nlp-group-161-problem-statement-18

March 10, 2024

Name	BITS Id	Contribution percentage
RAVURU JEEVAN KUMAR REDDY	2022da04668	100% Contributed
PREM RANJAN	2022da04012	100% Contributed
CHANDUPATLA ANIRUDH REDDY	2022da04387	100% Contributed

0.1 Performing EDA

```
[1]: import numpy as np
     import pandas as pd

     # Reading data
     df = pd.read_csv('./archive/brown.csv')
```

```
[2]: df.head()
```

```
[2]:    filename  para_id  sent_id  \
     0     cd05        0        0
     1     cd05        0        1
     2     cd05        0        2
     3     cd05        0        3
     4     cd05        0        4


                                                    raw_text  \
     0  Furthermore/rb ,/, as/cs an/at encouragement/n…
     1  The/at Unitarian/jj clergy/nns were/bed an/at …
     2  Ezra/np Stiles/np Gannett/np ,/, an/at honorab…
     3  Even/rb so/rb ,/, Gannett/np judiciously/rb ar…
     4  We/ppss today/nr are/ber not/* entitled/vbn to…


                                              tokenized_text  \
     0  Furthermore , as an encouragement to revisioni…
     1  The Unitarian clergy were an exclusive club of…
     2  Ezra Stiles Gannett , an honorable representat…
     3  Even so , Gannett judiciously argued , the Ass…
     4  We today are not entitled to excoriate honest …
```

```
                            tokenized_pos      label
    0  rb , cs at nn in nn nn , pps rb bez jj to vb c…  religion
    1  at jj nns bed at jj nn in vbn nns -- cs at nn …  religion
    2  np np np , at jj nn in at nn , vbd ppl rb in a…  religion
    3  rb rb , np rb vbd , at nn-tl md rb vb cs np ``…  religion
    4  ppss nr ber * vbn to vb jj nns wps vbd np to b…  religion
```

[11]:
```python
unique_tags = set()
for pos in df.tokenized_pos:
    for tag in set(pos.split()):
        unique_tags.add(tag)

print('number of unique POS tags present in data set:', len(unique_tags))
```

number of unique POS tags present in data set: 523

[14]:
```python
' | '.join(sorted(unique_tags)) # list of all unique tags separated by |
```

[14]: "' | '' | (| (-hl |) |)-hl | * | *-hl | *-nc | *-tl | , | ,-hl | ,-nc | ,-tl
| -- | ---hl | . | .-hl | .-nc | .-tl | 1 | 16 | 16'' | 16-inch | 2 | 2% | 2'' |
2-foot | 2-inch | 2-story | 20th | 29 | 3 | 3% | 32'' | 4 | 4'' | 4-inch | 50th
| 64'' | 7 | 7074 | 8 | 8'' | 8-inch | 8-inch-thick | 9 | : | :-hl | :-tl |
Output | `` | abl | abn | abn-hl | abn-nc | abn-tl | abx | active | ap | ap$ |
ap+ap-nc | ap-hl | ap-nc | ap-tl | at | at-hl | at-nc | at-tl | at-tl-hl | be |
be-hl | be-tl | bed | bed* | bed-nc | bedz | bedz* | bedz-hl | bedz-nc | beg |
bem | bem* | bem-nc | ben | ben-tl | ber | ber* | ber*-nc | ber-hl | ber-nc |
ber-tl | bez | bez* | bez-hl | bez-nc | bez-tl | c | cc | cc-hl | cc-nc | cc-tl
| cc-tl-hl | cd | cd$ | cd-hl | cd-nc | cd-tl | cd-tl-hl | cell | chamber |
coating | cs | cs-hl | cs-nc | cs-tl | cwt | day | destination | do | do* |
do*-hl | do+ppss | do-hl | do-nc | do-tl | dod | dod* | dod*-tl | dod-nc | doz |
doz* | doz*-tl | doz-hl | doz-tl | dt | dt$ | dt+bez | dt+bez-nc | dt+md | dt-hl
| dt-nc | dt-tl | dti | dti-hl | dti-tl | dts | dts+bez | dts-hl | dtx | ex |
ex+bez | ex+hvd | ex+hvz | ex+md | ex-hl | ex-nc | ft | fw-* | fw-*-tl | fw-at |
fw-at+nn-tl | fw-at+np-tl | fw-at-hl | fw-at-tl | fw-be | fw-ber | fw-bez | fw-
cc | fw-cc-tl | fw-cd | fw-cd-tl | fw-cs | fw-dt | fw-dt+bez | fw-dts | fw-hv |
fw-in | fw-in+at | fw-in+at-t | fw-in+at-tl | fw-in+nn | fw-in+nn-tl | fw-in+np-
tl | fw-in-tl | fw-jj | fw-jj-nc | fw-jj-tl | fw-jjr | fw-jjt | fw-nn | fw-nn$ |
fw-nn$-tl | fw-nn-nc | fw-nn-tl | fw-nn-tl-nc | fw-nns | fw-nns-nc | fw-nns-tl |
fw-np | fw-np-tl | fw-nps | fw-nps-tl | fw-nr | fw-nr-tl | fw-od-nc | fw-od-tl |
fw-pn | fw-pp$ | fw-pp$-nc | fw-pp$-tl | fw-ppl | fw-ppl+vbz | fw-ppo | fw-
ppo+in | fw-pps | fw-ppss | fw-ppss+hv | fw-ql | fw-rb | fw-rb+cc | fw-rb-tl |
fw-to+vb | fw-uh | fw-uh-nc | fw-uh-tl | fw-vb | fw-vb-nc | fw-vb-tl | fw-vbd |
fw-vbd-tl | fw-vbg | fw-vbg-tl | fw-vbn | fw-vbz | fw-wdt | fw-wpo | fw-wps | g
| hp | hr. | hv | hv* | hv+to | hv-hl | hv-nc | hv-tl | hvd | hvd* | hvd-hl |
hvg | hvg-hl | hvn | hvz | hvz* | hvz-nc | hvz-tl | in | in+in | in+ppo | in-hl
| in-nc | in-tl | in-tl-hl | jj | jj$-tl | jj+jj-nc | jj-hl | jj-nc | jj-tl |
jj-tl-hl | jj-tl-nc | jjr | jjr+cs | jjr-hl | jjr-nc | jjr-tl | jjs | jjs-hl |

2
```

```
jjs-tl | jjt | jjt-hl | jjt-nc | jjt-tl | l | lBOD | m. | mbf | md | md* |
md*-hl | md+hv | md+ppss | md+to | md-hl | md-nc | md-tl | nil | nn | nn$ |
nn$-hl | nn$-tl | nn+bez | nn+bez-tl | nn+hvd-tl | nn+hvz | nn+hvz-tl | nn+in |
nn+md | nn+nn-nc | nn-hl | nn-nc | nn-tl | nn-tl-hl | nn-tl-nc | nns | nns$ |
nns$-hl | nns$-nc | nns$-tl | nns$-tl-hl | nns+md | nns-hl | nns-nc | nns-tl |
nns-tl-hl | nns-tl-nc | np | np$ | np$-hl | np$-tl | np+bez | np+bez-nc | np+hvz
| np+hvz-nc | np+md | np-hl | np-nc | np-tl | np-tl-hl | nps | nps$ | nps$-hl |
nps$-tl | nps-hl | nps-nc | nps-tl | nr | nr$ | nr$-tl | nr+md | nr-hl | nr-nc |
nr-tl | nr-tl-hl | nrs | nrs-tl | od | od-hl | od-nc | od-tl | or | output | pn
| pn$ | pn+bez | pn+hvd | pn+hvz | pn+md | pn-hl | pn-nc | pn-tl | pp$ | pp$$ |
pp$-hl | pp$-nc | pp$-tl | ppl | ppl-hl | ppl-nc | ppl-tl | ppls | ppo | ppo-hl
| ppo-nc | ppo-tl | pps | pps+bez | pps+bez-hl | pps+bez-nc | pps+hvd | pps+hvz
| pps+md | pps-hl | pps-nc | pps-tl | ppss | ppss+bem | ppss+ber | ppss+ber-n |
ppss+ber-nc | ppss+ber-tl | ppss+bez | ppss+bez* | ppss+hv | ppss+hv-tl |
ppss+hvd | ppss+md | ppss+md-nc | ppss+vb | ppss-hl | ppss-nc | ppss-tl | ql |
ql-hl | ql-nc | ql-tl | qlp | r | rb | rb$ | rb+bez | rb+bez-hl | rb+bez-nc |
rb+cs | rb-hl | rb-nc | rb-tl | rbr | rbr+cs | rbr-nc | rbt | rn | rp | rp+in |
rp-hl | rp-nc | rp-tl | sec | sq. | to | to+vb | to-hl | to-nc | to-tl | uh |
uh-hl | uh-nc | uh-tl | under | vb | vb+at | vb+in | vb+jj-nc | vb+ppo | vb+rp |
vb+to | vb+vb-nc | vb-hl | vb-nc | vb-tl | vbd | vbd-hl | vbd-nc | vbd-tl | vbg
| vbg+to | vbg-hl | vbg-nc | vbg-tl | vbn | vbn+to | vbn-hl | vbn-nc | vbn-tl |
vbn-tl-hl | vbn-tl-nc | vbz | vbz-hl | vbz-nc | vbz-tl | watt | wdt | wdt+ber |
wdt+ber+pp | wdt+bez | wdt+bez-hl | wdt+bez-nc | wdt+bez-tl | wdt+do+pps |
wdt+dod | wdt+hvz | wdt-hl | wdt-nc | wp$ | wpo | wpo-nc | wpo-tl | wps |
wps+bez | wps+bez-nc | wps+bez-tl | wps+hvd | wps+hvz | wps+md | wps-hl | wps-nc
| wps-tl | wql | wql-tl | wrb | wrb+ber | wrb+bez | wrb+bez-tl | wrb+do |
wrb+dod | wrb+dod* | wrb+doz | wrb+in | wrb+md | wrb-hl | wrb-nc | wrb-tl"
```

```python
[3]: set(df.label) #what all labels are present
```

```
[3]: {'adventure',
 'belles_lettres',
 'editorial',
 'fiction',
 'government',
 'hobbies',
 'humor',
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction'}
```

```python
[4]: from collections import Counter
```

```
[5]: Counter(df.label)
```

```
[5]: Counter({'learned': 7734,
              'belles_lettres': 7209,
              'lore': 4881,
              'adventure': 4637,
              'news': 4623,
              'romance': 4431,
              'fiction': 4249,
              'hobbies': 4193,
              'mystery': 3886,
              'government': 3032,
              'editorial': 2997,
              'reviews': 1751,
              'religion': 1716,
              'humor': 1053,
              'science_fiction': 948})
```

*We observe that top 3 labels that occured in most sentences are 'learned',
'belles_lettres' and 'lore'*

```
[7]: file_para_sent_df = df[['filename','para_id','sent_id']]
     file_df = file_para_sent_df.groupby(['filename'])
     file_para_df = file_para_sent_df.groupby(['filename','para_id'])
```

```
[87]: len(file_df.groups.keys())
```

```
[87]: 500
```

```
[104]: file_para_sent_count = []
       for file in file_df.groups.keys():
           n_para = max(file_df.get_group(file).para_id)+1
           file_para_sent_count.append([file, n_para,sum([max(file_para_df.
        ↪get_group((file, p)).sent_id)+1 for p in range(n_para)])])

       from prettytable import PrettyTable
       table = PrettyTable()
       table.field_names = ['file1','n_para1','n_sent1'] + \
       ['file2','n_para2','n_sent2']
       for ls in [file_para_sent_count[i:i+2] for i in␣
        ↪range(0,len(file_para_sent_count),2)]:
           table.add_row(sum(ls, []))
       table
```

