP'),
('has', 'VBZ'), ('no', 'DT'), ('bearing', 'NN'), ('on', 'IN'), ('our', 'PRP$'),
('work', 'NN'), ('force', 'NN'), ('today', 'NN'), ('.', '.')], [('Yields',
'NNS'), ('on', 'IN'), ('money-market', 'JJ'), ('mutual', 'JJ'), ('funds',
'NNS'), ('continued', 'VBD'), ('*-1', '-NONE-'), ('to', 'TO'), ('slide', 'VB'),
(',', ','), ('amid', 'IN'), ('signs', 'NNS'), ('that', 'IN'), ('portfolio',
'NN'), ('managers', 'NNS'), ('expect', 'VBP'), ('further', 'JJ'), ('declines',
'NNS'), ('in', 'IN'), ('interest', 'NN'), ('rates', 'NNS'), ('.', '.')],
[('The', 'DT'), ('average', 'JJ'), ('seven-day', 'JJ'), ('compound', 'NN'),
('yield', 'NN'), ('of', 'IN'), ('the', 'DT'), ('400', 'CD'), ('taxable', 'JJ'),
('funds', 'NNS'), ('tracked', 'VBN'), ('*', '-NONE-'), ('by', 'IN'), ('IBC',
'NNP'), ("'s", 'POS'), ('Money', 'NNP'), ('Fund', 'NNP'), ('Report', 'NNP'),
('eased', 'VBD'), ('a', 'DT'), ('fraction', 'NN'), ('of', 'IN'), ('a', 'DT'),
('percentage', 'NN'), ('point', 'NN'), ('to', 'TO'), ('8.45', 'CD'), ('%',
'NN'), ('from', 'IN'), ('8.47', 'CD'), ('%', 'NN'), ('for', 'IN'), ('the',
'DT'), ('week', 'NN'), ('ended', 'VBD'), ('Tuesday', 'NNP'), ('.', '.')],
[('Compound', 'NN'), ('yields', 'NNS'), ('assume', 'VBP'), ('reinvestment',
'NN'), ('of', 'IN'), ('dividends', 'NNS'), ('and', 'CC'), ('that', 'IN'),
('the', 'DT'), ('current', 'JJ'), ('yield', 'NN'), ('continues', 'VBZ'), ('for',
'IN'), ('a', 'DT'), ('year', 'NN'), ('.', '.')], [('Average', 'JJ'),
('maturity', 'NN'), ('of', 'IN'), ('the', 'DT'), ('funds', 'NNS'), ("'", 'POS'),
('investments', 'NNS'), ('lengthened', 'VBD'), ('by', 'IN'), ('a', 'DT'),
('day', 'NN'), ('to', 'TO'), ('41', 'CD'), ('days', 'NNS'), (',', ','), ('the',
'DT'), ('longest', 'JJS'), ('since', 'IN'), ('early', 'JJ'), ('August', 'NNP'),
(',', ','), ('according', 'VBG'), ('to', 'TO'), ('Donoghue', 'NNP'), ("'s",
'POS'), ('.', '.')], [('Longer', 'JJR'), ('maturities', 'NNS'), ('are', 'VBP'),
('thought', 'VBN'), ('*-1', '-NONE-'), ('to', 'TO'), ('indicate', 'VB'),
('declining', 'VBG'), ('interest', 'NN'), ('rates', 'NNS'), ('because', 'IN'),
('they', 'PRP'), ('permit', 'VBP'), ('portfolio', 'NN'), ('managers', 'NNS'),
('to', 'TO'), ('retain', 'VB'), ('relatively', 'RB'), ('higher', 'JJR'),
('rates', 'NNS'), ('for', 'IN'), ('a', 'DT'), ('longer', 'JJR'), ('period',
'NN'), ('.', '.')], [('Shorter', 'JJR'), ('maturities', 'NNS'), ('are', 'VBP'),
('considered', 'VBN'), ('*-9', '-NONE-'), ('a', 'DT'), ('sign', 'NN'), ('of',
'IN'), ('rising', 'VBG'), ('rates', 'NNS'), ('because', 'IN'), ('portfolio',
'NN'), ('managers', 'NNS'), ('can', 'MD'), ('capture', 'VB'), ('higher', 'JJR'),
('rates', 'NNS'), ('sooner', 'RB'), ('.', '.')], [('The', 'DT'), ('average',
'JJ'), ('maturity', 'NN'), ('for', 'IN'), ('funds', 'NNS'), ('open', 'JJ'),
('only', 'RB'), ('to', 'TO'), ('institutions', 'NNS'), (',', ','),
('considered', 'VBN'), ('by', 'IN'), ('some', 'DT'), ('*', '-NONE-'), ('to',
'TO'), ('be', 'VB'), ('a', 'DT'), ('stronger', 'JJR'), ('indicator', 'NN'),
('because', 'IN'), ('those', 'DT'), ('managers', 'NNS'), ('watch', 'VBP'),
('the', 'DT'), ('market', 'NN'), ('closely', 'RB'), (',', ','), ('reached',
'VBD'), ('a', 'DT'), ('high', 'JJ'), ('point', 'NN'), ('for', 'IN'), ('the',
'DT'), ('year', 'NN'), ('--', ':'), ('33', 'CD'), ('days', 'NNS'), ('.', '.')]]
```

[4]:
```
#Train Test Split
```

[5]:
```
#In this step, we will split the dataset into a 70:30 ratio
```

[6]:
```
# Splitting into train and test
random.seed(1234)
train_set, test_set = train_test_split(wsj,test_size=0.3)
print(len(train_set))
print(len(test_set))
print(train_set[:40])
```

