HMM

```
import nltk as nl
from sklearn.model_selection import train_test_split
import numpy as np
import pandas as pd
import random
import pprint, time
```

```
nl.download('treebank')
nl.download('universal_tagset')
```

```
[nltk_data] Downloading package treebank to /root/nltk_data...
[nltk_data]   Package treebank is already up-to-date!
[nltk_data] Downloading package universal_tagset to /root/nltk_data...
[nltk_data]   Package universal_tagset is already up-to-date!
True
```

```
nl_data = list(nl.corpus.treebank.tagged_sents(tagset='universal'))

tr_set,ts_set =train_test_split(nl_data,train_size=0.75,test_size=0.25)

tr_tg_word = [ tup for sent in tr_set for tup in sent ]

ts_tg_word = [ tup for sent in ts_set for tup in sent ]

tags = {tag for word,tag in tr_tg_word}


def word_given_tag(word, tag, tr_bag = tr_tg_word):

    tg_lis = [pair for pair in tr_bag if pair[1]==tag]

    ct_tag = len(tg_lis)

    w_given_tg_lis = [pair[0] for pair in tg_lis if pair[0]==word]

    ct_w_given_tg = len(w_given_tg_lis)

    return (ct_w_given_tg, ct_tag)


def t2_with_t1(t2, t1, tr_bag = tr_tg_word):

    tags = [pair[1] for pair in tr_bag]

    ct_t1 = len([t for t in tags if t==t1])

    ct_t2_t1 = 0

    for index in range(len(tags)-1):

        if tags[index]==t1 and tags[index+1] == t2:

            ct_t2_t1 += 1

    return (ct_t2_t1, ct_t1)
```

```
tgs_mtx = np.zeros((len(tags), len(tags)), dtype='float32')
for i, t1 in enumerate(list(tags)):
    for j, t2 in enumerate(list(tags)):
        tgs_mtx[i, j] = t2_with_t1(t2, t1)[0]/t2_with_t1(t2, t1)[1]
```

```
tags_df = pd.DataFrame(tgs_mtx, columns = list(tags), index=list(tags))
display(tags_df)
```

	PRT	ADV	DET	NUM	PRON	ADP	CONJ	.	X	NOUN	VERB	ADJ
PRT	0.001254	0.007525	0.099916	0.057692	0.017140	0.019649	0.002508	0.043478	0.013796	0.239967	0.407609	0.089465
ADV	0.013831	0.084241	0.067896	0.027242	0.014250	0.120285	0.008382	0.133697	0.024728	0.031433	0.340319	0.133697
DET	0.000305	0.013134	0.005345	0.023060	0.003513	0.009163	0.000458	0.017104	0.047190	0.634698	0.041845	0.204184
NUM	0.030314	0.002246	0.002994	0.181512	0.001497	0.035928	0.013473	0.122006	0.209581	0.349177	0.018713	0.032560
PRON	0.012364	0.035608	0.008408	0.006924	0.007418	0.023244	0.005440	0.041048	0.093966	0.200791	0.491592	0.073195
ADP	0.001209	0.012497	0.329616	0.064499	0.065843	0.018140	0.000806	0.037759	0.032787	0.320747	0.008600	0.107498
CONJ	0.005396	0.055755	0.125300	0.038969	0.055755	0.057554	0.000000	0.035971	0.009592	0.335731	0.157674	0.122302
.	0.002511	0.051935	0.172469	0.081726	0.067458	0.091085	0.058213	0.093254	0.027851	0.222349	0.086520	0.044515
X	0.184450	0.026321	0.054667	0.001417	0.055477	0.142539	0.010326	0.161774	0.073294	0.063373	0.207532	0.018830
VERB	0.031783	0.082990	0.132235	0.023053	0.034432	0.090151	0.005592	0.033843	0.217775	0.112713	0.169610	0.065823
ADJ	0.011123	0.004737	0.004737	0.020803	0.000618	0.081771	0.015654	0.063852	0.022245	0.697013	0.011535	0.065911