```
[104]: +-------+---------+---------+-------+---------+---------+
       | file1 | n_para1 | n_sent1 | file2 | n_para2 | n_sent2 |
       +-------+---------+---------+-------+---------+---------+
```

ca01	67	98	ca02	60	98
ca03	52	112	ca04	50	88
ca05	53	84	ca06	49	99
ca07	55	120	ca08	62	101
ca09	68	113	ca10	75	106
ca11	60	100	ca12	53	104
ca13	61	108	ca14	58	133
ca15	61	110	ca16	27	116
ca17	71	104	ca18	49	105
ca19	70	113	ca20	51	97
ca21	73	112	ca22	69	106
ca23	59	113	ca24	59	100
ca25	68	107	ca26	62	107
ca27	54	115	ca28	36	93
ca29	56	119	ca30	59	118
ca31	65	113	ca32	57	124
ca33	49	93	ca34	30	94
ca35	19	74	ca36	25	86
ca37	21	95	ca38	25	95
ca39	30	126	ca40	21	96
ca41	17	103	ca42	30	104
ca43	75	129	ca44	23	92
cb01	43	103	cb02	46	95
cb03	38	96	cb04	49	117
cb05	43	105	cb06	31	99
cb07	26	92	cb08	40	128
cb09	39	125	cb10	51	136
cb11	57	110	cb12	50	105
cb13	42	143	cb14	45	119
cb15	39	102	cb16	48	120
cb17	40	126	cb18	40	115
cb19	41	138	cb20	15	110
cb21	20	78	cb22	26	94
cb23	22	133	cb24	22	109
cb25	21	79	cb26	32	112
cb27	37	108	cc01	44	116
cc02	37	99	cc03	38	114
cc04	51	102	cc05	32	105
cc06	29	83	cc07	47	120
cc08	42	91	cc09	35	90
cc10	36	110	cc11	39	109
cc12	42	92	cc13	32	94
cc14	40	102	cc15	34	127
cc16	24	96	cc17	27	101
cd01	13	84	cd02	19	67
cd03	20	84	cd04	14	96
cd05	10	71	cd06	37	88

cd07	21	122	cd08	17	72
cd09	33	170	cd10	16	96
cd11	7	67	cd12	24	81
cd13	15	104	cd14	19	93
cd15	17	99	cd16	46	171
cd17	41	151	ce01	43	104
ce02	26	124	ce03	40	99
ce04	17	96	ce05	46	112
ce06	33	94	ce07	27	132
ce08	39	130	ce09	21	142
ce10	47	131	ce11	24	113
ce12	44	121	ce13	22	75
ce14	38	168	ce15	20	199
ce16	24	130	ce17	20	110
ce18	28	83	ce19	35	108
ce20	35	123	ce21	27	112
ce22	23	107	ce23	22	89
ce24	19	124	ce25	25	93
ce26	13	90	ce27	25	187
ce28	46	125	ce29	37	89
ce30	72	152	ce31	31	103
ce32	29	119	ce33	37	113
ce34	31	109	ce35	36	112
ce36	17	75	cf01	14	73
cf02	30	124	cf03	28	75
cf04	41	124	cf05	24	89
cf06	36	166	cf07	41	91
cf08	44	113	cf09	75	132
cf10	30	102	cf11	45	117
cf12	24	96	cf13	42	138
cf14	10	70	cf15	29	82
cf16	32	123	cf17	15	74
cf18	7	101	cf19	16	88
cf20	15	90	cf21	23	107
cf22	29	116	cf23	25	101
cf24	41	148	cf25	22	109
cf26	15	122	cf27	12	85
cf28	16	96	cf29	20	98
cf30	26	95	cf31	26	105
cf32	23	112	cf33	10	90
cf34	50	114	cf35	15	99
cf36	35	127	cf37	18	93
cf38	17	60	cf39	19	76
cf40	18	86	cf41	25	99
cf42	13	106	cf43	12	97
cf44	13	96	cf45	21	100
cf46	32	112	cf47	22	107

cf48	7	57	cg01	16	100
cg02	13	90	cg03	27	121
cg04	24	147	cg05	17	96
cg06	32	108	cg07	29	86
cg08	17	88	cg09	14	106
cg10	17	71	cg11	10	92
cg12	23	145	cg13	12	104
cg14	9	97	cg15	24	92
cg16	11	73	cg17	22	112
cg18	15	107	cg19	12	117
cg20	16	88	cg21	4	89
cg22	10	95	cg23	32	126
cg24	10	73	cg25	15	105
cg26	16	85	cg27	23	102
cg28	12	64	cg29	16	79
cg30	19	121	cg31	12	71
cg32	13	101	cg33	9	74
cg34	45	123	cg35	43	100
cg36	52	146	cg37	30	113
cg38	10	75	cg39	22	100
cg40	29	119	cg41	33	109
cg42	22	100	cg43	11	79
cg44	12	72	cg45	27	102
cg46	15	85	cg47	55	135
cg48	18	103	cg49	20	84
cg50	18	102	cg51	20	107
cg52	20	86	cg53	28	102
cg54	30	118	cg55	28	111
cg56	15	87	cg57	14	86
cg58	15	75	cg59	10	50
cg60	24	119	cg61	16	82
cg62	10	82	cg63	13	107
cg64	8	71	cg65	15	91
cg66	7	71	cg67	14	83
cg68	6	80	cg69	9	56
cg70	16	89	cg71	12	74
cg72	19	83	cg73	16	100
cg74	9	99	cg75	18	128
ch01	9	149	ch02	19	116
ch03	31	83	ch04	34	123
ch05	34	83	ch06	22	107
ch07	40	133	ch08	38	81
ch09	14	104	ch10	22	107
ch11	20	101	ch12	4	79
ch13	45	108	ch14	33	112
ch15	57	156	ch16	20	79
ch17	13	106	ch18	20	85

ch19	41	92	ch20	34	107
ch21	31	90	ch22	31	83
ch23	12	70	ch24	39	127
ch25	23	81	ch26	27	70
ch27	45	109	ch28	43	119
ch29	19	73	ch30	31	99
cj01	13	71	cj02	13	111
cj03	15	84	cj04	19	113
cj05	19	95	cj06	16	81
cj07	15	102	cj08	26	113
cj09	20	97	cj10	12	97
cj11	22	102	cj12	10	102
cj13	14	76	cj14	7	108
cj15	21	115	cj16	16	93
cj17	14	86	cj18	26	126
cj19	23	127	cj20	24	108
cj21	19	78	cj22	15	86
cj23	18	110	cj24	10	65
cj25	29	120	cj26	21	100
cj27	13	95	cj28	16	88
cj29	14	96	cj30	22	104
cj31	14	61	cj32	16	99
cj33	7	97	cj34	15	103
cj35	22	104	cj36	20	82
cj37	13	83	cj38	26	95
cj39	24	95	cj40	11	70
cj41	15	84	cj42	10	91
cj43	12	95	cj44	17	78
cj45	36	108	cj46	25	90
cj47	23	115	cj48	23	114
cj49	20	81	cj50	15	70
cj51	15	100	cj52	10	96
cj53	10	71	cj54	15	97
cj55	16	86	cj56	30	100
cj57	12	87	cj58	21	82
cj59	13	72	cj60	15	79
cj61	14	89	cj62	12	85
cj63	14	88	cj64	25	112
cj65	15	82	cj66	19	113
cj67	9	78	cj68	11	57
cj69	29	102	cj70	20	118
cj71	18	126	cj72	12	112
cj73	5	185	cj74	26	108
cj75	22	92	cj76	14	98
cj77	32	122	cj78	30	120
cj79	16	97	cj80	32	116
ck01	45	172	ck02	21	141

ck03	22	137	ck04	26	129
ck05	48	174	ck06	47	205
ck07	18	240	ck08	54	142
ck09	25	93	ck10	42	156
ck11	42	152	ck12	21	139
ck13	42	139	ck14	38	140
ck15	40	127	ck16	23	127
ck17	68	187	ck18	35	151
ck19	27	118	ck20	50	138
ck21	43	207	ck22	22	134
ck23	23	110	ck24	80	220
ck25	5	67	ck26	47	136
ck27	20	96	ck28	54	187
ck29	15	85	cl01	55	150
cl02	72	147	cl03	64	189
cl04	46	168	cl05	53	151
cl06	29	144	cl07	55	144
cl08	29	143	cl09	51	162
cl10	96	180	cl11	41	147
cl12	52	196	cl13	51	177
cl14	31	197	cl15	46	153
cl16	38	183	cl17	50	122
cl18	47	135	cl19	42	146
cl20	53	175	cl21	72	209
cl22	13	140	cl23	40	160
cl24	38	168	cm01	57	174
cm02	41	152	cm03	52	144
cm04	81	197	cm05	44	120
cm06	60	161	cn01	58	173
cn02	37	172	cn03	61	212
cn04	42	178	cn05	45	196
cn06	5	127	cn07	81	201
cn08	32	114	cn09	55	181
cn10	59	199	cn11	39	127
cn12	70	186	cn13	47	122
cn14	45	168	cn15	40	176
cn16	58	129	cn17	58	188
cn18	37	126	cn19	22	146
cn20	53	159	cn21	47	143
cn22	42	142	cn23	35	129
cn24	54	179	cn25	44	195
cn26	49	129	cn27	51	127
cn28	47	121	cn29	74	192
cp01	33	131	cp02	40	98
cp03	44	146	cp04	47	133
cp05	29	131	cp06	57	182
cp07	50	144	cp08	22	175

```
| cp09  |    9    |    69    | cp10  |   15    |   123    |
| cp11  |   40    |   119    | cp12  |   41    |   170    |
| cp13  |   34    |   112    | cp14  |   53    |   231    |
| cp15  |   46    |   179    | cp16  |   64    |   207    |
| cp17  |   37    |   116    | cp18  |   48    |   185    |
| cp19  |   48    |   174    | cp20  |   61    |   182    |
| cp21  |   21    |   120    | cp22  |   66    |   176    |
| cp23  |   46    |   172    | cp24  |   88    |   235    |
| cp25  |   29    |    99    | cp26  |   34    |   109    |
| cp27  |   67    |   207    | cp28  |   48    |   156    |
| cp29  |   36    |   150    | cr01  |   32    |   113    |
| cr02  |   19    |   113    | cr03  |   12    |    92    |
| cr04  |   28    |   118    | cr05  |   45    |   141    |
| cr06  |   44    |   123    | cr07  |   52    |   184    |
| cr08  |   12    |    70    | cr09  |   10    |    99    |
+-------+---------+---------+-------+---------+---------+
```

0.1.1 *displayed number of paragraph, sentences are present per file*

0.2 *Preprocessing and Train Test split*

```python
[24]: # convert lowercase and delete whitespace
      def clean(text):
          text = text.lower()
          text = " ".join(text.split())
          return text

      # removing words with more than one '/' in them as they are creating an
       ↪ambiguity and not providing any insight.
      list_of_word_tags = []
      for i in range(len(df.raw_text)):
          sentence = []
          for word_tag in df.raw_text.loc[i].split():
              if(word_tag.count('/')>1):
                  pass
              else:
                  word , tag = word_tag.split('/')
                  word = clean(word)
                  sentence.append([word, tag])
          list_of_word_tags.append(sentence)
      len(list_of_word_tags)
```

```
[24]: 57340
```

```python
[42]: train_test_dictionary = {i:list_of_word_tags[i] for i in
       ↪range(len(list_of_word_tags))}
```

```
[44]: import random
      total_size = len(list_of_word_tags)
      test_indices = random.sample(range(total_size), int(total_size*(1-0.9)))
```

```
[45]: test = [train_test_dictionary[test_key] for test_key in test_indices]
```

```
[47]: train = [train_test_dictionary[i] for i in range(len(list_of_word_tags)) if i␣
       ↪not in test_indices]
```

```
[49]: len(train),len(test)
```

```
[49]: (51607, 5733)
```

```
[54]: total_size == sum([51607, 5733])
```

```
[54]: True
```

0.3 About Viterbi Algorithm

The Viterbi Algorithm is a powerful technique used in the fields of digital communication, signal processing, and natural language processing. Named after Andrew Viterbi, who introduced it in 1967, this algorithm plays a crucial role in decoding and error correction. Let's dive into the intricacies of the Viterbi Algorithm:

1. Background and Theoretical Foundations:

- The Viterbi Algorithm is primarily used to decode the most likely sequence of hidden states in a Hidden Markov Model (HMM).
- Hidden Markov Models are statistical models that represent systems transitioning between various states, where each state produces an observable output.
- The central problem the algorithm addresses is determining the sequence of hidden states (or path) that most likely led to a given sequence of observed events.
- It optimizes efficiency by using dynamic programming, breaking down complex problems into simpler subproblems and avoiding redundant calculations.

2. Key Components and Workflow:

- Initialization: Set up the probabilities of the initial state and consider the first observation.
- Recursion: For each new observation, calculate the most probable path to each state, considering the previous state's most likely path and the new observation.
- Termination: After processing all observations, identify the final state with the highest probability.
- Path Backtracking: Trace back through the states from the final state to determine the most probable sequence of hidden states.

3. Applications:

- Digital Communications: The Viterbi Algorithm is widely used in decoding convolutional codes for error correction. These codes ensure data integrity in communication mediums like satellites, mobile networks, and deep-space communication.

11

- Natural Language Processing: In NLP, the algorithm can be applied to tasks such as part-of-speech tagging, speech recognition, and machine translation.

4. Impact:

- The Viterbi Algorithm has revolutionized technology and information theory by providing a clear path through uncertainty. It transforms chaotic sequences into meaningful state sequences, making it a cornerstone in various fields.

In summary, the Viterbi Algorithm is a beacon of clarity, unraveling hidden patterns and enabling efficient decoding in complex systems

Pseudocode as per https://en.wikipedia.org/wiki/Viterbi_algorithm

```
function Viterbi(states, init, trans, emit, obs) is
    input states: S hidden states
    input init: initial probabilities of each state
    input trans: S × S transition matrix
    input emit: S × O emission matrix
    input obs: sequence of T observations

    prob ← T × S matrix of zeroes
    prev ← empty T × S matrix
    for each state s in states do
        prob[0][s] = init[s] * emit[s][obs[0]]
    end
    for t = 1 to T − 1 do
        for each state s in states do
            for each state r in states do
                new_prob ← prob[t − 1][r] * trans[r][s] * emit[s][obs[t]]
                if new_prob > prob[t][s] then
                    prob[t][s] ← new_prob
                    prev[t][s] ← i
                end
            end
        end
    end
    path ← empty array of length T
    path[T − 1] ← the state s with maximum prob[T − 1][s]
    for t = T − 1 to 1 do
        path[t − 1] ← prev[t][path[t]]
    end
    return path
end
```

0.4 Train a Hidden Markov Model (HMM) for part-of-speech (POS) tagging using viterbi algorithm

```
[55]: from joblib import Parallel,delayed
      from collections import defaultdict, Counter
      import random
      random.seed(42)

      class POSTagger():
          def __init__(self, corpus):
              """
              Args:
                  corpus: list of sentences comprising the training corpus. Each␣
          ↪sentence is a list
                            of (word, POS tag) tuples.
              """
              # Create a Python Counter object of (tag, word)-frequecy key-value pairs
              try:
                  self.tag_word_cnt = Counter([(tag, word) for sent in corpus for␣
          ↪word, tag in sent])
              except:
                  pass
              # Create a tag-only corpus. Adding the bos token for computing the␣
          ↪initial probability.
              self.tag_corpus = [["<bos>"]+[word_tag[1] for word_tag in sent] for␣
          ↪sent in corpus]
              # Count the unigrams and bigrams for pos tags
              self.tag_unigram_cnt = self._count_ngrams(self.tag_corpus, 1)
              self.tag_bigram_cnt = self._count_ngrams(self.tag_corpus, 2)
              self.all_tags = sorted(list(set(self.tag_unigram_cnt.keys())))

              # Compute the transition and emission probability
              self.tran_prob = self.compute_tran_prob()
              self.emis_prob = self.compute_emis_prob()

          def _get_ngrams(self, sent, n):
              """
              Given a text sentence and the argument n, we convert it to a list of␣
          ↪n-grams.
              Args:
                  sent (list of str): input text sentence.
                  n (int): the order of n-grams to return (i.e. 1 for unigram, 2 for␣
          ↪bigram, etc.).
              Returns:
                  ngrams: a list of n-gram (tuples if n != 1, otherwise strings)
              """
              ngrams = []
```

```python
        for i in range(len(sent) - n + 1):
            ngram = tuple(sent[i:i+n]) if n != 1 else sent[i]
            ngrams.append(ngram)
        return ngrams

    def _count_ngrams(self, corpus, n):
        """
        Given a training corpus, count the frequency of each n-gram.
        Args:
            corpus (list of str): list of sentences comprising the training␣
↪corpus with <bos> inserted.
            n (int): the order of n-grams to count (i.e. 1 for unigram, 2 for␣
↪bigram, etc.).
        Returns:
            ngram_freq (Counter): Python Counter object of (ngram (tuple or␣
↪str), frequency (int)) key-value pairs.
        """
        corpus_ngrams = []
        for sent in corpus:
            sent_ngrams = self._get_ngrams(sent, n)
            corpus_ngrams += sent_ngrams
        ngram_cnt = Counter(corpus_ngrams)
        return ngram_cnt

    def compute_tran_prob(self):
        """
        Compute the transition probability.