```
2739
1175
[[('In', 'IN'), ('an', 'DT'), ('era', 'NN'), ('when', 'WRB'), ('every', 'DT'),
('government', 'NN'), ('agency', 'NN'), ('has', 'VBZ'), ('a', 'DT'), ('public-
relations', 'NNS'), ('machine', 'NN'), ('that', 'WDT'), ('*T*-2', '-NONE-'),
('sends', 'VBZ'), ('you', 'PRP'), ('stuff', 'NN'), ('whether', 'IN'), ('you',
'PRP'), ('want', 'VBP'), ('it', 'PRP'), ('or', 'CC'), ('not', 'RB'), ('*T*-1',
'-NONE-'), (',', ','), ('this', 'DT'), ('does', 'VBZ'), ('seem', 'VB'), ('odd',
'JJ'), ('.', '.')], [('--', ':'), ('Of', 'IN'), ('all', 'DT'), ('scenes',
'NNS'), ('that', 'WDT'), ('*T*-219', '-NONE-'), ('evoke', 'VBP'), ('rural',
'JJ'), ('England', 'NNP'), (',', ','), ('this', 'DT'), ('is', 'VBZ'), ('one',
'CD'), ('of', 'IN'), ('the', 'DT'), ('loveliest', 'JJS'), ('*T*-2', '-NONE-'),
(':', ':'), ('An', 'DT'), ('ancient', 'JJ'), ('stone', 'NN'), ('church', 'NN'),
('stands', 'VBZ'), ('amid', 'IN'), ('the', 'DT'), ('fields', 'NNS'), (',', ','),
('the', 'DT'), ('sound', 'NN'), ('of', 'IN'), ('bells', 'NNS'), ('cascading',
'VBG'), ('from', 'IN'), ('its', 'PRP$'), ('tower', 'NN'), (',', ','), ('*-1',
'-NONE-'), ('calling', 'VBG'), ('the', 'DT'), ('faithful', 'NN'), ('to', 'TO'),
('evensong', 'NN'), ('.', '.')], [('A', 'DT'), ('50-state', 'JJ'), ('study',
'NN'), ('released', 'VBN'), ('*', '-NONE-'), ('in', 'IN'), ('September', 'NNP'),
('by', 'IN'), ('Friends', 'NNPS'), ('for', 'IN'), ('Education', 'NNP'), (',',
','), ('an', 'DT'), ('Albuquerque', 'NNP'), (',', ','), ('N.M.', 'NNP'), (',',
','), ('school-research', 'JJ'), ('group', 'NN'), (',', ','), ('concluded',
'VBD'), ('that', 'IN'), ('``', '``'), ('outright', 'JJ'), ('cheating', 'NN'),
('by', 'IN'), ('American', 'JJ'), ('educators', 'NNS'), ("''", "''"), ('is',
'VBZ'), ('``', '``'), ('common', 'JJ'), ('.', '.'), ("''", "''")], [('Mr.',
'NNP'), ('Dinkins', 'NNP'), ('did', 'VBD'), ('fail', 'VB'), ('*-1', '-NONE-'),
('to', 'TO'), ('file', 'VB'), ('his', 'PRP$'), ('income', 'NN'), ('taxes',
'NNS'), ('for', 'IN'), ('four', 'CD'), ('years', 'NNS'), (',', ','), ('but',
'CC'), ('he', 'PRP'), ('insists', 'VBZ'), ('0', '-NONE-'), ('he', 'PRP'),
('voluntarily', 'RB'), ('admitted', 'VBD'), ('the', 'DT'), ('``', '``'),
('oversight', 'NN'), ("''", "''"), ('when', 'WRB'), ('he', 'PRP'), ('was',
'VBD'), ('being', 'VBG'), ('considered', 'VBN'), ('*-2', '-NONE-'), ('for',
'IN'), ('a', 'DT'), ('city', 'NN'), ('job', 'NN'), ('*T*-3', '-NONE-'), ('.',
'.')], [('Soon', 'RB'), (',', ','), ('T-shirts', 'NNS'), ('*ICH*-1', '-NONE-'),
('appeared', 'VBD'), ('in', 'IN'), ('the', 'DT'), ('corridors', 'NNS'), ('that',
'WDT'), ('*T*-2', '-NONE-'), ('carried', 'VBD'), ('the', 'DT'), ('school',
'NN'), ("'s", 'POS'), ('familiar', 'JJ'), ('red-and-white', 'JJ'), ('GHS',
'NNP'), ('logo', 'NN'), ('on', 'IN'), ('the', 'DT'), ('front', 'NN'), ('.',
'.')], [('There', 'EX'), ('is', 'VBZ'), ('$', '$'), ('81.8', 'CD'), ('million',
'CD'), ('*U*', '-NONE-'), ('of', 'IN'), ('7.20', 'CD'), ('%', 'NN'), ('term',
'NN'), ('bonds', 'NNS'), ('due', 'JJ'), ('2009', 'CD'), ('priced', 'VBN'), ('*',
'-NONE-'), ('at', 'IN'), ('99', 'CD'), ('1\\/4', 'CD'), ('*', '-NONE-'), ('to',
'TO'), ('yield', 'VB'), ('7.272', 'CD'), ('%', 'NN'), ('.', '.')],
[('McDermott', 'NNP'), ('International', 'NNP'), ('Inc.', 'NNP'), ('said',
'VBD'), ('0', '-NONE-'), ('its', 'PRP$'), ('Babcock', 'NNP'), ('&', 'CC'),
('Wilcox', 'NNP'), ('unit', 'NN'), ('completed', 'VBD'), ('the', 'DT'), ('sale',
'NN'), ('of', 'IN'), ('its', 'PRP$'), ('Bailey', 'NNP'), ('Controls', 'NNP'),
('Operations', 'NNP'), ('to', 'TO'), ('Finmeccanica', 'NNP'), ('S.p', 'NNP'),
('.', '.'), ('A.', 'NNP'), ('for', 'IN'), ('$', '$'), ('295', 'CD'), ('million',
'CD'), ('*U*', '-NONE-'), ('.', '.')], [('A', 'DT'), ('buffet', 'NN'),
('breakfast', 'NN'), ('was', 'VBD'), ('held', 'VBN'), ('*-1', '-NONE-'), ('in',
'IN'), ('the', 'DT'), ('museum', 'NN'), (',', ','), ('where', 'WRB'), ('food',
'NN'), ('and', 'CC'), ('drinks', 'NNS'), ('are', 'VBP'), ('banned', 'VBN'),
('*-2', '-NONE-'), ('to', 'TO'), ('everyday', 'JJ'), ('visitors', 'NNS'),
('*T*-3', '-NONE-'), ('.', '.')], [('In', 'IN'), ('the', 'DT'), ('year-ago',
'JJ'), ('quarter', 'NN'), (',', ','), ('the', 'DT'), ('company', 'NN'),
('reported', 'VBD'), ('net', 'JJ'), ('income', 'NN'), ('of', 'IN'), ('$', '$'),
('1.9', 'CD'), ('million', 'CD'), ('*U*', '-NONE-'), (',', ','), ('or', 'CC'),
('29', 'CD'), ('cents', 'NNS'), ('a', 'DT'), ('share', 'NN'), ('.', '.')],
[('Mr.', 'NNP'), ('Baldwin', 'NNP'), ('is', 'VBZ'), ('also', 'RB'),
('attacking', 'VBG'), ('the', 'DT'), ('greater', 'JJR'), ('problem', 'NN'),
(':', ':'), ('lack', 'NN'), ('of', 'IN'), ('ringers', 'NNS'), ('.', '.')],
[('The', 'DT'), ('rest', 'NN'), ('were', 'VBD'), ('history', 'NN'), (',', ','),
('sociology', 'NN'), (',', ','), ('finance', 'NN'), ('--', ':'), ('subjects',
'NNS'), ('0', '-NONE-'), ('they', 'PRP'), ('never', 'RB'), ('had', 'VBD'),
('*T*-1', '-NONE-'), ('.', '.'), ("''", "''")], [('During', 'IN'), ('the',
'DT'), ('current', 'JJ'), ('crop', 'NN'), ('year', 'NN'), (',', ','), ('Brazil',
'NNP'), ('was', 'VBD'), ('expected', 'VBN'), ('*-1', '-NONE-'), ('to', 'TO'),
('produce', 'VB'), ('6.9', 'CD'), ('million', 'CD'), ('tons', 'NNS'), ('of',
'IN'), ('sugar', 'NN'), (',', ','), ('a', 'DT'), ('drop', 'NN'), ('from', 'IN'),
('8.1', 'CD'), ('million', 'CD'), ('tons', 'NNS'), ('in', 'IN'), ('1988-89',
'CD'), ('.', '.')], [('Hudson', 'NNP'), ('General', 'NNP'), (',', ','),
('which', 'WDT'), ('*T*-195', '-NONE-'), ('provides', 'VBZ'), ('maintenance',
'NN'), (',', ','), ('fueling', 'NN'), ('and', 'CC'), ('other', 'JJ'),
('services', 'NNS'), ('to', 'TO'), ('airlines', 'NNS'), ('and', 'CC'),
('airports', 'NNS'), (',', ','), ('reported', 'VBD'), ('a', 'DT'), ('loss',
'NN'), ('for', 'IN'), ('its', 'PRP$'), ('most', 'RBS'), ('recent', 'JJ'),
('fiscal', 'NN'), ('year', 'NN'), ('and', 'CC'), ('last', 'JJ'), ('month',
'NN'), ('omitted', 'VBD'), ('the', 'DT'), ('semiannual', 'JJ'), ('dividend',