        P(tagB | tagA) = P(tagA | tagB) * P(tagB) / P(tagA) = P(tagB, tagA) /␣
↪P(tagA)

        Returns:
            tran_prob: a dictionary that maps each (tagA, tagB) tuple to its␣
↪transition probability P(tagB|tagA).
        """
        tran_prob = defaultdict(lambda: 0) # if a tuple is unseen during␣
↪training, we set its transition praobility to 0
        for tag_bigram in self.tag_bigram_cnt:
            tagA, _ = tag_bigram
            tran_prob[tag_bigram] = self.tag_bigram_cnt[tag_bigram] / self.
↪tag_unigram_cnt[tagA]
        return tran_prob

    def compute_emis_prob(self):
        """
        Compute the emission probability.
```

```python
        P(word|tag) = P(tag|word)P(word)/P(tag) = P(tag & word) / P(tag)

        Returns:
            emis_prob: a dictionary that maps each (tagA, wordA) tuple to its
↪emission probability P(wordA|tagA).
        """
        emis_prob = defaultdict(lambda: 0) # if a tuple is unseen during
↪training, we set its emission probability to 0
        for tag, word in self.tag_word_cnt:
            emis_prob[tag, word] = self.tag_word_cnt[tag, word] / self.
↪tag_unigram_cnt[tag]
        return emis_prob

    def init_prob(self, tag):
        """
        Compute the initial probability for a given tag.

        P(tag|<bos>) = P(<bos>, tag) / P(<bos>)

        Returns:
            tag_init_prob (float): the initial probability for {tag}
        """
        return self.tag_bigram_cnt['<bos>', tag] / self.tag_unigram_cnt['<bos>']

    def viterbi(self, sent):
        """
        Given the computed initial/transition/emission probability, make
↪predictions for a given
        sentence using the Viterbi algorithm.

        Implementation based off of https://en.wikipedia.org/wiki/
↪Viterbi_algorithm

        Args:
            sent: a list of words (strings)
        Returns:
            pos_tag: a list of corresponding pos tags (strings)

        Example 1:
            Input: ['Eddie', 'shouted', '.']
            Output: ['NP', 'VBD', '.']
        Example 2:
            Input: ['Mike', 'caught', 'the', 'ball', 'just', 'as', 'the',
↪'catcher', 'slid', 'into', 'the', 'bag', '.']
```

```python
            Output: ['NP', 'VBD', 'AT', 'NN', 'RB', 'CS', 'AT', 'NN', 'VBD',␣
↪'IN', 'AT', 'NN', '.']
        """
        trellis = {}
        pointers = {}

        # Find tag probabilities for first word
        for tag in self.all_tags:
            trellis[tag, 0] = self.init_prob(tag) * self.emis_prob[tag, sent[0]]

        # Find tag probabilities for rest of the words
        for step in range(1, len(sent)):
            word = sent[step]
            for tag in self.all_tags:
                max_tag, max_prob = 'nn', trellis['nn', step - 1] * self.
↪tran_prob['nn', tag] * self.emis_prob[tag, word]
                for prev_tag in self.all_tags:
                    prob = trellis[prev_tag, step - 1] * self.
↪tran_prob[prev_tag, tag] * self.emis_prob[tag, word]
                    if prob > max_prob:
                        max_tag = prev_tag
                        max_prob = prob
                trellis[tag, step] = max_prob
                pointers[tag, step] = max_tag

        # Find most probable final tag
        last_step = len(sent) - 1
        max_tag, max_prob = 'nn', 0.0
        for tag, step in trellis:
            if step == last_step and trellis[tag, step] > max_prob:
                max_tag = tag
                max_prob = trellis[tag, step]

        # Backtrack to find most probable tags at each step
        prev_tag = max_tag
        pos_tag = [prev_tag]
        for step in range(len(sent)-1, 0, -1):
            prev_tag = pointers[prev_tag, step]
            pos_tag.append(prev_tag)
        pos_tag = pos_tag[::-1]

        return pos_tag

    def test_acc(self, corpus):
        """
        Given a training corpus, we compute the model prediction accuracy.
        Args:
```

```
            corpus: list of sentences comprising with each sentence being a list
                    of (word, POS tag) tuples
        Returns:
            acc: model prediction accuracy (float)
        """
        list_cor_tot = Parallel(n_jobs=-1)(delayed(self.predict_tags)(data) for
    ↪data in corpus)
        list_cor,list_tot = list(zip(*list_cor_tot))
        cor = sum(list_cor)
        tot = sum(list_tot)
        acc = cor/tot
        return acc

    def predict_tags(self,data):
        tot = cor = 0
        sent, gold_tags = zip(*data)
        pred_tags = self.viterbi(sent)
        for gold_tag, pred_tag in zip(gold_tags, pred_tags):
            cor += (gold_tag==pred_tag)
            tot += 1
        return cor,tot
```

```
[62]: # training POS tagger i.e building trainsition and emission probability tables
      pos_tagger = POSTagger(train)
```

```
[57]: # number of unique tag to tag transitions
      len(pos_tagger.tran_prob)
```

```
[57]: 7987
```

```
[58]: # probabilities of tag to tag transitions
      pos_tagger.tran_prob
```

```
[58]: defaultdict(<function __main__.POSTagger.compute_tran_prob.<locals>.<lambda>()>,
                  {('<bos>', 'rb'): 0.06572751758482376,
                   ('rb', ','): 0.13050749536290934,
                   (',', 'cs'): 0.05141202027516292,
                   ('cs', 'at'): 0.25392998844859627,
                   ('at', 'nn'): 0.4938574102433214,
                   ('nn', 'in'): 0.2768281655631071,
                   ('in', 'nn'): 0.1427004102705942,
                   ('nn', 'nn'): 0.07648834699413704,
                   ('nn', ','): 0.1202636875419304,
                   (',', 'pps'): 0.046648119211860205,
                   ('pps', 'rb'): 0.03800475059382423,
                   ('rb', 'bez'): 0.004834737130173017,
                   ('bez', 'jj'): 0.12681958535509483,
```

```
('jj', 'to'): 0.019779952375158606,
('to', 'vb'): 0.8215060195917147,
('vb', 'cs'): 0.041050577222205683,
('cs', 'dti'): 0.009944251921048667,
('dti', 'nn'): 0.39826676714393366,
('nn', 'hvz'): 0.00492226467928711,
('hvz', 'at'): 0.0958904109589041,
('at', 'jj'): 0.19933072429243945,
('jj', 'nn'): 0.444405819268941304,
('nn', 'to'): 0.016808622349269317,
('vb', 'to'): 0.035559524990903375,
('vb', 'nns'): 0.024908206807581622,
('nns', 'pps'): 0.004576705175507571,
('pps', 'vbz'): 0.075503502040319142,
('vbz', '.'): 0.028515683625994298,
('<bos>', 'at'): 0.1439727168794931,
('jj', 'nns'): 0.18830931813046425,
('nns', 'bed'): 0.021774632553075667,
('bed', 'at'): 0.058229552339279705,
('in', 'vbn'): 0.005079979716959388,
('vbn', 'nns'): 0.04291976367448066,
('nns', '--'): 0.008629206233996652,
('--', 'cs'): 0.050724637681159424,
('nn', 'bedz'): 0.017282618207274742,
('bedz', 'rb'): 0.11340088807924399,
('rb', 'vbn'): 0.08389333171161857,
('vbn', 'in'): 0.3594816085382123,
('in', 'at'): 0.3589176231964228,
('at', 'jj-tl'): 0.014542567069366456,
('jj-tl', 'nn-tl'): 0.4289972899728997,
('nn-tl', '--'): 0.004986702127659575,
('--', 'cc'): 0.1274703557312253,
('cc', 'np'): 0.05211275894538606,
('np', 'bedz'): 0.02745627572016461,
('rb', '*'): 0.002402164989205461,
('*', 'at'): 0.06983511154219205,
(',', 'cc'): 0.15507450741263004,
('cc', 'in'): 0.053848870056497175,
('nn', 'cc'): 0.05913280634716915,
('in', 'nns'): 0.04464112847462315,
('nns', '.'): 0.1394987802173431,
('<bos>', 'np'): 0.06570814036855466,
('np', 'np'): 0.22173996913580246,
('np', ','): 0.16316229423868311,
(',', 'at'): 0.09221006898128739,
(',', 'vbd'): 0.02585845497160715,
('vbd', 'ppl'): 0.007769380996858283,
```

```
('ppl', 'rb'): 0.04809437386569873,
('rb', 'in'): 0.144485967099461794,
('in', 'cd'): 0.03330106485963214,
('cd', ','): 0.10286232780664852,
(',', 'vbg'): 0.03969282365943824,
('vbg', 'cs'): 0.018787466782028306,
('cs', 'np'): 0.05298578675104214,
('np', 'md'): 0.016814557613168725,
('md', '*'): 0.054833796254815874,
('*', 'be'): 0.042919495635305525,
('be', 'vbn'): 0.48241294407316215,
('vbn', 'cc'): 0.03159900895749952,
('cc', 'vbn'): 0.026865583804143126,
('cc', 'cs'): 0.020185969868173257,
('cs', 'in'): 0.03550801064738084,
('in', 'dt'): 0.02732678744295395,
('dt', 'nn'): 0.4872332552115456,
('nn', 'at'): 0.006774494647492926,
('vb', 'ppo'): 0.09129701299990077,
('ppo', 'in'): 0.15785280350562694,
('nn', 'md'): 0.018237902167254908,
('md', 'vb'): 0.412239046680405,
('vb', '.'): 0.03850352287387119,
('rb', 'rb'): 0.027609693799981756,
(',', 'np'): 0.06389344106101605,
('np', 'rb'): 0.014564043209876544,
('rb', 'vbd'): 0.046857420865387536,
('vbd', ','): 0.054685267895049673,
('at', 'nn-tl'): 0.02652146786909421,
('nn-tl', 'md'): 0.014378324468085107,
('md', 'rb'): 0.08440103933339306,
('rb', 'vb'): 0.048859062851582692,
('np', '``'): 0.0013503086419753086,
('``', 'md'): 0.005681818181818182,
('in', 'vbg'): 0.031872032452864976,
('vbg', 'pp$'): 0.03633891601260738,
('pp$', 'nns'): 0.1917446920743769,
('nns', 'in'): 0.26276739450392145,
('in', 'dts'): 0.008528096621029825,
('dts', 'wps'): 0.09386446886446886,
('wps', 'vb'): 0.10623229461756374,
('vb', 'in'): 0.139955674638616,
('in', 'ppo'): 0.03325496704005901,
('in', 'pp$'): 0.0634121606048034,
('pp$', 'nn'): 0.5117367796386655,
('nn', "''"): 0.017960796896420966,
("''", '.'): 0.5265223446005823,
```

```
('<bos>', 'ppss'): 0.05514755750188928,
('ppss', 'nr'): 8.07297973682086e-05,
('nr', 'ber'): 0.0021156558533145277,
('ber', '*'): 0.046624203382165605,
('*', 'vbn'): 0.08050436469447139,
('vbn', 'to'): 0.05290642271774347,
('vb', 'jj'): 0.038106579339089015,
('nns', 'wps'): 0.01746002943607734,
('wps', 'vbd'): 0.24277620396600566,
('vbd', 'np'): 0.025897936656194276,
('np', 'to'): 0.006526491769547325,
('to', 'be'): 0.11590518208330218,
('be', 'ql'): 0.038515652479774884,
('ql', 'jj'): 0.5019717593181529,
('jj', 'cc'): 0.04294926390071785,
('cc', 'wps'): 0.0022069209039548023,
('vbd', 'pp$'): 0.04687101978432538,
('vb', 'at'): 0.17941847772154412,
('at', 'nns'): 0.07316658158924622,
('<bos>', 'pn'): 0.0048443040672776945,
('pn', 'md'): 0.08208955223880597,
('vb', '--'): 0.002017796301809401,
('cs', 'dt'): 0.02445884184621566,
('dt', 'bez'): 0.06833600592080918,
('bez', 'at'): 0.19739744155271283,
('nn', '--'): 0.007015138698480296,
('nn', 'vbd'): 0.021555873173292887,
('vbd', 'in'): 0.16748747558801053,
('pp$', 'vbg'): 0.014044573387841224,
('vbg', 'at'): 0.13868116927260368,
('nn-tl', 'in'): 0.06358045212765957,
('at', 'od-tl'): 0.0008848051727071635,
('od-tl', 'nn-tl'): 0.6783625730994152,
('nn-tl', ','): 0.12150930851063829,
('cc', 'rb'): 0.0650011770244821,
('rb', 'to'): 0.014899504363426278,
('ppo', ','): 0.11821531719948213,
(',', 'rb'): 0.05811959297229315,
('rb', 'cs'): 0.04694864232067382,
('cs', 'vb'): 0.0011049168801165186,
('vb', 'np'): 0.014025338228970263,
('np', 'vb'): 0.005304783950617284,
('ppo', 'cs'): 0.04332237824917837,
('at', 'vbg'): 0.01600589870115138,
('vbg', 'nn'): 0.1543785921760089,
('nn', 'wrb'): 0.004586821456698655,
('wrb', 'pp$'): 0.02340477950234048,
```

```
('ppo', '.'): 0.1766756299173389,
('pps', 'dod'): 0.014799926913941166,
('dod', 'vb'): 0.09503695881731784,
('jj', 'cs'): 0.023951471329497855,
('cs', 'nn'): 0.0448495806338205,
('dts', 'nns'): 0.48305860805860806,
('nns', ','): 0.13419625395673299,
('cs', 'np$'): 0.0045703380041183265,
('np$', 'nn'): 0.501737619461338,
(',', 'ppss'): 0.03048896680513739,
('ppss', 'hvd'): 0.042867522402518767,
('hvd', 'rb'): 0.08223534756928669,
('in', 'nps'): 0.0008758585718895496,
('nps', '.'): 0.10614035087719298,
('<bos>', 'pp$'): 0.018098319995349468,
('in', 'jj'): 0.06662978841100815,
('nn', 'hvd'): 0.004564944724790712,
('vbn', 'cs'): 0.0397179340575567,
('cs', 'ppss'): 0.109688112098883984,
('ppss', 'md'): 0.15217566803907323,
('vb', '``'): 0.002447818464490093,
('``', 'jj'): 0.069318181818182,
("''", 'in'): 0.044689201164704394,
('vbg', 'np'): 0.011247759718187998,
('np', 'at'): 0.004854681069958847,
('at', '``'): 0.006193636208950144,
('jj', "''"): 0.008760189804112422,
("''", 'nn'): 0.025826053930877325,
('nn', '.'): 0.13028323075576817,
('<bos>', 'vb'): 0.01441664890421842,
('vb', 'abn'): 0.005325659091660878,
('abn', 'dt'): 0.029076186970923814,
('dt', '--'): 0.002220303441470334,
('--', 'rb'): 0.06686429512516469,
(',', 'wrb'): 0.011547696177445787,
('wrb', 'jj'): 0.0349839862034984,
('jj', 'np'): 0.011645490414197069,
('np', 'cc'): 0.05623070987654321,
('cc', 'at'): 0.10604990583804143,
('at', 'np-tl'): 0.008246837955873177,
('np-tl', 'nn-tl'): 0.37596471885336274,
('nn-tl', 'nn-tl'): 0.14145611702127658,
('nn-tl', 'vb'): 0.0029920212765957447,
('nns', 'np'): 0.0017742293191394987,
('np', 'vbd'): 0.1110789609053498,
('in', 'abn'): 0.008823122666297884,
('abn', 'np'): 0.006256900993743099,
```

```
('vbd', 'at'): 0.14995329880275113,
('nn', 'pps'): 0.0059723478108683606,
('pps', 'vbd'): 0.3404592240696754,
('ppss', 'vb'): 0.23225962702833616,
('<bos>', 'wrb'): 0.014784816013331524,
('wrb', 'jj-tl'): 0.0014781966001478197,
('jj-tl', 'np-tl'): 0.2157181571815718,
('np-tl', 'nn'): 0.031422271223814774,
('nn', 'bez'): 0.0242248344660061898,
('bez', 'vbn'): 0.16707101896779886,
('in', 'np'): 0.054499469875074909,
('np', 'cd'): 0.01890432098765432,
(',', 'cd'): 0.017893212393765005,
('cd', '--'): 0.0029695619896066533,
('cs', 'pps'): 0.1368087991562453,
('pps', 'md'): 0.11023813874170169,
('cs', 'nn-tl'): 0.0058259253678870975,
('nn-tl', 'np'): 0.08194813829787234,
('np', 'hvd'): 0.020640432098765434,
('hvd', 'at'): 0.08814175374829623,
('vbg', ','): 0.040472677832025215,
(',', '``'): 0.018483936125614542,
('``', 'at'): 0.055808080808081,
('nps', 'cc'): 0.09473684210526316,
('cc', 'nps'): 0.001559557438794727,
('nps', 'bez'): 0.005263157894736842,
('np', '.'): 0.07420267489711935,
('.', '.'): 0.09696525167323737,
('nn-tl', 'bez'): 0.013214760638297872,
('at', 'np'): 0.022573875560093017,
('np', "'"): 0.017779063786008232,
("'", '--'): 0.020002531966071654,
('--', 'in'): 0.08333333333333333,
('dts', 'vbn'): 0.027472527472527472,
('vbn', 'jj'): 0.025004764627406136,
('nn', 'vbz'): 0.013169792608581513,
('vbz', 'rb'): 0.04682575416479064,
('rb', 'at'): 0.056952595250402895,
(',', 'in'): 0.07915698006783795,
('nn-tl', 'rb'): 0.008560505319148936,
('nn-tl', 'dod'): 0.0006648936170212766,
('dod', '*'): 0.418162261879619854,
('*', 'rb'): 0.14500484966052377,
('vbd', 'ql'): 0.008873227477286236,
('ql', 'rb'): 0.23355807149217656,
('in', 'wdt'): 0.011801041810722352,
('wdt', 'md'): 0.06598680263947211,
```

```
('md', 'hv'): 0.07902517695546994,
('hv', 'ben'): 0.1656372824256036,
('ben', 'at'): 0.0896121266161391,
('vbd', 'np$'): 0.002844527468795109,
('cs', 'dod'): 0.0008537994073627643,
('dod', 'rb'): 0.027455121436114043,
('rb', 'np'): 0.007176087815854289,
('<bos>', 'cs'): 0.038153738833879125,
('vbd', 'rp'): 0.05697546064362741,
('rp', 'in'): 0.3367837338262477,
('vbg', 'cc'): 0.025832766825288892,
('cc', 'vbd'): 0.059586864406779966,
('in', 'in'): 0.011008159314064445,
('cs', '*'): 0.004670784993219828,
('*', 'in'): 0.055286129970902036,
('nn-tl', '.'): 0.08626994680851063,
('<bos>', 'in'): 0.08378708314763501,
(',', 'nn-tl'): 0.007069629177941233,
('vbd', 'ppo'): 0.07595312898021568,
('ppo', 'rb'): 0.06353948809879494,
('ppo', '--'): 0.010955084154964646,
('cs', 'pp$'): 0.028928732861232483,
('pp$', 'jj'): 0.18172227350652775,
('<bos>', 'abn'): 0.0035072761447090513,
('dt', 'jj'): 0.0843715307758727,
(',', 'wdt'): 0.013720035062311826,
('wdt', 'in'): 0.01999600079984003,
('ppo', 'vbz'): 0.0011951000896325067,
('vbz', 'in'): 0.17574666066336486,
(',', 'vbz'): 0.00851785510118526,
('vbz', 'ppo'): 0.04937715743659012,
('ppo', 'to'): 0.045911761776715465,
('nn', 'rp'): 0.002056412799346615,
('rp', 'cc'): 0.05545286506469501,
('cc', 'rp'): 0.001942090395480226,
('rp', 'rb'): 0.05970425138632163,
('rb', 'vbg'): 0.025025085900203727,
('vbg', 'ppls'): 0.0016068228168839998,
('ppls', 'in'): 0.23961661341853036,
('<bos>', 'pps'): 0.10322243106555312,
('pps', 'bedz'): 0.1510445215908399,
('in', 'jj-tl'): 0.005753007882727147,
('np-tl', '('): 0.007717750826901874,
('(', 'wdt'): 0.015255905511811024,
('wdt', 'ppss'): 0.09278144371125775,
('vb', ')'): 0.0005954153021732659,
(')', 'dod'): 0.0019579050416054823,
```

```
('vbn', '.'): 0.0664379645511721,
('nn', 'cs'): 0.021825686200157512,
('bedz', 'vbn'): 0.22156438574518958,
('vbn', ','): 0.05542214598818372,
('vb', 'ppls'): 0.0035394131851410804,
('vbg', 'pps'): 0.0020394289598912306,
('abn', 'dts'): 0.02208317997791682,
('bez', 'rb'): 0.1143581826202029,
(',', 'dti'): 0.0024200617401577804,
('dti', 'ap'): 0.044084400904295405,
('ap', 'cs'): 0.034167245772527215,
('cs', 'bez'): 0.0014062578474210235,
('in', 'np-tl'): 0.004563684137740285,
('in', 'hvg'): 0.0006084912183653713,
('hvg', 'vbn'): 0.30078125,
('vbn', 'at'): 0.044787497617686294,
('np', 'np$'): 0.014949845679012346,
('np$', 'at-tl'): 0.009122502172024326,
('at-tl', 'nn-tl'): 0.4785185185185185,
('in', 'abx'): 0.001880791038583875,
('abx', 'nns'): 0.2579185520361991,
('nns', 'at'): 0.0074799895159176595,
('vbz', ':'): 0.01020561308719796,
(':', 'nn'): 0.060431654676258995,
('vbd', 'nn'): 0.020378704254054514,
('cd', 'jj'): 0.048420357997195414,
(',', 'abx'): 0.0016959487785357674,
('abx', 'jj'): 0.09200603318250378,
('cc', 'wrb'): 0.007444679849340866,
('wrb', 'vbn'): 0.02044838630204484,
('vbd', 'rb'): 0.06924513882992273,
('rb', '.'): 0.07756864414510292,
('vb', 'ppss'): 0.006979590486586616,
('ppss', 'abx'): 0.0013724065552595463,
('abx', 'hvd'): 0.00904977375565611,
('hvd', 'vbn'): 0.44366197183098594,
('nns', 'wdt'): 0.009334865622290772,
('wdt', 'pp$'): 0.012397520495900819,
('nns', 'vbd'): 0.025081150829653825,
('vbd', 'jj'): 0.02475163454190371,
('jj', ':'): 0.0010602610675612258,
(':', 'bez'): 0.0035971223021582736,
('bez', 'dt'): 0.01179973533303926,
('dt', 'at'): 0.0032071049710127053,
('nn', 'nns'): 0.039553113128956042,
('nns', 'vb'): 0.027581201233895843,
('<bos>', 'bez'): 0.001046369678531982,
```

```
('bez', 'pps'): 0.008270842523158359,
('pps', 'in'): 0.0014617211766855473,
('abn', 'jj'): 0.05152741994847258,
('wrb', 'at'): 0.17935452081793546,
('vb', 'pp$'): 0.04988257087096027,
('nn', 'ppss'): 0.004820173263716711,
('ppss', 'ber'): 0.05126342132881247,
('ber', 'vbn'): 0.23388535031847132,
('nn', 'ql'): 0.003259633054283464,
('hvz', 'rb'): 0.091324200913242,
('in', '.'): 0.0013276172037062647,
(',', 'ex'): 0.005678570067456839,
('ex', 'bez'): 0.25914634146341464,
('at', 'cd'): 0.010163915830072033,
('cd', 'nns'): 0.26247628474799967,
('vbg', 'ppo'): 0.037327730053766765,
('ppo', ':'): 0.003386116920625436,
(':', 'pps'): 0.02877697841726619,
('cs', 'cs'): 0.030033649741349002,
(',', 'vb'): 0.0149014825260109,
('ppo', 'jj'): 0.026989343690867442,
('rb', 'pps'): 0.016936783531486605,
('bedz', 'in'): 0.04838893316634407,
('rb', 'vbz'): 0.01730166935263174,
('vbz', 'at'): 0.18370103556956327,
('abn', 'in'): 0.1431726168568274,
('ppo', 'ppss'): 0.0055771337516183645,
('ppss', 'hv'): 0.055594574957616857,
('hv', 'cc'): 0.0019651880965749578,
('ppo', 'cc'): 0.03196892739766956,
('cc', 'jj'): 0.089159604519774,
('jj', ','): 0.054229746406410234,
(',', 'bedz'): 0.007565074888524715,
('bedz', 'at'): 0.1611066833655926,
('jj', 'cd'): 0.00285053795213182,
('cd', '.'): 0.07679617256454672,
('pps', 'md*'): 0.009744807844570314,
('md*', 'be'): 0.14857881136950904,
('vbn', 'rp'): 0.030836668572517628,
('in', 'rb'): 0.0084727792375420648,
('nns', 'cs'): 0.019677816085001715,
('cs', 'jj'): 0.03922454924413641,
('vbg', 'in'): 0.1632161176688709,
(':', 'dt'): 0.007194244604316547,
('bedz', 'jj'): 0.10975748605260162,
('rb', 'jj'): 0.04126250494116216,
(',', 'jj'): 0.04257021990167308,
```

```
('jj', 'in'): 0.0586793666243721,
('jj', 'jj'): 0.047120782855057095,
('cc', 'ql'): 0.010946327683615819,
('ql', 'vbn'): 0.061315354280625857,
('nns', 'wrb'): 0.004395249904231941,
('bed', 'vbn'): 0.30461124200605855,
('wrb', 'vbg'): 0.01084010840108401,
('vbg', 'nns'): 0.08757184352017798,
('nns', 'vbn'): 0.021048811467973143,
('nn-tl', 'cc'): 0.03316156914893617,
('cc', 'nn'): 0.1140830979284369,
('vbg', 'rb'): 0.04122118534083184,
('bed', 'jj'): 0.11141029956243689,
('nns', 'bez'): 0.006290449404221859,
('jj', '.'): 0.04060278448890202,
('<bos>', 'cc'): 0.049373147053694263,
('np', 'cs'): 0.009002057613168725,
('nn', 'wdt'): 0.006380713473149958,
('wdt', 'hvd'): 0.02659468106378724,
('vbn', 'ppl'): 0.001791499904707452,
('ppl', 'at'): 0.043557168784029036,
('wdt', 'bez'): 0.07038592281543692,
('bez', '*'): 0.051830613145125715,
('*', 'to'): 0.03516003879728419,
('ppo', 'rp'): 0.03147096902698934,
('rp', 'cs'): 0.025878003696857672,
('nns', 'hv'): 0.00967761446803363,
('hv', 'vbn'): 0.321729365524986,
(',', 'nns'): 0.0196844439193566827,
('vbd', 'bez'): 8.491126772522713e-05,
('bez', 'ql'): 0.053705337450374944,
('pps', 'hvz'): 0.023935684268225836,
('hvz', 'ben'): 0.2319634703196347,
('ben', 'rb'): 0.0664288898796255,
('cc', 'ppss'): 0.02739524482109228,
('rb', ':'): 0.001854836257487761,
(':', 'in'): 0.02877697841726619,
('wdt', 'nn'): 0.047390521895620874,
('nn', 'do'): 0.00021147507511011288,
('do', 'at'): 0.06162695152013147,
('jj', 'vb'): 0.0005388211982688196,
('ppls', 'vbn'): 0.03194888178913738,
('wdt', 'do'): 0.004999000199960008,
('do', 'ppss'): 0.09367296631059983,
('ppss', ','): 0.0063776539920884796,
('at', 'od'): 0.012761613067891782,
('od', 'nn'): 0.5910138248847926,
```