'NN'), ('on', 'IN'), ('its', 'PRP$'), ('common', 'JJ'), ('shares', 'NNS'), ('.',
'.')], [('They', 'PRP'), ('point', 'VBP'), ('out', 'RP'), ('that', 'IN'),
('these', 'DT'), ('institutions', 'NNS'), ('want', 'VBP'), ('*-1', '-NONE-'),
('to', 'TO'), ('lock', 'VB'), ('in', 'RP'), ('returns', 'NNS'), ('on', 'IN'),
('high-yield', 'JJ'), ('U.S.', 'NNP'), ('Treasury', 'NNP'), ('debt', 'NN'),
('and', 'CC'), ('suggest', 'VBP'), ('0', '-NONE-'), ('demand', 'NN'), ('for',
'IN'), ('the', 'DT'), ('U.S.', 'NNP'), ('unit', 'NN'), ('will', 'MD'),
('continue', 'VB'), ('*-2', '-NONE-'), ('unabated', 'JJ'), ('until', 'IN'),
('rates', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('U.S.', 'NNP'), ('recede',
'VBP'), ('.', '.')], [('Except', 'IN'), ('where', 'WRB'), ('*', '-NONE-'),
('noted', 'VBN'), ('*-3', '-NONE-'), ('*T*-1', '-NONE-'), (',', ','), ('none',
'NN'), ('of', 'IN'), ('these', 'DT'), ('people', 'NNS'), ('could', 'MD'), ('be',
'VB'), ('reached', 'VBN'), ('*-2', '-NONE-'), ('for', 'IN'), ('comment', 'NN'),
('or', 'CC'), ('had', 'VBD'), ('any', 'DT'), ('comment', 'NN'), ('.', '.')],
[('McGraw-Hill', 'NNP'), ('was', 'VBD'), ('outraged', 'JJ'), ('.', '.')],
[('American', 'NNP'), ('Express', 'NNP'), ('also', 'RB'), ('represents', 'VBZ'),
('the', 'DT'), ('upscale', 'NN'), ('image', 'NN'), ('0', '-NONE-'), ('``',
'``'), ('we', 'PRP'), ("'re", 'VBP'), ('trying', 'VBG'), ('*-2', '-NONE-'),
('to', 'TO'), ('project', 'VB'), ('*T*-1', '-NONE-'), (',', ','), ("''", "''"),
('she', 'PRP'), ('adds', 'VBZ'), ('*T*-3', '-NONE-'), ('.', '.')], [('When',
'WRB'), ('it', 'PRP'), ("'s", 'VBZ'), ('time', 'NN'), ('for', 'IN'), ('their',
'PRP$'), ('biannual', 'JJ'), ('powwow', 'NN'), ('*T*-1', '-NONE-'), (',', ','),
('the', 'DT'), ('nation', 'NN'), ("'s", 'POS'), ('manufacturing', 'VBG'),
('titans', 'NNS'), ('typically', 'RB'), ('jet', 'VBP'), ('off', 'RP'), ('to',
'TO'), ('the', 'DT'), ('sunny', 'JJ'), ('confines', 'NNS'), ('of', 'IN'),
('resort', 'NN'), ('towns', 'NNS'), ('like', 'IN'), ('Boca', 'NNP'), ('Raton',
'NNP'), ('and', 'CC'), ('Hot', 'NNP'), ('Springs', 'NNP'), ('.', '.')],
[('Scott', 'NNP'), ('Taccetta', 'NNP'), (',', ','), ('a', 'DT'), ('Chicago',
'NNP'), ('accountant', 'NN'), (',', ','), ('is', 'VBZ'), ('going', 'VBG'),
('into', 'IN'), ('money-market', 'JJ'), ('funds', 'NNS'), ('.', '.')], [('In',
'IN'), ('October', 'NNP'), (',', ','), ('before', 'IN'), ('the', 'DT'),
('market', 'NN'), ('dropped', 'VBD'), (',', ','), ('Mrs.', 'NNP'), ('Arighi',
'NNP'), ('of', 'IN'), ('Arnold', 'NNP'), (',', ','), ('Calif.', 'NNP'), (',',
','), ('moved', 'VBD'), ('*-1', '-NONE-'), ('to', 'TO'), ('sell', 'VB'), ('the',
'DT'), ('``', '``'), ('speculative', 'JJ'), ('stocks', 'NNS'), ("''", "''"),
('in', 'IN'), ('her', 'PRP$'), ('family', 'NN'), ('trust', 'NN'), ('``', '``'),
('so', 'IN'), ('we', 'PRP'), ('will', 'MD'), ('be', 'VB'), ('able', 'JJ'),
('*-2', '-NONE-'), ('to', 'TO'), ('withstand', 'VB'), ('all', 'PDT'), ('this',
'DT'), ('flim-flammery', 'NN'), ("''", "''"), ('caused', 'VBN'), ('*',
'-NONE-'), ('by', 'IN'), ('program', 'NN'), ('trading', 'NN'), ('.', '.')],
[('Without', 'IN'), ('the', 'DT'), ('Cray-3', 'NNP'), ('research', 'NN'),
('and', 'CC'), ('development', 'NN'), ('expenses', 'NNS'), (',', ','), ('the',
'DT'), ('company', 'NN'), ('would', 'MD'), ('have', 'VB'), ('been', 'VBN'),
('able', 'JJ'), ('*-2', '-NONE-'), ('to', 'TO'), ('report', 'VB'), ('a', 'DT'),
('profit', 'NN'), ('of', 'IN'), ('$', '$'), ('19.3', 'CD'), ('million', 'CD'),
('*U*', '-NONE-'), ('*ICH*-3', '-NONE-'), ('for', 'IN'), ('the', 'DT'),
('first', 'JJ'), ('half', 'DT'), ('of', 'IN'), ('1989', 'CD'), ('rather', 'RB'),
('than', 'IN'), ('the', 'DT'), ('$', '$'), ('5.9', 'CD'), ('million', 'CD'),
('*U*', '-NONE-'), ('0', '-NONE-'), ('it', 'PRP'), ('posted', 'VBD'), ('*T*-1',
'-NONE-'), ('.', '.')], [('The', 'DT'), ('Treasury', 'NNP'), ('plans', 'VBZ'),
('*-1', '-NONE-'), ('to', 'TO'), ('sell', 'VB'), ('$', '$'), ('30', 'CD'),
('billion', 'CD'), ('*U*', '-NONE-'), ('in', 'IN'), ('notes', 'NNS'), ('and',
'CC'), ('bonds', 'NNS'), ('next', 'IN'), ('week', 'NN'), ('but', 'CC'), ('will',
'MD'), ('delay', 'VB'), ('the', 'DT'), ('auction', 'NN'), ('unless', 'IN'),
('Congress', 'NNP'), ('quickly', 'RB'), ('raises', 'VBZ'), ('the', 'DT'),
('debt', 'NN'), ('ceiling', 'NN'), ('.', '.')], [('B.A.T', 'NNP'),
('Industries', 'NNPS'), (',', ','), ('which', 'WDT'), ('*T*-2', '-NONE-'),
('is', 'VBZ'), ('being', 'VBG'), ('pursued', 'VBN'), ('*-1', '-NONE-'), ('by',
'IN'), ('Sir', 'NNP'), ('James', 'NNP'), ('Goldsmith', 'NNP'), ("'s", 'POS'),
('Hoylake', 'NNP'), ('Investments', 'NNPS'), (',', ','), ('rose', 'VBD'), ('9',
'CD'), ('to', 'TO'), ('753', 'CD'), ('on', 'IN'), ('speculation', 'NN'),
('that', 'IN'), ('Hoylake', 'NNP'), ('will', 'MD'), ('sweeten', 'VB'), ('its',
'PRP$'), ('bid', 'NN'), (',', ','), ('dealers', 'NNS'), ('said', 'VBD'), ('0',
'-NONE-'), ('*T*-3', '-NONE-'), ('.', '.')], [('If', 'IN'), ('we', 'PRP'),
('look', 'VBP'), ('to', 'TO'), ('the', 'DT'), ('future', 'NN'), (',', ','),
('*', '-NONE-'), ('preventing', 'VBG'), ('homelessness', 'NN'), ('is', 'VBZ'),
('an', 'DT'), ('important', 'JJ'), ('objective', 'NN'), ('.', '.')], [('With',
'IN'), ('the', 'DT'), ('harvest', 'NN'), ('winding', 'VBG'), ('down', 'IN'),
(',', ','), ('however', 'RB'), (',', ','), ('some', 'DT'), ('analysts', 'NNS'),
('are', 'VBP'), ('speculating', 'VBG'), ('that', 'IN'), ('prices', 'NNS'),
('might', 'MD'), ('jump', 'VB'), ('in', 'IN'), ('some', 'DT'), ('regions',
'NNS'), ('as', 'IN'), ('U.S.', 'NNP'), ('exporters', 'NNS'), ('try', 'VBP'),
('*-1', '-NONE-'), ('to', 'TO'), ('gather', 'VB'), ('the', 'DT'), ('corn',
'NN'), ('0', '-NONE-'), ('they', 'PRP'), ('are', 'VBP'), ('obligated', 'VBN'),
('*-3', '-NONE-'), ('to', 'TO'), ('deliver', 'VB'), ('*T*-2', '-NONE-'), ('.',
'.')], [('According', 'VBG'), ('to', 'TO'), ('an', 'DT'), ('American', 'JJ'),
('member', 'NN'), ('of', 'IN'), ('the', 'DT'), ('Nixon', 'NNP'), ('party',