```
('in', 'ppls'): 0.0005900520905361176,
('ppls', '.'): 0.14057507987220447,
('<bos>', 'ex'): 0.014125990660181758,
('ex', 'md'): 0.09095528455284553,
('md', 'be'): 0.2693307051339486,
('be', 'at'): 0.09813577207175518,
('np', 'in'): 0.047228652263374485,
(',', 'vbn'): 0.01793132360227143,
('abn', 'nps'): 0.003312476996687523,
('nps', ','): 0.17982456140350878,
('pp$', 'ap'): 0.008044309639984175,
('ap', 'nns'): 0.24785730831596015,
('np', 'dod'): 0.0026363168724279837,
('*', 'vb'): 0.251454898157129,
('nn', ':'): 0.002946066563602952,
(':', '``'): 0.2827338129496403,
('vbn', 'ql'): 0.007928340003811702,
('nns', 'cc'): 0.07633218411661526,
('cc', 'wdt'): 0.005120056497175141,
('vb', 'rbr'): 0.0016539313949257386,
('rbr', 'in'): 0.13295346628679963,
('cd', 'dt'): 0.004206879485275922,
('dt', 'cs'): 0.0029604045886271126,
('cs', 'ap'): 0.01009492240470092,
('<bos>', 'nn'): 0.02412463425504292,
('cc', 'vbz'): 0.009121939736346516,
('vbz', 'rbr'): 0.0019510730901996097,
('<bos>', 'ql'): 0.004921812932354138,
('ql', 'ql'): 0.03790866301997201,
('vbd', '--'): 0.0019529591576802241,
('--', 'vb'): 0.01383399209486166,
('vb', ','): 0.03463332340974496,
(',', 'dt'): 0.007393574450245817,
('bez', 'np'): 0.0056241729157476845,
(',', '*'): 0.004592400625023819,
('*', 'np'): 0.004849660523763337,
('vbd', 'dt'): 0.008406215504797487,
('``', 'od'): 0.0007575757575757576,
('od', 'jj-tl'): 0.006912442396313645,
('nn-tl', "'"): 0.0343251329787234,
("'", 'wdt'): 0.0031649575895683,
('wdt', 'vbd'): 0.07598480303939212,
('jj-tl', 'nns'): 0.036585365853658534,
('nns', '('): 0.004657351962741184,
('(', 'in'): 0.09104330708661418,
('wdt', 'bed'): 0.009198160367926415,
('bed', 'rb'): 0.1120834735779199,
```

('nns', ')'): 0.003225871489344543,
(')', ','): 0.16837983357807146,
('bedz', '*'): 0.03654787657975635,
('*', 'pn'): 0.0021823472356935015,
('pn', 'wdt'): 0.005706760316066725,
('wdt', 'np'): 0.03099380123975205,
('vbn', 'pn'): 0.0012959786544692205,
('md', '--'): 0.001075172475584625,
('--', 'np'): 0.03458498023715415,
('np$', 'jjr'): 0.0017376194613379669,
('jjr', 'nn'): 0.34799774393683025,
('pps', 'wps'): 0.0009744807844570315,
('vb', 'wql'): 0.0007277298137673249,
('wql', 'ap'): 0.28846153846153844,
('ap', 'nn'): 0.2804030576789437,
('ppo', 'nns'): 0.005875908774026491,
('at', 'vbn'): 0.01478078384663377,
('vbn', 'nn'): 0.07882599580712789,
('md', ','): 0.009138966042469313,
('ppss', 'bem'): 0.015015742310486802,
('bem', 'jj'): 0.1625615763546798,
(',', 'wql'): 0.00043827889782385,
('wql', 'jj'): 0.358974358974359,
('jj', 'pps'): 0.001651226252759286,
('pps', 'be'): 0.0015226262257141117,
('be', ','): 0.009321139641224059,
('cc', 'pp$'): 0.023305084745762712,
('nns', 'bedz'): 0.003467811851045384,
('in', 'ql'): 0.0033835799566680497,
('ql', 'ap'): 0.06475003180256965,
('hvd', 'to'): 0.05406633348477965,
('in', ','): 0.0015027889180841746,
('at', 'ap'): 0.0307980262038455,
('ap', ','): 0.020847810979847115,
(',', 'nn'): 0.04129349441670795,
('<bos>', 'wdt'): 0.005910050962078788,
('wdt', 'rb'): 0.028394321135772844,
('rb', 'md'): 0.007358530726426856,
('md', '``'): 0.0012543678881820626,
('cc', 'nns'): 0.05373116760828625,
('nns', "''"): 0.014415613218008428,
("''", 'do'): 0.000126598303582732,
('do', 'wrb'): 0.003286770747740345,
('hvd', '``'): 0.0031803725579282144,
('``', 'vbn'): 0.012626262626262626,
(',', "''"): 0.0005716681275963261,
("''", 'jj'): 0.0812720848056537,

```
('jj', 'bez'): 0.0008864477777970904,
('nn-tl', "'"): 0.0007480053191489362,
("'", '``'): 0.0035335689045936395,
('``', '.'): 0.0006313131313131314,
('cc', ','): 0.012329331450094162,
('wrb', 'ppss'): 0.13944321261394432,
('jj', 'bed'): 0.0001911946187405489,
('jj', 'jj-tl'): 0.0022943354248865866,
('jj-tl', 'nn'): 0.07073170731707316,
(',', 'md'): 0.008594077518198103,
('md', 'ppss'): 0.017650748140847595,
('ppss', 'be'): 0.0019375151368370065,
('``', 'vb'): 0.04621212121212121,
('at', 'nps'): 0.006012137711984573,
('nps', 'vbg'): 0.014912280701754385,
('nps', 'in'): 0.10526315789473684,
('ppss', 'cc'): 0.0029870025026237186,
('ppss', 'do'): 0.014692823121013967,
('ex', 'ber'): 0.15701219512195122,
('ber', 'at'): 0.063694267516592357,
('at', 'ql'): 0.013771198457262776,
('wps', 'hv'): 0.03541076487252125,
('hv', 'rb'): 0.062324536777091152,
('rb', 'hvn'): 0.0006081430352418888,
('hvn', 'at'): 0.34101382488479265,
(',', 'wps'): 0.010156637066961393,
('vb', '*'): 0.0013893023717376203,
('ql', 'in'): 0.01933596234575754,
('in', 'nr'): 0.0016779006324620845,
('nr', '.'): 0.16290550070521861,
('np', 'hvz'): 0.007844650205761318,
('ben', 'vbn'): 0.45786892554614356,
('cc', 'ap'): 0.014712806026365348,
('wdt', 'to'): 0.011797640471905619,
('vb', 'cc'): 0.017300122390923225,
('vb', "''"): 0.011180576229697993,
('<bos>', 'dt'): 0.028639525645745734,
('dt', 'bedz'): 0.033921302578018994,
('bedz', 'nn'): 0.015826027553227828,
('cc', 'pps'): 0.03695856873822975,
('vbd', 'vbn'): 0.005519232402139764,
('``', 'rb'): 0.054924242424242424,
('cc', 'ber'): 0.0022951977401129945,
('ber', 'jjr'): 0.0035668789808917197,
('jjr', 'cs'): 0.12013536379018612,
('cs', 'dts'): 0.007885088644467882,
('ppo', "''"): 0.045911761776715465,
```

('--', 'wdt'): 0.016469038208168644,
('vbd', 'cs'): 0.04109705357900993,
('ppss', 'bed'): 0.04448211834988294,
('bed', 'ql'): 0.04039044092898014,
('jj', 'qlp'): 0.00260719934646203007,
('qlp', '.'): 0.11637931034482758,
('bedz', '``'): 0.005123534099965843,
('``', 'nn'): 0.08623737373737374,
("''", 'pps'): 0.002658564375237372,
('nns', 'hvd'): 0.005080747595717656,
('vbn', 'ppo'): 0.019955403087478559,
('nn', 'wps'): 0.008867368666686113,
('wdt', 'pps'): 0.09018196360727855,
('vbn', '--'): 0.003316180674671241,
('rb', 'cc'): 0.023291878249764343,
('at', 'jjt'): 0.006885599228631388,
('jjt', 'nn'): 0.42572062084257206,
('<bos>', '``'): 0.07270331544170365,
('``', 'cs'): 0.01893939393939394,
('ppo', 'nn'): 0.011552634199780898,
("''", ','): 0.23243448537789593,
('``', 'ppss'): 0.11186868686868687,
('vb', 'rp'): 0.040256690152492475,
('rb', "''"): 0.01730166935263174,
('hv', 'to'): 0.06513194834362718,
('ppls', 'cs'): 0.04792332268370607,
('cs', 'wrb'): 0.003164080156697303,
('wrb', 'np'): 0.053954175905395542,
('vbd', 'wdt'): 0.003226628173558631,
('``', 'cc'): 0.028535353535353535,
('nns', 'to'): 0.016613238170124396,
('vbd', 'ap'): 0.00725991339050692,
('nns', 'jjr'): 0.0006451742978689087,
('ppl', 'rp'): 0.023593466424682397,
('np-tl', ','): 0.09867695700110253,
('in', 'ap'): 0.012704559074355783,
('ap', 'nns$'): 0.0006949270326615705,
('nns$', 'nns'): 0.3876651982378855,
('vb', 'rb'): 0.04372994608183652,
('in', 'wrb'): 0.000912736827548057,
('jj', 'bedz'): 0.0004692958823631655,
('cd', 'vbd'): 0.0080001319805328713,
('vbd', 'to'): 0.04602190710707311,
('dt', 'in'): 0.0437893178734427,
('cs', 'vbn'): 0.015067048365225252,
('at', 'cd-tl'): 0.00030627871362940275,
('cd-tl', 'nps'): 0.003778337531486146,

```
('pps', 'bez'): 0.09951885011267433,
('vb', 'wrb'): 0.00790579206774503,
('ber', 'ql'): 0.05961783439490446,
('ql', 'rbr'): 0.022251621931052029,
('rbr', 'vbn'): 0.06362773029439696,
('cs', 'pn'): 0.010898498317512932,
('pn', ','): 0.04697102721685689,
('bez', 'in'): 0.041354212615791794,
('cs', 'cc'): 0.0020089397820300336,
('cc', '*'): 0.0066207627118644065,
('cd', 'rb'): 0.017798234760372845,
('od', 'in'): 0.05933179723502304,
('rb', 'cd'): 0.018426733967829234,
('hv', 'at'): 0.13447501403705783,
('rb', 'ppss'): 0.012375710767172438,
(',', 'to'): 0.007603186097031137,
('cc', 'to'): 0.007532956685499058,
('rp', ','): 0.08650646950092422,
('vbg', 'abx'): 0.0004944070205796922,
('abx', 'ppo'): 0.0015082956259426848,
('hv', '*'): 0.011510387422796182,
('vb', 'wdt'): 0.007111904998180676,
('wdt', 'bedz'): 0.04239152169566087,
('bedz', 'vbg'): 0.08357053398610953,
('at', 'rb'): 0.0035165333787079574,
('nn-tl', 'wps'): 0.002160904255319149,
('wps', 'md'): 0.1028328611898017,
('pp$', 'ql'): 0.005209020176711064,
('cs', ','): 0.009241122997338154,
('ap', 'jj'): 0.07841093351864721,
('vbd', 'nns'): 0.013246157765135434,
('*', 'wrb'): 0.001454898157129001,
('vbd', 'wrb'): 0.004542752823299652,
('wps', 'hvd'): 0.07053824362606233,
('vbn', 'rb'): 0.044368210405946254,
('rb', 'ap'): 0.0030407151762094444,
('wps', 'dod'): 0.009065155807365438,
('in', 'cs'): 0.00125386069238925,
('cs', 'vbg'): 0.019436492391140574,
('nn', 'nr'): 0.0010938365953971356,
('nr', 'bez'): 0.014809590973201692,
('bez', 'cs'): 0.02337891486546096,
('vb', 'vb'): 0.0038040422083291986,
('pn', 'wps'): 0.03467954345917471,
('wps', 'vbz'): 0.10793201133144475,
('jj', 'np-tl'): 0.0010428797385848123,
('np-tl', 'nns-tl'): 0.04437706725468578,
```

```
('nns-tl', "'''"): 0.04630541871921182,
('pn', 'hvz'): 0.014486391571553995,
('hvz', 'vbn'): 0.3771689497716895,
('cc', 'hv'): 0.0018243879472693032,
('hv', 'nn'): 0.025547445255474453,
('to', 'do'): 0.026396470500261723,
('do', 'rb'): 0.07066557107641742,
('vbz', ','): 0.044424433438391114,
('be', 'jj'): 0.143510376636299682,
('cc', 'vb'): 0.04140183615819209,
('nn', 'rb'): 0.01646588688271155,
('nr', 'in'): 0.2143864598025388,
('vb', 'nn'): 0.04329992391915583,
('in', 'dti'): 0.009182685658968833,
('dti', '``'): 0.0015071590052750565,
('wdt', 'vbz'): 0.07118576284743051,
('cs', 'nns'): 0.0267691225955502,
('bez', 'wrb'): 0.003970004411116012,
('cc', 'vbg'): 0.027954331450094162,
('nns', 'rb'): 0.026331176031774834,
('np', 'wdt'): 0.0007716049382716049,
('dti', 'jjs-tl'): 0.00037678975131876413,
('jjs-tl', 'nn-tl'): 0.875,
('bez', 'wdt'): 0.003859726510807234,
('jj', 'md'): 0.00064310917212732008,
('be', '``'): 0.0047485051002462185,
('vbz', 'cd'): 0.00945520036019811,
('<bos>', 'cd'): 0.012459550061038232,
('cd', 'bez'): 0.0035469768209189143,
('cs', 'wdt'): 0.0021093867711315353,
('jj', 'hvz'): 0.0001911946187405489,
('hvz', 'to'): 0.023287671232876714,
('do', 'in'): 0.09778142974527526,
('dti', 'jj'): 0.15749811605124342,
('in', '``'): 0.003319043009265662,
("'''", 'cc'): 0.026332447145208254,
('jj', '``'): 0.002520292701579963,
('``', 'jjr'): 0.0007575757575757576,
('jjr', 'nns'): 0.16187253243090807,
(',', 'bez'): 0.008212965433133884,
('jj', 'nns-tl'): 0.0005388211982688196,
('nns-tl', 'in-tl'): 0.06748768472906404,
('in-tl', 'np-tl'): 0.27786144578313254,
('np-tl', 'in'): 0.03721058434399118,
('nns', 'md'): 0.022944010967963065,
('cs', 'rb'): 0.0178795640600673,
('abn', 'nns'): 0.1409642988590357,
```

```
('<bos>', 'wql'): 0.000697579785687988,
('wql', 'vbn'): 0.04487179487179487,
('vbn', 'dt'): 0.004917095483133219,
('dt', 'md'): 0.0235598865178241,
('*', 'nn'): 0.010426770126091174,
('vb', 'dt'): 0.016109291786576694,
('dt', 'ql'): 0.004934007647711854,
('<bos>', 'ap'): 0.008157808049295638,
('nns', 'do'): 0.0015726123510554648,
('do', '*'): 0.14872637633525063,
('nn', 'vbg'): 0.007226613773590409,
('rb', 'nn'): 0.008666038252196917,
('vbg', 'dt'): 0.009393733391014153,
('dt', ','): 0.028863944739114345,
('hv', 'jj'): 0.026108927568781585,
(',', 'do'): 0.000724112961622013,
('*', 'hv'): 0.02061105722599418,
('vbg', 'jj'): 0.035473703726559292,
('wql', 'rb'): 0.2564102564102564,
('nns', 'nns'): 0.0026815056755176516,
('wdt', 'at'): 0.08378324335132974,
('nn-tl', 'vbz'): 0.010804521276595746,
('in', 'fw-nn'): 9.219563914626838e-05,
('fw-nn', '.'): 0.08846153846153847,
('at', 'fw-nn'): 0.000771368612103681,
('fw-nn', 'in'): 0.07307692307692308,
('in', 'nns-tl'): 0.0007836629327432812,
('nns-tl', 'cc'): 0.053694581280788176,
('at', 'nns-tl'): 0.0029380069196301966,
('nns-tl', 'vbz'): 0.00935960591133005,
('vbz', 'vbg'): 0.0060033018159988,
('nn', 'vbn'): 0.01108421083335764,
('jj', 'vbg'): 0.006604905011037144,
('np', 'vbz'): 0.016107253086419752,
('``', 'pp$'): 0.013383838383838383,
('dt', 'vbz'): 0.021709633649932156,
('abx', 'in'): 0.13574660633484162,
('vb', 'hvg'): 0.00026462902318811815,
('hvg', 'nn'): 0.0703125,
('in', 'wpo'): 0.0008113216244871618,
('wpo', ','): 0.027559055118110236,
('to', 'hv'): 0.023405369027144247,
('hv', 'dti'): 0.011510387422796182,
('abn', '.'): 0.044534412955465584,
('vbz', 'cc'): 0.013357346540597328,
('vbz', 'ppl'): 0.007354044724598529,
('ppl', 'cs'): 0.039927404718693285,
```

```
('jj', 'nn-tl'): 0.004049849651504354,
('nn-tl', 'in-tl'): 0.0890126329787234,
('np-tl', 'cc'): 0.026460859977949284,
('jj', 'rb'): 0.004606052178749587,
('rb', 'ql'): 0.0096669474260346034,
('rbr', 'vbz'): 0.00949667616334283,
('wps', 'ber'): 0.04759206798866856,
('ber', 'vbg'): 0.06828025477707006,
('np-tl', 'md'): 0.006339581036383682,
('rb', 'be'): 0.010095174385015355,
('ap', '.'): 0.02374334028260366,
('<bos>', 'dtx'): 0.000348789892843994,
('dtx', 'bez'): 0.010752688172043012,
('vbz', 'wrb'): 0.0057031367251988594,
('wrb', 'dtx'): 0.0004927322000492732,
('dtx', 'nn'): 0.3870967741935484,
('vbz', 'rp'): 0.02656461053579469,
('rp', 'at'): 0.10499075785582256,
('cc', 'dt'): 0.010298964218455744,
('wdt', 'ql'): 0.003599280143971206,
('<bos>', 'jj-hl'): 0.0032359951169415,
('jj-hl', 'nns-hl'): 0.2540983606557377,
('in-tl', 'nns-tl'): 0.09186746987951808,
('nns-tl', '.'): 0.1167487684729064,
('wrb', 'pps'): 0.18649913771864993,
('pps', 'hvd'): 0.07168524270662038,
(',', 'pp$'): 0.011985975075269637,
('bed', 'nns'): 0.029956243688993606,
('pp$', '``'): 0.00527495714097323,
("''", 'nns'): 0.010634257500949488,
('wps', 'bez'): 0.056090651558073655,
('bez', ','): 0.020291133656815175,
('vb', 'abx'): 0.0010585160927524726,
('abx', 'at'): 0.10256410256410256,
('abx', '.'): 0.02262443438914027,
('dt', 'ap'): 0.02269643517947453,
('cd', 'nn'): 0.11705023509032418,
('ber', 'nns'): 0.028789808917197453,
('bez', 'vbg'): 0.03694309660344067,
('vbg', '.'): 0.03602991162474507,
('bedz', 'ql'): 0.044290105886371396,
('pps', 'ql'): 0.0010962908825141605,
('ql', 'vbd'): 0.0008904719501335708,
('vbd', 'cc'): 0.014477371147151226,
('nn', 'np'): 0.0042367937461715716,
('vb', 'ql'): 0.0118752274155566802,
('np-tl', 'cs'): 0.005237045203969129,
```

```
('vbz', 'jj'): 0.04007203962179198,
(',', 'abn'): 0.0024200617401577804,
('abn', 'at'): 0.16709606183290393,
('pps', 'doz'): 0.0056641695596556496,
('doz', 'ap'): 0.004842615012106538,
('ap', 'cc'): 0.010655547834144082,
('cc', 'pn'): 0.003295668549905838,
('pn', 'to'): 0.03292361720807726,
('rp', 'vbg'): 0.007024029574861368,
('cs', 'cd'): 0.010546933855657676,
('cd', 'in'): 0.13577497319145426,
('--', 'bez'): 0.006258234519104084,
('in', 'ppl'): 0.0022495735951689484,
('ppl', 'in'): 0.1823956442831216,
('vbz', 'ql'): 0.019660813447396067,
('dti', 'in'): 0.1333835719668425,
('nns', 'ber'): 0.02826669892538156,
('ber', 'jj'): 0.14445859872611466,
('nn', 'vb'): 0.0039742729632276259,
('nns', 'vbg'): 0.01330671989354624,
('vbz', 'nns'): 0.01785982290259643,
('ap', 'at'): 0.009381514940931202,
('in', 'wql'): 0.00013829345871940258,
('jj', 'abn'): 0.0003650079085046843,
('bez', '.'): 0.0034186149095721216,
('vbg', 'to'): 0.03535010197144799,
('cs', 'nps'): 0.0015569283310732762,
('*', 'vbg'): 0.03588748787584869,
('ppo', 'vb'): 0.039836669654416894,
('ap', 'in'): 0.1296038915913829,
('wdt', 'jj'): 0.014597080583883223,
('<bos>', 'to'): 0.003933574902629488,
('rb', 'wdt'): 0.0030711223279715387,
('ppss', 'rb'): 0.04270606280778235,
('cs', 'abl'): 0.001255587363768771,
('abl', 'at'): 0.9660493827160493,
('vbz', 'to'): 0.06243433888638751,
('vb', 'ppl'): 0.005689523998544541,
('ppl', '*'): 0.0018148820326678765,
('hv', 'in'): 0.014879281302638967,
('ppo', 'pn'): 0.001693058460312718,
('pn', 'dt'): 0.0030728709394205445,
('<bos>', 'vbg'): 0.010424942352781599,
('ap', 'vbg'): 0.007065091498725967,
('``', 'do'): 0.00404040404040404,
('np', 'nn-tl'): 0.004275977366255144,
('``', 'wrb'): 0.0208333333333332,
```

```
('wrb', ','): 0.04607046070460705,
('``', 'dt'): 0.012752525252525253,
('cs', '``'): 0.011099392295715936,
('pp$', "'"): 0.0001978108927864961,
('at', 'jjr'): 0.00654528954682094,
('vbz', '*'): 0.0037520636349992497,
('bez', 'pp$'): 0.0105866784296427,
('*', 'pp$'): 0.005334626576613967,
('vbz', 'wdt'): 0.005402971634398919,
('bez', 'to'): 0.025033083370092635,
('be', 'cc'): 0.0015828350334154062,
('cc', 'do'): 0.0009121939736346516,
('do', '.'): 0.06820049301561217,
('jj-hl', 'nn-hl'): 0.5136612021857924,
('nn-hl', 'in-hl'): 0.1530993278566094,
('in-hl', 'at-hl'): 0.28820960698689957,
('at-hl', 'jj-hl'): 0.19805194805194806,
('vbz', 'pp$'): 0.032267747260993546,
('be', 'rb'): 0.04484699261343651,
('rb', 'dt'): 0.004956365737221394,
('nns', 'vbz'): 0.0024194036170084076,
('in', 'pn'): 0.0023325496704005902,
('pn', 'jj'): 0.05267778753292362,
('*', 'cs'): 0.01600387972841901,
('hvz', 'pp$'): 0.011872146118721462,
('in', 'beg'): 0.0014474715345964137,
('beg', 'at'): 0.10952380952380952,
('jj', 'ql'): 0.0006778718300801279,
('cs', 'abn'): 0.006679724775249862,
('abn', 'cc'): 0.004416635995583364,
('cc', 'dti'): 0.003266242937853107,
('vb', 'dti'): 0.008831993648903443,
('<bos>', 'cd-hl'): 0.0038366888212839345,
('cd-hl', '.-hl'): 0.5147783251231527,
('<bos>', 'at-hl'): 0.0020733621407948532,
('at-hl', 'nn-hl'): 0.4935064935064935,
('nn-hl', 'nn-hl'): 0.13666915608663183,
('nn', 'jj'): 0.006103608202316016,
(',', 'pn'): 0.002744007012462365,
('dts', 'nn'): 0.0173992673992674,
('wdt', 'vb'): 0.03179364127174565,
('cc', 'dts'): 0.0026188794726930032,
('dts', 'ap'): 0.011904761904761904,
('od', ','): 0.051267281105990783,
("'", 'rb'): 0.004810735536143816,
('(', 'cc'): 0.04675196850393701,
(')', 'in'): 0.09887420460107685,
```

```
('rb', 'ber'): 0.0018244291057256667,
('vbg', '``'): 0.0028428403683332304,
('--', 'at'): 0.13076416337285904,
('at', 'nn$'): 0.009335828937666661,
('nn$', '``'): 0.004447739065974796,
('ex', 'vbz'): 0.009146341463414634,
('wdt', ','): 0.021395720855828835,
('pps', 'bed'): 0.00395882818685669,
('nn-tl', 'hvz'): 0.006399601063829787,
('vbn', 'vbn'): 0.002287021154945683,
('vbz', '``'): 0.005402971634398919,
('nn', '('): 0.003770090132135461,
('(', 'rb'): 0.05265748031496063,
('nn', ')'): 0.002851267392001867,
(')', 'wdt'): 0.004894762604013705,
('<bos>', ')'): 0.003080977386788614,
(')', 'hvz'): 0.011257953989231522,
('dts', 'jj'): 0.10073260073260074,
('nns', ':'): 0.003971854271255469,
(':', 'rb'): 0.017266187050359712,
('rb', 'dts'): 0.0014291361328184389,
('ber', 'rb'): 0.12713375796178344,
(':', 'at'): 0.10431654676258993,
('nns', '``'): 0.0008467912659529426,
('vbn', "''"): 0.007356584715075281,
('be', 'nns'): 0.007914175167077032,
('nn-tl', 'nn'): 0.03873005319148936,
('nn', '``'): 0.0013636496222617625,
("''", 'hvz'): 0.0015191796429927839,
('cs', 'to'): 0.0065792777861483605,
('dti', 'cd'): 0.022230595327807082,
('jj', '('): 0.0010950237255140528,
('(', 'jj'): 0.03641732283464567,
(',', 'ber'): 0.003201341514539426,
('dt', '``'): 0.003083754779819909,
("''", 'vbz'): 0.006836308393467528,
('vbz', 'ap'): 0.013357346540597328,
('ap', 'to'): 0.013435255964790364,
('nn', 'ber'): 0.0032669252982527783,
('abn', ','): 0.05042326094957674,
('rb', 'abn'): 0.002675829355064311,
('vb', 'cd'): 0.009592802090569284,
('(', 'vbg'): 0.009350393700787402,
(')', 'vbz'): 0.016152716593245228,
('vbz', 'abx'): 0.00210115563559958,
('cc', '``'): 0.005620291902071563,
("''", ')'): 0.004430940625395619,
```

(',', 'uh'): 0.0015053927360036588,
('uh', ','): 0.5581818181818182,
('cd', 'ql'): 0.0020621958261155648,
("''", 'ber'): 0.0017723762501582478,
('in', '('): 0.0002581477896095515,
('vbn', '``'): 0.0039260529826567565,
('in', ')'): 6.453694740238787e-05,
(')', 'at'): 0.0298580518844836,
('cs', 'ber'): 0.0008537994073627643,
('dti', 'vbg'): 0.009796533534287867,
("''", 'vbn'): 0.003418154196733764,
('dti', 'vbn'): 0.0146948003014318,
('(', 'np'): 0.18799212598425197,
('np', ')'): 0.011541923868312758,
(')', 'cs'): 0.022026431718061675,
(',', '('): 0.0007622241701284348,
('cc', '('): 0.0013830037664783427,
(',', 'hvz'): 0.002477228552917413,
('dti', 'rb'): 0.007912584777694047,
('ql', 'jjr'): 0.02455158376796845,
('jjr', 'cc'): 0.053581500282200789,
('dti', 'jjr'): 0.006782215523737754,
('jjr', '.'): 0.043993231810490696,
('be', 'vbg'): 0.032887794583186775,
(',', 'ql'): 0.009108578833034795,
('vbn', 'vbg'): 0.0038879359634076617,
('<bos>', 'jj'): 0.018311469374309688,
('vbg', 'vbn'): 0.0038316544094926146,
("''", 'md'): 0.004557538928978352,
('at-hl', 'nns-hl'): 0.08116883116883117,
('nns-hl', 'in-hl'): 0.15892857142857142,
('vbg', 'wdt'): 0.003151844756195538,
('*', 'ql'): 0.03491755577109602,
('vbn', '('): 0.0013340956737183153,
('bed', 'to'): 0.020531807472231572,
('hv', 'pp$'): 0.01347557551937114,
('rb', 'doz'): 0.0006385501870039834,
('doz', 'hv'): 0.014527845036319613,
('dti', 'nns'): 0.12773172569706104,
('md', 'cc'): 0.003673505958247469,
('cc', 'md'): 0.007532956685499058,
('vb', 'dts'): 0.004498693394198009,
('dt', 'vbn'): 0.008634513383495744,
('<bos>', 'nns'): 0.016974441451741044,
('nn-hl', 'nns-hl'): 0.06646751306945482,
('<bos>', 'dts'): 0.007247078884647431,
('dts', 'ber'): 0.028846153846153848,