'NN'), (',', ','), ('the', 'DT'), ('former', 'JJ'), ('president', 'NN'),
('raised', 'VBD'), ('a', 'DT'), ('number', 'NN'), ('of', 'IN'),
('controversial', 'JJ'), ('issues', 'NNS'), ('in', 'IN'), ('his', 'PRP$'),
('20', 'CD'), ('hours', 'NNS'), ('of', 'IN'), ('talks', 'NNS'), ('with', 'IN'),
('top-level', 'JJ'), ('Chinese', 'JJ'), ('officials', 'NNS'), ('.', '.')],
[('That', 'DT'), ("'s", 'VBZ'), ('not', 'RB'), ('*', '-NONE-'), ('to', 'TO'),
('say', 'VB'), ('that', 'IN'), ('the', 'DT'), ('nutty', 'JJ'), ('plot', 'NN'),
('of', 'IN'), ('``', '``'), ('A', 'DT'), ('Wild', 'NNP'), ('Sheep', 'NNP'),
('Chase', 'NNP'), ("''", "''"), ('is', 'VBZ'), ('rooted', 'VBN'), ('*-57',
'-NONE-'), ('in', 'IN'), ('reality', 'NN'), ('.', '.')], [('J.P.', 'NNP'),
('Bolduc', 'NNP'), (',', ','), ('vice', 'NN'), ('chairman', 'NN'), ('of', 'IN'),
('W.R.', 'NNP'), ('Grace', 'NNP'), ('&', 'CC'), ('Co.', 'NNP'), (',', ','),
('which', 'WDT'), ('*T*-10', '-NONE-'), ('holds', 'VBZ'), ('a', 'DT'), ('83.4',
'CD'), ('%', 'NN'), ('interest', 'NN'), ('in', 'IN'), ('this', 'DT'), ('energy-
services', 'JJ'), ('company', 'NN'), (',', ','), ('was', 'VBD'), ('elected',
'VBN'), ('*-10', '-NONE-'), ('a', 'DT'), ('director', 'NN'), ('.', '.')],
[('In', 'IN'), ('June', 'NNP'), ('1988', 'CD'), (',', ','), ('I', 'PRP'),
('wrote', 'VBD'), ('in', 'IN'), ('this', 'DT'), ('space', 'NN'), ('about',
'IN'), ('this', 'DT'), ('issue', 'NN'), ('.', '.')], [('The', 'DT'), ('other',
'JJ'), ('concern', 'NN'), ('was', 'VBD'), ("n't", 'RB'), ('identified', 'VBD'),
('.', '.')], [('Tokyu', 'NNP'), ('Group', 'NNP'), (',', ','), ('Mitsubishi',
'NNP'), ('Estate', 'NNP'), ('and', 'CC'), ('Bridgestone\\/Firestone', 'NNP'),
(',', ','), ('which', 'WDT'), ('*T*-1', '-NONE-'), ('advanced', 'VBD'),
('Tuesday', 'NNP'), (',', ','), ('declined', 'VBD'), ('on', 'IN'), ('profit-
taking', 'NN'), ('.', '.')], [('Along', 'IN'), ('with', 'IN'), ('the', 'DT'),
('note', 'NN'), (',', ','), ('Cray', 'NNP'), ('Research', 'NNP'), ('is', 'VBZ'),
('transferring', 'VBG'), ('about', 'IN'), ('$', '$'), ('53', 'CD'), ('million',
'CD'), ('*U*', '-NONE-'), ('in', 'IN'), ('assets', 'NNS'), (',', ','),
('primarily', 'RB'), ('those', 'DT'), ('related', 'VBN'), ('to', 'TO'), ('the',
'DT'), ('Cray-3', 'CD'), ('development', 'NN'), (',', ','), ('which', 'WDT'),
('*T*-25', '-NONE-'), ('has', 'VBZ'), ('been', 'VBN'), ('a', 'DT'), ('drain',
'NN'), ('on', 'IN'), ('Cray', 'NNP'), ('Research', 'NNP'), ("'s", 'POS'),
('earnings', 'NNS'), ('.', '.')], [('Some', 'DT'), ('long-tenured', 'JJ'),
('employees', 'NNS'), ('will', 'MD'), ('receive', 'VB'), ('additional', 'JJ'),
('benefits', 'NNS'), (',', ','), ('the', 'DT'), ('company', 'NN'), ('said',
'VBD'), ('0', '-NONE-'), ('*T*-1', '-NONE-'), ('.', '.')], [('It', 'PRP'),
('rose', 'VBD'), ('largely', 'RB'), ('throughout', 'IN'), ('the', 'DT'),
('session', 'NN'), ('after', 'IN'), ('*-1', '-NONE-'), ('posting', 'VBG'),
('an', 'DT'), ('intraday', 'NN'), ('low', 'JJ'), ('of', 'IN'), ('2141.7', 'CD'),
('in', 'IN'), ('the', 'DT'), ('first', 'JJ'), ('40', 'CD'), ('minutes', 'NNS'),
('of', 'IN'), ('trading', 'NN'), ('.', '.')], [('``', '``'), ('What', 'WP'),
('sector', 'NN'), ('is', 'VBZ'), ('*T*-46', '-NONE-'), ('stepping', 'VBG'),
('forward', 'RB'), ('*-2', '-NONE-'), ('to', 'TO'), ('pick', 'VB'), ('up',
'RP'), ('the', 'DT'), ('slack', 'NN'), ('?', '.'), ("''", "''"), ('he', 'PRP'),
('asked', 'VBD'), ('*T*-1', '-NONE-'), ('.', '.')], [('And', 'CC'), ('I',
'PRP'), ('apparently', 'RB'), ('had', 'VBD'), ('no', 'DT'), ('right', 'NN'),
('*', '-NONE-'), ('to', 'TO'), ('print', 'VB'), ('hither', 'RB'), ('what',
'WP'), ('the', 'DT'), ('Voice', 'NNP'), ('was', 'VBD'), ('booming', 'VBG'),
('*T*-2', '-NONE-'), ('to', 'TO'), ('yon', 'RB'), ('.', '.')], [('But', 'CC'),
('the', 'DT'), ('administration', 'NN'), ("'s", 'POS'), ('handling', 'NN'),
('of', 'IN'), ('the', 'DT'), ('fetal-tissue', 'JJ'), ('transplant', 'NN'),
('issue', 'NN'), ('disturbs', 'VBZ'), ('many', 'JJ'), ('scientists', 'NNS'),
('.', '.')], [('But', 'CC'), ('he', 'PRP'), ('has', 'VBZ'), ('not', 'RB'),
('said', 'VBD'), ('before', 'IN'), ('that', 'IN'), ('the', 'DT'), ('country',
'NN'), ('wants', 'VBZ'), ('half', 'PDT'), ('the', 'DT'), ('debt', 'NN'),
('forgiven', 'VBN'), ('*-2', '-NONE-'), ('.', '.')], [('``', '``'), ('If',
'IN'), ('you', 'PRP'), ('continue', 'VBP'), ('*-2', '-NONE-'), ('to', 'TO'),
('do', 'VB'), ('this', 'DT'), (',', ','), ('the', 'DT'), ('investor', 'NN'),
('*ICH*-1', '-NONE-'), ('becomes', 'VBZ'), ('frightened', 'VBN'), ('--', ':'),
('any', 'DT'), ('investor', 'NN'), (':', ':'), ('the', 'DT'), ('odd', 'JJ'),
('lotter', 'NN'), (',', ','), ('mutual', 'JJ'), ('funds', 'NNS'), ('and', 'CC'),
('pension', 'NN'), ('funds', 'NNS'), (',', ','), ("''", "''"), ('says', 'VBZ'),
('*T*-3', '-NONE-'), ('Larry', 'NNP'), ('Zicklin', 'NNP'), (',', ','),
('managing', 'VBG'), ('partner', 'NN'), ('at', 'IN'), ('Neuberger', 'NNP'),
('&', 'CC'), ('Berman', 'NNP'), ('.', '.')], [('Meanwhile', 'RB'), (',', ','),
('most', 'RBS'), ('investment-grade', 'JJ'), ('bonds', 'NNS'), ('ended', 'VBD'),
('unchanged', 'JJ'), ('to', 'TO'), ('as', 'RB'), ('much', 'JJ'), ('as', 'IN'),
('1\\/8', 'CD'), ('point', 'NN'), ('higher', 'JJR'), ('.', '.')]]
```

[7]:
```
#From the above output, we can observe that the total number of training␣
↪records is 2739, and the test set has 1175.
```

[9]:
```
#will check the number of tagged words in the training set to understand how␣
↪much data will be used for training the POS tagger
```

[10]:
```
# Getting list of tagged words
train_tagged_words = [tup for sent in train_set for tup in sent]
len(train_tagged_words)
```

[10]:
```
69935
```

[11]:
```
#will create a tokens variable that will contain all the tokens from the␣
↪train_tagged_words
```

[12]:
```
# tokens
tokens = [pair[0] for pair in train_tagged_words]
# vocabulary
V = set(tokens)
print("Total vocabularies: ",len(V))
# number of tags
T = set([pair[1] for pair in train_tagged_words])
print("Total tags: ",len(T))
```