```
…})
```

[59]: `# emission/observation probabilities of word and its corresponding observed tag`
`pos_tagger.emis_prob`

[59]:
```
defaultdict(<function __main__.POSTagger.compute_emis_prob.<locals>.<lambda>()>,
            {('rb', 'furthermore'): 0.0010946574634354,
             (',', ','): 0.9999428331872404,
             ('cs', 'as'): 0.2737180453015921,
             ('at', 'an'): 0.03787646758550281,
             ('nn', 'encouragement'): 0.00010209141557039932,
             ('in', 'to'): 0.09145807403309823,
             ('nn', 'revisionist'): 7.292243969314237e-06,
             ('nn', 'thinking'): 0.00024793629495668406,
             ('pps', 'it'): 0.3190206468116207,
             ('rb', 'manifestly'): 0.0001520357588104722,
             ('bez', 'is'): 0.9998897220996912,
             ('jj', 'fair'): 0.000938591764726331,
             ('to', 'to'): 0.9999252224631721,
             ('vb', 'admit'): 0.001157751976448017,
             ('cs', 'that'): 0.2908442569433981,
             ('dti', 'any'): 0.4547852298417483,
             ('nn', 'fraternity'): 3.646121984657119e-05,
             ('hvz', 'has'): 0.9986301369863013,
             ('at', 'a'): 0.23484771141738983,
             ('jj', 'constitutional'): 0.0003823892374810978,
             ('nn', 'right'): 0.0007802701047166234,
             ('vb', 'refuse'): 0.000430022162680692,
             ('vb', 'accept'): 0.0020508749297079155,
             ('nns', 'persons'): 0.0021371398616907597,
             ('vbz', 'dislikes'): 0.00045024763619990995,
             ('.', '.'): 0.8060694966535252,
             ('at', 'the'): 0.7037604219840055,
             ('jj', 'unitarian'): 0.00012166930283489475,
             ('nns', 'clergy'): 4.032339361680679e-05,
             ('bed', 'were'): 0.9996634129922585,
             ('jj', 'exclusive'): 0.00039977056645751135,
             ('nn', 'club'): 0.0004958725899133681,
             ('in', 'of'): 0.2900751394459042,
             ('vbn', 'cultivated'): 0.00026681913474366307,
             ('nns', 'gentlemen'): 0.0003629105425512611,
             ('--', '--'): 1.0,
             ('nn', 'term'): 0.0004667036140361112,
             ('bedz', 'was'): 1.0,
             ('rb', 'then'): 0.036488582114513335,
             ('vbn', 'understood'): 0.0012959786544692205,
             ('in', 'in'): 0.17160374314294935,
```

```
('jj-tl', 'back'): 0.0005420054200542005,
('nn-tl', 'bay'): 0.0024933510638297874,
('cc', 'and'): 0.7571209981167608,
('np', 'parker'): 0.001768261316872428,
('rb', 'definitely'): 0.0005473287317177,
('*', 'not'): 0.9992725509214355,
('nn', 'gentleman'): 0.00018230609923285593,
('cc', 'either'): 0.005031779661016949,
('nn', 'theology'): 0.00013126039144765629,
('cc', 'or'): 0.1100517890772128,
('nns', 'manners'): 0.00028226375531764756,
('np', 'ezra'): 6.430041152263375e-05,
('np', 'stiles'): 3.2150205761316875e-05,
('np', 'gannett'): 9.645061728395061e-05,
('jj', 'honorable'): 0.00015643196078772182,
('nn', 'representative'): 0.00016042936732491324,
('nn', 'sanhedrin'): 7.292243969314237e-06,
('vbd', 'addressed'): 0.000297189437038295,
('ppl', 'himself'): 0.49183303085299457,
('rb', 'frankly'): 0.0003344786693830389,
('nn', 'issue'): 0.0008750692763177085,
('cd', '1845'): 0.00041243916522312957,
('vbg', 'insisting'): 0.00037080526543476915,
('md', 'should'): 0.07078218797598781,
('be', 'be'): 1.0,
('vbn', 'persecuted'): 0.00011435105774728416,
('vbn', 'calumniated'): 3.8117019249094724e-05,
('dt', 'this'): 0.5732083384729246,
('nn', 'republic'): 4.3753463815885424e-05,
('at', 'no'): 0.018524190346548692,
('nn', 'power'): 0.0021512119709477,
('vb', 'restrain'): 0.0003307862789851477,
('ppo', 'him'): 0.2351359426351957,
('in', 'by'): 0.0433872677822339,
('nn', 'force'): 0.0011521745471516495,
('md', 'could'): 0.13027506495833707,
('vb', 'exist'): 0.0016539313949257386,
('rb', 'even'): 0.027183993675312434,
('rb', 'so'): 0.015750904612764922,
('rb', 'judiciously'): 6.081430352418889e-05,
('vbd', 'argued'): 0.0006368345079392036,
('nn-tl', 'association'): 0.006399601063829787,
('rb', 'legitimately'): 6.081430352418889e-05,
('vb', 'decide'): 0.0010585160927524726,
('``', '``'): 1.0,
('vbn', 'encouraged'): 0.0008004574042309891,
('cc', 'nor'): 0.005296610169491525,
```

('vbn', 'assisted'): 0.00011435105774728416,
('vbg', 'diffusing'): 0.00012360175514492306,
('pp$', 'his'): 0.41144665699591193,
('nns', 'opinions'): 0.000766144478719329,
('dts', 'those'): 0.35210622710622713,
('wps', 'who'): 0.569971671388102,
('vb', 'differ'): 0.0005292580463762363,
('in', 'from'): 0.03626054487622735,
('nn', 'regard'): 0.00034273546655776913,
('pp$', 'their'): 0.15831465119345906,
('nn', 'correctness'): 1.4584487938628475e-05,
("'''", "'''"): 1.0,
('ppss', 'we'): 0.19173326874949545,
('nr', 'today'): 0.17912552891396333,
('ber', 'are'): 0.9984713375796178,
('vbn', 'entitled'): 0.001905850962454736,
('vb', 'excoriate'): 3.307862789851477e-05,
('jj', 'honest'): 0.0007473971459857821,
('nns', 'men'): 0.013891409100989939,
('vbd', 'believed'): 0.0019105035238176106,
('ql', 'downright'): 0.0006360513929525506,
('jj', 'pernicious'): 1.7381328976413536e-05,
('vbd', 'barred'): 8.491126772522713e-05,
('nns', 'pulpits'): 2.0161696808403396e-05,
('in', 'against'): 0.00519061448393491,
('nn', 'demand'): 0.0005323338097599393,
('vb', 'poison'): 3.307862789851477e-05,
('nns', 'minds'): 0.0010685699308453798,
('nns', 'congregations'): 0.0003427488457428577,
('pn', 'one'): 0.22036874451273047,
('md', 'can'): 0.1419227667771705,
('vb', 'argue'): 0.000860044325361384,
('cs', 'though'): 0.0179800110491688,
('jj', 'delicate'): 0.0004692958823631655,
('nn', 'matter'): 0.0018522299682058162,
('at', 'every'): 0.004968521354432534,
('nn', 'justification'): 0.00010938365953971357,
('vbd', 'existed'): 0.0010189352127027256,
('in', 'for'): 0.07332319181302724,
('vbg', 'returning'): 0.0020394289598912306,
('jj-tl', 'public'): 0.007588075880758808,
('nn-tl', 'lecture'): 8.311170212765957e-05,
('od-tl', 'first'): 0.2631578947368421,
('nn-tl', 'church'): 0.008976063829787235,
('vb', 'suppress'): 0.00016539313949257386,
('ppo', 'it'): 0.25605019420376457,
('rb', 'rather'): 0.0038313011220239,

```
('cs', 'than'): 0.058861935613479986,
('vb', 'let'): 0.009758195230061857,
('vb', 'use'): 0.006946511858688101,
('vbg', 'sounding'): 0.00012360175514492306,
('nn', 'board'): 0.0010719598634891928,
('nn', 'propaganda'): 0.00016042936732491324,
('wrb', 'when'): 0.5112096575511209,
('nn', 'turn'): 0.0005906717615144532,
('vb', 'come'): 0.013032979392014482,
('vb', 'occupy'): 0.0005292580463762363,
('rb', 'finally'): 0.005047587192507677,
('dod', 'did'): 0.9978880675818373,
('vb', 'seem'): 0.0066488042076014685,
('jj', 'clear'): 0.003076495228825196,
('nn', 'day'): 0.004003441939153516,
('dts', 'these'): 0.6428571428571429,
('nns', 'clergymen'): 0.00010080848404201697,
('np$', "gannett's"): 0.0004344048653344917,
('nn', 'son'): 0.0008896537642563369,
('vbd', 'explained'): 0.002462426764031587,
('nn', 'biography'): 8.750692763177085e-05,
('nn', 'father'): 0.0010719598634891928,
('ppss', 'they'): 0.26398643739404215,
('hvd', 'had'): 1.0,
('rb', 'always'): 0.012771003740079667,
('vbn', 'contended'): 0.0001905850962454736,
('nn', 'propriety'): 5.104570778519966e-05,
('nn', 'claim'): 0.00045211912609748273,
('nn', 'title'): 0.00033544322258845493,
('nps', 'christians'): 0.013157894736842105,
('jj-tl', 'calvinist'): 0.00027100271002710027,
('nn-tl', 'orthodoxy'): 0.0002493351063829787,
('jj', 'intellectual'): 0.0009038291067735039,
('nn', 'liberty'): 0.0002698130268646268,
('rb', 'never'): 0.019460577127740443,
('vbn', 'meant'): 0.0011435105774728416,
('md', 'would'): 0.2174536331869904,
('vb', 'follow'): 0.002712447487678211,
('jj', 'free'): 0.0034762657952827074,
('nn', 'inquiry'): 0.0001166759035090278,
('nn', 'extreme'): 5.104570778519966e-05,
('vbg', 'proclaiming'): 0.00018540263271738457,
('np', 'christianity'): 0.0009002057613168725,
('jj', 'natural'): 0.002329098082839414,
('nn', 'religion'): 0.0007583933728086807,
('vb', 'grant'): 0.0003969435347821772,
('abn', 'all'): 0.9109311740890689,
```

```
('rb', 'still'): 0.020159941618268618,
('jj', 'modern'): 0.003076495228825196,
('np', 'unitarianism'): 3.2150205761316875e-05,
('np-tl', 'harvard'): 0.003031973539140022,
('nn-tl', 'divinity'): 0.0003324468085106383,
('nn-tl', 'school'): 0.005319148936170213,
('vb', 'recall'): 0.0008931229532598987,
('in', 'with'): 0.06033282625731803,
('jj', 'humorous'): 0.00020857594771696243,
('nn', 'affection'): 0.0001166759035090278,
('nns', 'insults'): 4.032339361680679e-05,
('vbd', 'lavished'): 4.245563386261357e-05,
('in', 'upon'): 0.003706264693679989,
('ppo', 'them'): 0.15904790359525944,
('rb', 'else'): 0.00465229421960045,
('in', 'after'): 0.005633153551836998,
('vbd', 'received'): 0.002462426764031587,
('nn', 'treatment'): 0.0007948545926552519,
('pps', 'he'): 0.5226871307631402,
('vbd', 'invited'): 0.0004245563386261357,
('vb', 'betray'): 0.00013231451159405908,
('jj', 'uneasy'): 0.0003302452505518572,
('nn', 'conscience'): 0.00026252078289531257,
('wrb', 'whenever'): 0.010347376201034738,
('jj-tl', 'new'): 0.13333333333333333,
('np-tl', 'england'): 0.014608599779492833,
('nn', 'liberalism'): 8.021468366245662e-05,
('vbn', 'reminded'): 0.00034305317324185246,
('jj', 'dramatic'): 0.000938591764726331,
('nn', 'confrontation'): 5.104570778519966e-05,
('in', 'on'): 0.050984188447886415,
('np', 'january'): 0.001511059670781893,
('cd', '23'): 0.0031345376556957847,
('cd', '1843'): 8.248783304462591e-05,
('cs', 'while'): 0.026919793079200245,
('md', 'may'): 0.10438132783800735,
('vb', 'defend'): 0.0006284939300717806,
('nn', 'privilege'): 0.00013126039144765629,
('np', 'chandler'): 0.0009323559670781893,
('np', 'robbins'): 0.00019290123456790122,
('vb', 'demand'): 0.0005954153021732659,
('vb', 'leave'): 0.0056564453706460256,
('vb', 'plead'): 0.00016539313949257386,
('nn-tl', 'dr.'): 0.014212101063829786,
('np', 'n.'): 0.0007716049382716049,
('np', 'l.'): 0.0016075102880658437,
('np', 'frothingham'): 6.430041152263375e-05,
```