```
Total vocabularies:  10262
Total tags:  45
```

[13]:
```
#will use HMM algorithm to tag the words.
```

[14]:
```
#P(w/t) is basically the probability that given a tag (say NN), what is the␣
↪probability of it being w (say 'building').
```

```
#This can be computed by computing the fraction of all NNs which are equal to␣
 ↪w, i.e.P(w/t) = count(w, t) / count(t).
```

[15]:

```
#The term P(t) is the probability of tag t, and in a tagging task, we assume␣
 ↪that a tag will depend only on the previous tag
```

[16]:

```
#Emission probabilities
```

[17]:

```
# computing P(w/t) and storing in T x V matrix
t = len(T)
v = len(V)
w_given_t = np.zeros((t, v))
# compute word given tag: Emission Probability
def word_given_tag(word, tag, train_bag = train_tagged_words):
    tag_list = [pair for pair in train_bag if pair[1]==tag]
    count_tag = len(tag_list)
    w_given_tag_list = [pair[0] for pair in tag_list if pair[0]==word]
    count_w_given_tag = len(w_given_tag_list)

    return (count_w_given_tag, count_tag)
# examples
# large
print("\n", "large")
print(word_given_tag('large', 'JJ'))
print(word_given_tag('large', 'VB'))
print(word_given_tag('large', 'NN'), "\n")
# will
print("\n", "will")
print(word_given_tag('will', 'MD'))
print(word_given_tag('will', 'NN'))
print(word_given_tag('will', 'VB'))
# book
print("\n", "book")
print(word_given_tag('book', 'NN'))
print(word_given_tag('book', 'VB'))
```

```

 large
(19, 4042)
(0, 1765)
(0, 9045)



 will
(192, 633)
(1, 9045)
(0, 1765)
```

```
 book
(4, 9045)
(1, 1765)
```

[18]:

```
#Transition Probabilities
```

[19]:

```
# compute tag given tag: tag2(t2) given tag1 (t1), i.e. Transition Probability
def t2_given_t1(t2, t1, train_bag = train_tagged_words):
    tags = [pair[1] for pair in train_bag]
    count_t1 = len([t for t in tags if t==t1])
    count_t2_t1 = 0
    for index in range(len(tags)-1):
        if tags[index]==t1 and tags[index+1] == t2:
            count_t2_t1 += 1
    return (count_t2_t1, count_t1)
# examples
print(t2_given_t1(t2='NNP', t1='JJ'))
print(t2_given_t1('NN', 'JJ'))
print(t2_given_t1('NN', 'DT'))
print(t2_given_t1('NNP', 'VB'))
print(t2_given_t1(',', 'NNP'))
print(t2_given_t1('PRP', 'PRP'))
print(t2_given_t1('VBG', 'NNP'))
```

```
(152, 4042)
(1793, 4042)
(2662, 5658)
(62, 1765)
(1057, 6734)
(2, 1157)
(3, 6734)
```

[20]:

```
#Please note P(tag|start) is same as P(tag|'.')
print(t2_given_t1('DT', '.'))
print(t2_given_t1('VBG', '.'))
print(t2_given_t1('NN', '.'))
print(t2_given_t1('NNP', '.'))
```

```
(582, 2707)
(11, 2707)
(98, 2707)
(522, 2707)
```

[21]:

```
#Next, we will create a transition matrix of tags of dimension txt
```

[22]:

```
# creating t x t transition matrix of tags
# each column is t2, each row is t1
# thus M(i, j) represents P(tj given ti)
tags_matrix = np.zeros((len(T), len(T)), dtype='float32')
```

```
for i, t1 in enumerate(list(T)):
    for j, t2 in enumerate(list(T)):
        tags_matrix[i, j] = t2_given_t1(t2, t1)[0]/t2_given_t1(t2, t1)[1]
tags_matrix
```

```
[22]: array([[0.0000000e+00, 0.0000000e+00, 0.0000000e+00, …, 0.0000000e+00,
              0.0000000e+00, 3.9215688e-02],
             [0.0000000e+00, 0.0000000e+00, 6.3291140e-02, …, 0.0000000e+00,
              0.0000000e+00, 3.7974682e-02],
             [1.7686425e-02, 0.0000000e+00, 0.0000000e+00, …, 0.0000000e+00,
              2.3900573e-03, 2.8680688e-02],
             …,
             [0.0000000e+00, 0.0000000e+00, 1.1940298e-01, …, 0.0000000e+00,
              0.0000000e+00, 0.0000000e+00],
             [2.2662889e-02, 0.0000000e+00, 5.6657224e-04, …, 0.0000000e+00,
              1.6997167e-03, 6.4589232e-02],
             [4.4223329e-04, 1.3266999e-03, 4.6323936e-02, …, 1.1055832e-04,
              9.9502492e-04, 1.2404644e-01]], dtype=float32)
```

```
[23]: #As tags are not visible in this matrix, we will now convert it into pandas␣
       ↪dataframe for better readability.
```

```
[24]: # convert the matrix to a df for better readability
      tags_df = pd.DataFrame(tags_matrix, columns = list(T), index=list(T))
      tags_df
```

```
[24]:               RP     -RRB-       VBD     -LRB-        MD    -NONE-        DT  \
      RP      0.000000  0.000000  0.000000  0.000000  0.000000  0.104575  0.189542
      -RRB-   0.000000  0.000000  0.063291  0.000000  0.000000  0.050633  0.063291
      VBD     0.017686  0.000000  0.000000  0.000000  0.000000  0.269598  0.129063
      -LRB-   0.000000  0.000000  0.000000  0.000000  0.000000  0.013158  0.092105
      MD      0.000000  0.000000  0.000000  0.000000  0.000000  0.004739  0.001580
      -NONE-  0.001089  0.004356  0.030930  0.001307  0.014376  0.070355  0.051623
      DT      0.000000  0.000177  0.001944  0.000353  0.001767  0.001944  0.001414
      CC      0.000000  0.000000  0.039128  0.000000  0.011546  0.008980  0.117383
      .       0.000000  0.003694  0.000000  0.003694  0.000000  0.020687  0.214998
      LS      0.000000  0.428571  0.000000  0.000000  0.000000  0.000000  0.000000
      WP      0.000000  0.000000  0.000000  0.000000  0.000000  0.800000  0.037500
      :       0.000000  0.000000  0.025707  0.000000  0.015424  0.030848  0.113111
      VBN     0.011572  0.000000  0.000681  0.000000  0.000000  0.563649  0.046971
      TO      0.000000  0.000000  0.000000  0.000000  0.000000  0.009138  0.129896
      NNP     0.000000  0.003267  0.062816  0.002673  0.009801  0.005643  0.002673
      WRB     0.000000  0.000000  0.015038  0.000000  0.007519  0.052632  0.300752
      VBZ     0.010239  0.000000  0.001365  0.000000  0.000000  0.181570  0.143345
      NNS     0.000239  0.000718  0.075377  0.003111  0.027758  0.041158  0.014836
      RB      0.000000  0.000000  0.065900  0.000000  0.006799  0.023013  0.056485
      VBG     0.016537  0.000973  0.001946  0.000000  0.000000  0.076848  0.184825
```

```
POS     0.000000  0.001721  0.003442  0.000000  0.000000  0.000000  0.000000
VBP     0.012022  0.000000  0.001093  0.000000  0.000000  0.171585  0.104918
JJS     0.000000  0.000000  0.008065  0.000000  0.000000  0.008065  0.016129
#       0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
WP$     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
``      0.000000  0.000000  0.004310  0.000000  0.008621  0.036638  0.174569
IN      0.000146  0.000000  0.000730  0.000000  0.000000  0.034156  0.318932
$       0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
JJR     0.000000  0.000000  0.003953  0.000000  0.000000  0.019763  0.011858
RBS     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
CD      0.000000  0.000789  0.007101  0.001183  0.001972  0.223274  0.000394
''      0.000000  0.002151  0.073118  0.004301  0.004301  0.015054  0.116129
UH      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
,       0.000000  0.000000  0.054402  0.000585  0.009067  0.031881  0.133372
JJ      0.000247  0.000247  0.000990  0.000247  0.000000  0.023008  0.003216
RBR     0.000000  0.000000  0.010526  0.010526  0.000000  0.031579  0.052632
PRP     0.003457  0.000864  0.260156  0.001729  0.131374  0.036301  0.010372
FW      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
WDT     0.000000  0.000000  0.006579  0.000000  0.003289  0.881579  0.019737
PDT     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.900000
PRP$    0.000000  0.000000  0.000000  0.001942  0.000000  0.000000  0.000000
NNPS    0.000000  0.006098  0.036585  0.000000  0.036585  0.012195  0.000000
EX      0.000000  0.000000  0.119403  0.000000  0.044776  0.000000  0.000000
VB      0.022663  0.000000  0.000567  0.001133  0.000000  0.078187  0.231161
NN      0.000442  0.001327  0.046324  0.001437  0.014704  0.040796  0.005528

              CC         .        LS  …       RBR       PRP        FW  \
RP      0.006536  0.026144  0.000000  …  0.000000  0.000000  0.000000
-RRB-   0.050633  0.126582  0.000000  …  0.000000  0.012658  0.000000
VBD     0.002390  0.007648  0.000000  …  0.003824  0.011950  0.000000
-LRB-   0.026316  0.000000  0.000000  …  0.000000  0.026316  0.000000
MD      0.000000  0.000000  0.000000  …  0.000000  0.001580  0.000000
-NONE-  0.011544  0.092355  0.000000  …  0.001307  0.048791  0.000000
DT      0.000177  0.001060  0.000000  …  0.001414  0.000353  0.000177
CC      0.000641  0.000000  0.000641  …  0.001283  0.042335  0.000000
.       0.050610  0.000000  0.001478  …  0.001108  0.058737  0.000000
LS      0.000000  0.285714  0.000000  …  0.000000  0.000000  0.000000
WP      0.000000  0.000000  0.000000  …  0.000000  0.043750  0.000000
:       0.053985  0.007712  0.002571  …  0.000000  0.028278  0.000000
VBN     0.007488  0.008850  0.000000  …  0.001361  0.002723  0.000000
TO      0.000000  0.000000  0.000000  …  0.001958  0.005875  0.000000
NNP     0.036828  0.050490  0.000000  …  0.000000  0.000594  0.000000
WRB     0.000000  0.000000  0.000000  …  0.000000  0.157895  0.000000
VBZ     0.004096  0.002048  0.000000  …  0.002048  0.015017  0.000000
NNS     0.055755  0.120124  0.000000  …  0.001436  0.001196  0.000000
RB      0.007845  0.042364  0.000000  …  0.006799  0.004707  0.000000
VBG     0.010700  0.016537  0.000000  …  0.002918  0.018482  0.000000
```

```
POS     0.005164  0.010327  0.000000  …  0.000000  0.000000  0.000000
VBP     0.002186  0.007650  0.000000  …  0.004372  0.018579  0.000000
JJS     0.000000  0.024194  0.000000  …  0.000000  0.000000  0.000000
#       0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
WP$     0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
``      0.019397  0.000000  0.000000  …  0.000000  0.196121  0.000000
IN      0.000584  0.002481  0.000000  …  0.000876  0.030652  0.000146
$       0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
JJR     0.031621  0.071146  0.000000  …  0.000000  0.000000  0.000000
RBS     0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
CD      0.012229  0.048126  0.000000  …  0.000394  0.000789  0.000000
''      0.055914  0.002151  0.000000  …  0.000000  0.096774  0.000000
UH      0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
,       0.084235  0.000000  0.000292  …  0.000877  0.037730  0.000000
JJ      0.014597  0.021277  0.000000  …  0.000495  0.000495  0.000000
RBR     0.010526  0.063158  0.000000  …  0.000000  0.000000  0.000000
PRP     0.008643  0.027658  0.000000  …  0.000864  0.001729  0.000000
FW      0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
WDT     0.000000  0.000000  0.000000  …  0.000000  0.026316  0.000000
PDT     0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
PRP$    0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
NNPS    0.067073  0.085366  0.000000  …  0.000000  0.000000  0.000000
EX      0.000000  0.000000  0.000000  …  0.000000  0.000000  0.000000
VB      0.008499  0.012465  0.000000  …  0.007932  0.026629  0.000000
NN      0.037922  0.105252  0.000000  …  0.000774  0.001437  0.000111

             WDT       PDT      PRP$      NNPS        EX        VB        NN
RP      0.000000  0.000000  0.071895  0.000000  0.000000  0.000000  0.039216
-RRB-   0.012658  0.000000  0.000000  0.000000  0.000000  0.000000  0.037975
VBD     0.000000  0.000956  0.016730  0.000000  0.000000  0.002390  0.028681
-LRB-   0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.039474
MD      0.000000  0.000000  0.000000  0.000000  0.000000  0.819905  0.000000
-NONE-  0.000000  0.000000  0.003267  0.000000  0.001307  0.009148  0.021128
DT      0.000353  0.000000  0.000000  0.003358  0.000000  0.000000  0.470484
CC      0.001283  0.000000  0.016036  0.002566  0.005773  0.033355  0.114817
.       0.000739  0.000739  0.007758  0.001847  0.005541  0.000739  0.036202
LS      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
WP      0.000000  0.006250  0.018750  0.000000  0.000000  0.000000  0.018750
:       0.005141  0.000000  0.000000  0.000000  0.000000  0.002571  0.048843
VBN     0.000000  0.000000  0.011572  0.000000  0.000000  0.000000  0.070796
TO      0.000000  0.000000  0.011749  0.000000  0.000000  0.582245  0.026762
NNP     0.000594  0.000000  0.000000  0.015890  0.000000  0.001040  0.056133
WRB     0.000000  0.007519  0.007519  0.000000  0.000000  0.000000  0.060150
VBZ     0.000000  0.000683  0.008191  0.000000  0.000000  0.002730  0.038908
NNS     0.014597  0.000000  0.000000  0.000000  0.000000  0.003829  0.021058
RB      0.000523  0.000523  0.001569  0.000523  0.001046  0.097803  0.016736
VBG     0.000000  0.000973  0.025292  0.000000  0.000000  0.000000  0.143969
```

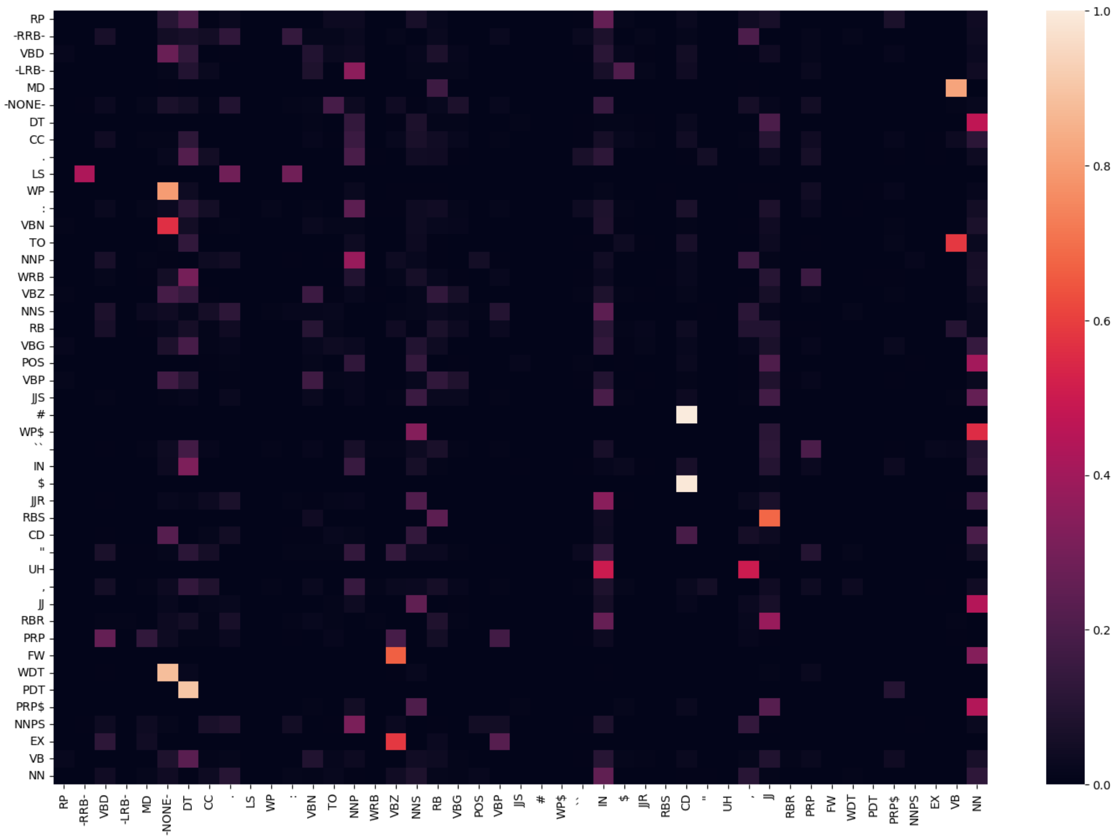
```
POS     0.000000  0.000000  0.000000  0.005164  0.000000  0.000000  0.406196
VBP     0.001093  0.000000  0.007650  0.000000  0.000000  0.001093  0.026230
JJS     0.000000  0.000000  0.000000  0.000000  0.000000  0.008065  0.258065
#       0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
WP$     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.555556
``      0.000000  0.000000  0.004310  0.000000  0.021552  0.019397  0.092672
IN      0.002773  0.000730  0.034010  0.002335  0.001314  0.000000  0.108305
$       0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
JJR     0.000000  0.000000  0.000000  0.000000  0.000000  0.003953  0.169960
RBS     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
CD      0.001972  0.000000  0.000394  0.000000  0.000000  0.000000  0.192505
''      0.012903  0.000000  0.002151  0.000000  0.000000  0.004301  0.053763
UH      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
,       0.035098  0.000000  0.003217  0.000000  0.004095  0.001170  0.045627
JJ      0.000000  0.000000  0.000000  0.001237  0.000000  0.000000  0.443592
RBR     0.000000  0.000000  0.000000  0.000000  0.000000  0.010526  0.000000
PRP     0.000000  0.000000  0.000000  0.000000  0.000000  0.006050  0.003457
FW      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.333333
WDT     0.000000  0.000000  0.000000  0.000000  0.003289  0.000000  0.003289
PDT     0.000000  0.000000  0.100000  0.000000  0.000000  0.000000  0.000000
PRP$    0.000000  0.000000  0.000000  0.001942  0.000000  0.000000  0.438835
NNPS    0.000000  0.000000  0.000000  0.006098  0.000000  0.000000  0.024390
EX      0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
VB      0.001133  0.003399  0.039660  0.001133  0.000000  0.001700  0.064589
NN      0.008402  0.000000  0.000111  0.000000  0.000111  0.000995  0.124046

[45 rows x 45 columns]
```

[25]:

```
#Next will create a heatmap of the tag matrix
```

[26]:

```
# heatmap of tags matrix
# T(i, j) means P(tag j given tag i)
plt.figure(figsize=(18, 12))
sns.heatmap(tags_df)
plt.show()
```

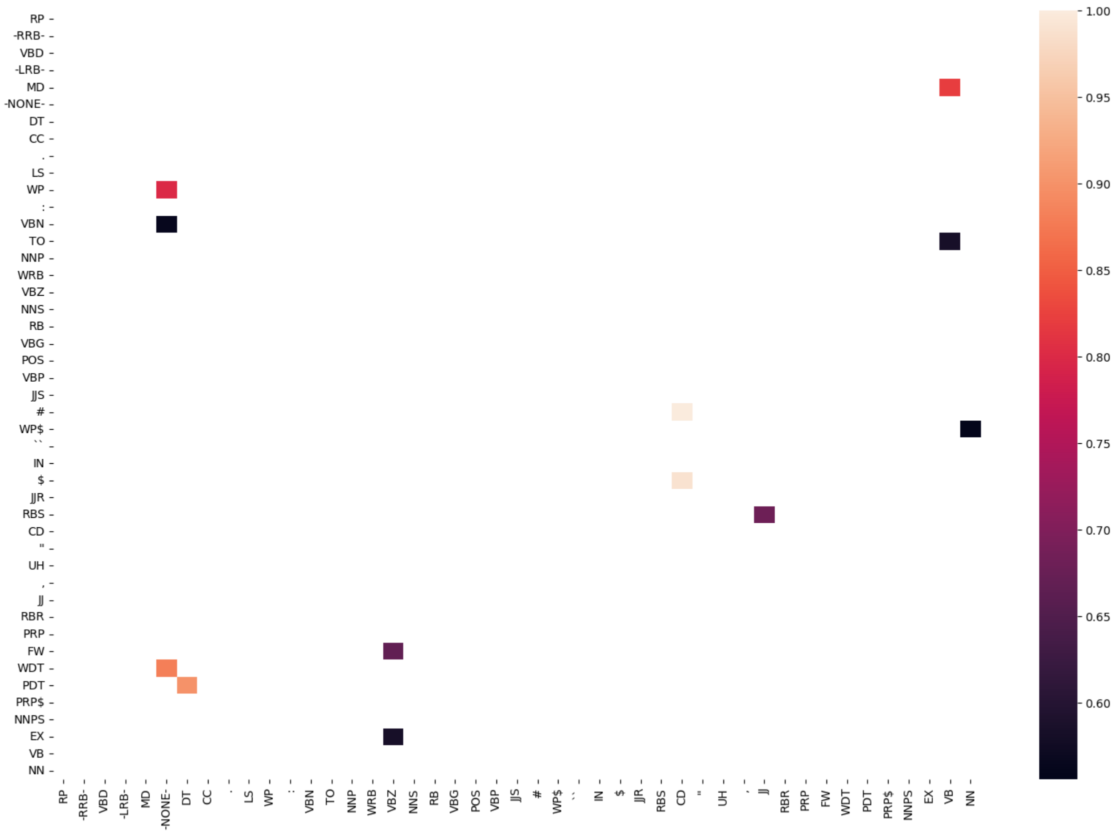


[27]:

```
#Now, in order to see the most frequent tags we have to filter the tags with >0.
↪5 probability
```

[28]:

```
# frequent tags
# filter the df to get P(t2, t1) > 0.5
tags_frequent = tags_df[tags_df>0.5]
plt.figure(figsize=(18, 12))
sns.heatmap(tags_frequent)
plt.show()
```



[29]:

```
#Viterbi Algorithm
```

[30]:

```
#Let's now use the computed probabilities P(w, tag) and P(t2, t1) to assign␣
↪tags to each word in the document. We'll run through each word w and compute␣
↪P(tag/w)=P(w/tag).P(tag) for each tag in the tag set, and then assign the␣
↪tag having
#the max P(tag/w)
```

[31]:

```
#Note: P(tag|start) = P(tag|'.')
```

[32]:

```
# Viterbi Heuristic
def Viterbi(words, train_bag = train_tagged_words):
    state = []
    T = list(set([pair[1] for pair in train_bag]))

    for key, word in enumerate(words):
        #initialise list of probability column for a given observation
        p = []
        for tag in T:
            if key == 0:
                transition_p = tags_df.loc['.', tag]
            else:
                transition_p = tags_df.loc[state[-1], tag]

            # compute emission and state probabilities
            emission_p = word_given_tag(words[key], tag)[0]/
 ↪word_given_tag(words[key], tag)[1]
            state_probability = emission_p * transition_p
            p.append(state_probability)

        pmax = max(p)
        # getting state for which probability is maximum
        state_max = T[p.index(pmax)]
        state.append(state_max)
    return list(zip(words, state))
```

[33]:

```
#Evaluating on Test Set
```

[34]:

```
# Running on entire test dataset would take more than 3-4hrs.
# Let's test our Viterbi algorithm on a few sample sentences of test dataset
random.seed(1234)
# choose random 5 sents
rndom = [random.randint(1,len(test_set)) for x in range(5)]
# list of sents
test_run = [test_set[i] for i in rndom]
# list of tagged words
test_run_base = [tup for sent in test_run for tup in sent]
# list of untagged words
test_tagged_words = [tup[0] for sent in test_run for tup in sent]
test_run
```

[34]:

```
[[('The', 'DT'),
  ('purchase', 'NN'),
  ('price', 'NN'),
  ('includes', 'VBZ'),
  ('two', 'CD'),
  ('ancillary', 'JJ'),
  ('companies', 'NNS'),
  ('.', '.')],
 [('He', 'PRP'),
  ('has', 'VBZ'),
  ('a', 'DT'),
  ('point', 'NN'),
  ('0', '-NONE-'),
  ('he', 'PRP'),
  ('wants', 'VBZ'),
  ('*-1', '-NONE-'),
  ('to', 'TO'),
  ('make', 'VB'),
  ('*T*-2', '-NONE-'),
  (',', ','),
  ('and', 'CC'),
  ('he', 'PRP'),
  ('makes', 'VBZ'),
  ('it', 'PRP'),
  (',', ','),
  ('with', 'IN'),
  ('a', 'DT'),
  ('great', 'JJ'),
  ('deal', 'NN'),
  ('of', 'IN'),
  ('force', 'NN'),
  ('.', '.')],
 [('The', 'DT'),
  ('new', 'JJ'),
  ('plant', 'NN'),
  (',', ','),
  ('located', 'VBN'),
  ('*', '-NONE-'),
  ('in', 'IN'),
  ('Chinchon', 'NNP'),
  ('about', 'IN'),
  ('60', 'CD'),
  ('miles', 'NNS'),
  ('from', 'IN'),
  ('Seoul', 'NNP'),
  (',', ','),
  ('will', 'MD'),
  ('help', 'VB'),
  ('*-2', '-NONE-'),
  ('meet', 'VB'),
  ('increasing', 'VBG'),
  ('and', 'CC'),
  ('diversifying', 'VBG'),
  ('demand', 'NN'),
  ('for', 'IN'),
  ('control', 'NN'),
  ('products', 'NNS'),
  ('in', 'IN'),
  ('South', 'NNP'),
  ('Korea', 'NNP'),
  (',', ','),
  ('the', 'DT'),
  ('company', 'NN'),
  ('said', 'VBD'),
  ('0', '-NONE-'),
  ('*T*-1', '-NONE-'),
  ('.', '.')],
 [('The', 'DT'),
  ('excision', 'NN'),
  ('of', 'IN'),
  ('unconstitutional', 'JJ'),
  ('conditions', 'NNS'),
  ('in', 'IN'),
  ('an', 'DT'),
  ('appropriations', 'NNS'),
  ('bill', 'NN'),
  ('would', 'MD'),
  ('be', 'VB'),
  ('a', 'DT'),
  ('power', 'NN'),
  ('of', 'IN'),
  ('far', 'RB'),
  ('more', 'RBR'),
  ('limited', 'VBN'),
  ('applicability', 'NN'),
  ('.', '.')],
 [('Pacific', 'NNP'),
  ('First', 'NNP'),
  ('Financial', 'NNP'),
  ('Corp.', 'NNP'),
  ('said', 'VBD'),
  ('0', '-NONE-'),
  ('shareholders', 'NNS'),
  ('approved', 'VBD'),
  ('its', 'PRP$'),
  ('acquisition', 'NN'),
  ('by', 'IN'),
  ('Royal', 'NNP'),
  ('Trustco', 'NNP'),
  ('Ltd.', 'NNP'),
  ('of', 'IN'),
  ('Toronto', 'NNP'),
  ('for', 'IN'),
  ('$', '$'),
  ('27', 'CD'),
  ('*U*', '-NONE-'),
  ('a', 'DT'),
  ('share', 'NN'),
  (',', ','),
  ('or', 'CC'),
  ('$', '$'),
  ('212', 'CD'),
  ('million', 'CD'),
  ('*U*', '-NONE-'),
  ('.', '.')]]
```

[35]:

```
#now, we will tag the test sentences using the Viterbi algorithm
```

[36]:

```
# tagging the test sentences
start = time.time()
tagged_seq = Viterbi(test_tagged_words)
end = time.time()
difference = end-start
print("Time taken in seconds: ", difference)
print(tagged_seq)
```

```
Time taken in seconds:  11.921807527542114
[('The', 'DT'), ('purchase', 'NN'), ('price', 'NN'), ('includes', 'VBZ'),
('two', 'CD'), ('ancillary', 'RP'), ('companies', 'NNS'), ('.', '.'), ('He',
'PRP'), ('has', 'VBZ'), ('a', 'DT'), ('point', 'NN'), ('0', '-NONE-'), ('he',
'PRP'), ('wants', 'VBZ'), ('*-1', '-NONE-'), ('to', 'TO'), ('make', 'VB'),
('*T*-2', '-NONE-'), (',', ','), ('and', 'CC'), ('he', 'PRP'), ('makes', 'VBZ'),
('it', 'PRP'), (',', ','), ('with', 'IN'), ('a', 'DT'), ('great', 'JJ'),
('deal', 'NN'), ('of', 'IN'), ('force', 'NN'), ('.', '.'), ('The', 'DT'),
('new', 'JJ'), ('plant', 'NN'), (',', ','), ('located', 'VBN'), ('*', '-NONE-'),
('in', 'IN'), ('Chinchon', 'RP'), ('about', 'IN'), ('60', 'CD'), ('miles',
'NNS'), ('from', 'IN'), ('Seoul', 'NNP'), (',', ','), ('will', 'MD'), ('help',
'VB'), ('*-2', '-NONE-'), ('meet', 'VBP'), ('increasing', 'VBG'), ('and', 'CC'),
('diversifying', 'RP'), ('demand', 'NN'), ('for', 'IN'), ('control', 'NN'),
('products', 'NNS'), ('in', 'IN'), ('South', 'NNP'), ('Korea', 'NNP'), (',',
','), ('the', 'DT'), ('company', 'NN'), ('said', 'VBD'), ('0', '-NONE-'),
('*T*-1', '-NONE-'), ('.', '.'), ('The', 'DT'), ('excision', 'NN'), ('of',
'IN'), ('unconstitutional', 'JJ'), ('conditions', 'NNS'), ('in', 'IN'), ('an',
'DT'), ('appropriations', 'NNS'), ('bill', 'NN'), ('would', 'MD'), ('be', 'VB'),
('a', 'DT'), ('power', 'NN'), ('of', 'IN'), ('far', 'RB'), ('more', 'JJR'),
('limited', 'JJ'), ('applicability', 'RP'), ('.', '.'), ('Pacific', 'NNP'),
('First', 'NNP'), ('Financial', 'NNP'), ('Corp.', 'NNP'), ('said', 'VBD'), ('0',
'-NONE-'), ('shareholders', 'NNS'), ('approved', 'VBD'), ('its', 'PRP$'),
('acquisition', 'NN'), ('by', 'IN'), ('Royal', 'RP'), ('Trustco', 'RP'),
('Ltd.', 'NNP'), ('of', 'IN'), ('Toronto', 'NNP'), ('for', 'IN'), ('$', '$'),
('27', 'CD'), ('*U*', '-NONE-'), ('a', 'DT'), ('share', 'NN'), (',', ','),
('or', 'CC'), ('$', '$'), ('212', 'RP'), ('million', 'CD'), ('*U*', '-NONE-'),
('.', '.')]
```

[37]:

```
#As we can see it has taken around 12 seconds and it has tagged all the words␣
 ↪in the test sentences. Now in order to check the accuracy we have to execute
#the below code
```

[38]:

```
# accuracy
check = [i for i, j in zip(tagged_seq, test_run_base) if i == j]
accuracy = len(check)/len(tagged_seq)
print(accuracy)
```

```
0.9130434782608695
```

[39]:

```
#Our POS tagger model, which is based on HMM, achieves a reasonably good␣
 ↪accuracy of 91.30% for POS tagging
```

[40]:

```
#Now let's test the model on a sample sentence.
```

[41]:

```
## Testing
sentence_test = 'Twitter is the best networking social site. Man is a social␣
 ↪animal. Data science is an emerging field. Data science jobs are high in␣
 ↪demand.'
words = word_tokenize(sentence_test)
start = time.time()
tagged_seq = Viterbi(words)
print(tagged_seq)
```

```
[('Twitter', 'RP'), ('is', 'RP'), ('the', 'DT'), ('best', 'JJS'), ('networking',
'RP'), ('social', 'JJ'), ('site', 'RP'), ('.', '.'), ('Man', 'NNP'), ('is',
'VBZ'), ('a', 'DT'), ('social', 'JJ'), ('animal', 'RP'), ('.', '.'), ('Data',
'NNP'), ('science', 'RP'), ('is', 'RP'), ('an', 'DT'), ('emerging', 'VBG'),
('field', 'NN'), ('.', '.'), ('Data', 'NNP'), ('science', 'RP'), ('jobs',
'NNS'), ('are', 'VBP'), ('high', 'JJ'), ('in', 'IN'), ('demand', 'NN'), ('.',
'.')]
```

[ ]:

```
#As we can see HMM model has done a reasonably good job of tagging a sample␣
 ↪sentence.
```