```
('nn', 'warrant'): 6.563019572382814e-05,
('vbg', 'stating'): 0.0008652122860144614,
('nn', 'difference'): 0.000984452935857422,
('in', 'between'): 0.006084912183653713,
('nps', 'trinitarians'): 0.0008771929824561404,
('nps', 'unitarians'): 0.0017543859649122807,
('.', ';'): 0.0912074814339415,
('np', 'mr.'): 0.023823302469135804,
('in', 'despite'): 0.0008574194440602959,
('vbn', 'supposed'): 0.0018296169239565465,
('jj', 'conclusive'): 0.00017381328976413536,
('nns', 'assurances'): 6.048509042521019e-05,
('nn', 'liberal'): 5.104570778519966e-05,
('vbz', 'heaves'): 0.00015008254539996998,
('rb', 'repeatedly'): 0.0005473287317177,
('nn', 'sigh'): 6.563019572382814e-05,
('nn', 'relief'): 0.00042295015022022575,
('jj', 'positive'): 0.0011297863834668798,
('nn', 'thanksgiving'): 7.292243969314237e-06,
('rb', 'quite'): 0.00039529297290722776,
('vbd', 'brought'): 0.005094676063513629,
('ppl', 'itself'): 0.24319419237749546,
('rb', 'officially'): 0.0005473287317177,
('vb', 'expel'): 6.615725779702954e-05,
('rb', 'happily'): 0.0005777358834797944,
('nn', 'posterity'): 4.3753463815885424e-05,
('np-tl', 'boston'): 0.003031973539140022,
('rb', 'actually'): 0.00465229421960045,
('vb', 'command'): 0.00026462902318811815,
('nn', 'room'): 0.002435609485750955,
('vbd', 'came'): 0.02343550989216269,
('ql', 'too'): 0.06322350845948353,
('rb', 'close'): 0.0026454222033022168,
('nn', 'comfort'): 0.00027710527083394103,
('wdt', 'what'): 0.3327334533093381,
('hv', 'have'): 0.9991577765300393,
('ben', 'been'): 1.0,
('jj', 'unforgivable'): 1.7381328976413536e-05,
('nn', 'brutality'): 7.292243969314237e-05,
('rb', 'fortunately'): 0.0005777358834797944,
('nn', 'honor'): 0.0003135664906805122,
('nn', 'denomination'): 5.83379517545139e-05,
('vb', 'attest'): 6.615725779702954e-05,
('np', 'cyrus'): 0.0001286008230452675,
('np', 'bartol'): 3.2150205761316875e-05,
('vbd', 'defended'): 0.00033964507090090854,
('np$', "parker's"): 0.0004344048653344917,
```

```
('nn', 'sincerity'): 9.479917160108509e-05,
('rb', 'also'): 0.02949493720923161,
('cs', 'whereupon'): 0.00030134096730450503,
('vbd', 'broke'): 0.0026322492994820413,
('rp', 'down'): 0.11534195933456562,
('in', 'into'): 0.014797400082976074,
('nns', 'convulsions'): 2.0161696808403396e-05,
('vbg', 'weeping'): 0.00030900438786230763,
('vbd', 'rushed'): 0.0008066570433896578,
('in', 'out'): 0.0036786060019361085,
('nn-tl', 'fellowship'): 0.00024933510638296787,
('nn', 'hall'): 0.0006854709331155383,
('nn', 'adjournment'): 2.916897587725695e-05,
('vbd', 'took'): 0.016005773966205315,
('rb', 'warmly'): 0.00018244291057256667,
('nn', 'hand'): 0.002727299244523525,
('vbd', 'requested'): 0.00016982253545045427,
('vb', 'visit'): 0.0015216168833316794,
('pp$', 'our'): 0.07299221943821707,
('jj', 'burly'): 3.476265795282707e-05,
('np', 'theodore'): 0.00038580246913580245,
('rb', 'again'): 0.015811718916628911,
('vbd', 'burst'): 0.0004670119724887493,
('nns', 'tears'): 0.0006250126010605053,
('jj', 'near'): 0.00026071993464620307,
('nn', 'tragedy'): 0.0003135664906805122,
('wdt', 'which'): 0.6456708658268346,
('ppo', 'us'): 0.059854596155576138,
('vbz', 'borders'): 0.00030016509079993996,
('nn', 'comedy'): 0.00024793629495668406,
('vbz', 'enables'): 0.0012006603631997599,
('vb', 'tell'): 0.007575005788759882,
('nn', 'story'): 0.000984452935857422,
('rp', 'over'): 0.06469500924214418,
('vbg', 'warming'): 0.0005562078981521537,
('ppls', 'ourselves'): 0.19808306709265175,
('nn', 'glow'): 8.021468366245662e-05,
('nn', 'complacency'): 2.1876731907942712e-05,
('rb', 'indeed'): 0.0038921154255480888,
('nn', 'thing'): 0.0021657964588863285,
('cc', 'but'): 0.11275894538606403,
('rb', 'somehow'): 0.0020068720162982334,
('jj', 'inherent'): 0.00039977056645751135,
('nn', 'decency'): 6.563019572382814e-05,
('(', '('): 0.999507874015748,
('vb', 'inherit'): 0.00013231451159405908,
(')', ')'): 0.9995105237395986,
```

```
('vb', 'triumph'): 3.307862789851477e-05,
('vbn', 'excommunicated'): 0.00011435105774728416,
('nn', 'extent'): 0.0007438088848700522,
('vbn', 'ostracized'): 3.8117019249094724e-05,
('vbn', 'reviled'): 3.8117019249094724e-05,
('vb', 'solace'): 3.307862789851477e-05,
('vbg', 'saying'): 0.006736295655398307,
('vbd', 'asked'): 0.011802666213806573,
('rb', 'yet'): 0.007814638002858271,
('nns', 'stratagems'): 2.0161696808403396e-05,
('jj', 'christian'): 0.0017902768845705942,
('nn', 'liberality'): 1.4584487938628475e-05,
('vbn', 'laid'): 0.001639031827711073,
('nn', 'rest'): 0.0008677770323483943,
('ap', 'more'): 0.1038915913829048,
('nn-tl', 'university'): 0.009474734042553192,
('hvg', 'having'): 0.98828125,
('vbn', 'done'): 0.010558414331999237,
('jj', 'abject'): 5.214398692924061e-05,
('nn', 'penance'): 2.1876731907942712e-05,
('pp$', 'its'): 0.11031254121060266,
('nn', 'rejection'): 7.292243969314237e-05,
('np', 'ralph'): 0.00064300411522263374,
('np', 'waldo'): 6.430041152263375e-05,
('np$', "emerson's"): 0.0008688097306689834,
('at-tl', 'the'): 0.9066666666666666,
('nn-tl', 'address'): 8.311170212765957e-05,
('vbg', 'naming'): 0.00018540263271738457,
('nn', 'philosophy'): 0.0005542105416678821,
('abx', 'both'): 1.0,
('nns', 'cases'): 0.002600858888284038,
('jj', 'stubborn'): 0.00020857594771696243,
('nn', 'fact'): 0.0029679432955108946,
('vbz', 'remains'): 0.010055530541797988,
(':', ':'): 1.0,
('vbd', 'gave'): 0.010783731001103846,
('nn', 'birth'): 0.00028439751480325524,
('cd', 'two'): 0.10319227913882702,
('jj', 'brilliant'): 0.0008343037908678497,
('nns', 'apostates'): 2.0161696808403396e-05,
('jj', 'legitimate'): 0.00022595727669337597,
('nns', 'offspring'): 4.032339361680679e-05,
('nns', 'loins'): 4.032339361680679e-05,
('vbn', 'brought'): 0.004154755098151324,
('nn', 'test'): 0.0006198407373917102,
('vbd', 'behaved'): 0.00038210070476352214,
('rb', 'shabbily'): 6.081430352418889e-05,
```

```
('vb', 'suppose'): 0.0028447619992722703,
('vbn', 'ventured'): 0.00011435105774728416,
('nns', 'realms'): 4.032339361680679e-05,
('nns', 'colleagues'): 0.00044355732978487467,
('vbd', 'thought'): 0.013288613398998048,
('jj', 'infidel'): 1.7381328976413536e-05,
('nn', 'way'): 0.005906717615144532,
('vb', 'settle'): 0.0006284939300717806,
('jj', 'frank'): 0.0002781012636226166,
('nns', 'differences'): 0.0014516421702050445,
('nn', 'opinion'): 0.0006490097132689671,
('.', '?'): 0.07554781333088842,
('jj', 'possible'): 0.00570107590426364,
('wrb', 'how'): 0.14461690071446168,
('nns', 'liberals'): 0.0003830722393596645,
('vb', 'trumpet'): 3.307862789851477e-05,
('nn', 'confidence'): 0.00035731995449639765,
('jj', 'human'): 0.004015086993551527,
('nn', 'dignity'): 0.00024064405098736983,
('vbn', 'exposed'): 0.0010672765389746523,
('nn', 'contagion'): 1.4584487938628475e-05,
('nn', 'fear'): 0.0006271329813610244,
('ql', 'more'): 0.11550693296018319,
('jj', 'insidious'): 1.7381328976413536e-05,
('nn', 'conservative'): 2.1876731907942712e-05,
('rb', 'ever'): 0.009182959832152521,
('vb', 'worry'): 0.0013231451159405909,
('in', 'about'): 0.010141520306089521,
('rb', 'however'): 0.011007388937878188,
('ex', 'there'): 0.9989837398373984,
('jj', 'crucial'): 0.0004692958823631655,
('nns', 'histories'): 0.00016129357446722717,
('np', 'emerson'): 0.0002250514403292181,
('vbd', 'evaded'): 4.245563386261357e-05,
('nn', 'problem'): 0.0020491205553773008,
('vbg', 'shoving'): 0.00030900438786230763,
('rb', 'aside'): 0.0017636148022014777,
('vbg', 'leaving'): 0.005005871083369384,
('in', 'behind'): 0.0019545475499008896,
('vbd', 'walked'): 0.005561688036002378,
('nn', 'communion'): 7.292243969314237e-05,
('cs', 'so'): 0.021244538194967605,
('vb', 'lick'): 9.92358836955443e-05,
('nn', 'wound'): 0.00010209141557039932,
('nn', 'departure'): 0.0001166759035090278,
('vb', 'preserve'): 0.0009262015811584136,
('nn', 'self-respect'): 2.916897587725695e-05,
```

```
('rb', 'eventually'): 0.0014595432845805334,
('vb', 'accord'): 6.615725579702954e-05,
('jj', 'pious'): 0.00015643196078772182,
('nn', 'veneration'): 2.1876731907942712e-05,
('rb', 'hence'): 0.0014899504363426277,
('nn', 'lifetime'): 5.104570778519966e-05,
('nn', 'memory'): 0.00045941137006679693,
('rb', 'afterwards'): 0.00039529297290722776,
('nn', 'canker'): 7.292243969314237e-06,
('in', 'within'): 0.002793527866131932,
('jj', 'liberal'): 0.0007647784749621956,
('nn', 'sensitivity'): 0.00019689058717148441,
('vbz', 'points'): 0.0028515683625994297,
('vbg', 'accusing'): 0.0004326061430072307,
('nn', 'finger'): 0.00027710527083394103,
('in', 'at'): 0.0447333324113769415,
('vbg', 'telling'): 0.0023484333477535382,
('cc', 'neither'): 0.002618879472693032,
('nn', 'courage'): 0.00021876731907942713,
('vb', 'support'): 0.0014223809996361352,
('nn', 'energy'): 0.0006490097132689671,
('vb', 'cut'): 0.0026132116039826667,
('nn', 'throat'): 0.00026252078289531257,
('nn', 'dispute'): 0.00020418283114079865,
('nn', 'society'): 0.001319896158445877,
('nn', 'time'): 0.010384155412303473,
('jj', 'ecclesiastical'): 0.00010428797385848122,
('jj', 'social'): 0.0055620252724523315,
('jj', 'real'): 0.0037022230719760832,
('cd', 'one'): 0.19995050730017322,
('jj', 'bitter'): 0.0008169224618914362,
('md*', 'cannot'): 0.3023255813953488,
('vbn', 'smoothed'): 0.00011435105774728416,
('rb', 'now'): 0.036154103445130294,
('vbg', 'cherishing'): 0.00012360175514492306,
('nns', 'sarcasms'): 2.0161696808403396e-05,
('jj', 'delightful'): 0.0003650079085046843,
('nns', 'bits'): 0.00020161696808403395,
('nn', 'self-deprecation'): 7.292243969314237e-06,
('rb', 'solemnly'): 0.00027366436585885,
('vbg', 'calling'): 0.002472035102898461,
('nn', 'reconsideration'): 2.916897587725695e-05,
('nn', 'justice'): 0.0005323338097599393,
('nns', 'objections'): 0.00024194036170084075,
('jj', 'incontestable'): 1.7381328976413536e-05,
('dt', 'that'): 0.25200444060688293,
('nn', 'world'): 0.004535775748913455,
```

('np', 'boston'): 0.0014146090534979425,
('jj', 'narrow-minded'): 1.7381328976413536e-05,
('rb', 'intellectually'): 0.0001520357588104722,
('jj', 'sterile'): 0.00015643196078772182,
('jj', 'smug'): 8.690664488206768e-05,
('jj', 'afraid'): 0.0008343037908678497,
('jj', 'logical'): 0.0004692958823631655,
('nns', 'consequences'): 0.0005846892074436985,
('jj', 'own'): 0.01174977838805555,
('jj', 'mild'): 0.0001911946187405489,
('nns', 'ventures'): 8.064678723361358e-05,
('nn', 'iconoclasm'): 7.292243969314237e-06,
('ql', 'quite'): 0.02506042488233049,
('vbn', 'prepared'): 0.0027825424051839145,
('vb', 'resort'): 0.0001984717673910886,
('jj', 'hysterical'): 0.00015643196078772182,
('nns', 'repressions'): 2.0161696808403396e-05,
('jj', 'brittle'): 5.214398692924061e-05,
('nns', 'foundations'): 0.0002621020585092441,
('vbn', 'threatened'): 0.0005336382694873261,
('rb', 'along'): 0.0041135372639644844,
('np', 'garrison'): 6.430041152263375e-05,
('np', 'charles'): 0.002539866255144033,
('np', 'sumner'): 6.430041152263375e-05,
('vbd', 'showed'): 0.005179587331238855,
('jj', 'magnificent'): 0.0004345332244103384,
('jj', 'moral'): 0.002137903464098865,
('nn', 'bravery'): 2.916897587725695e-05,
('vbg', 'facing'): 0.0019158272047463073,
('nns', 'mobs'): 8.064678723361358e-05,
('vbn', 'mobilized'): 0.00011435105774728416,
('nn', 'defense'): 0.0008531925444097658,
('jj-tl', 'mexican'): 0.0005420054200542005,
('nn-tl', 'war'): 0.01171875,
('nn', 'slavery'): 0.00019689058717148441,
('rb', 'nevertheless'): 0.0019460577127740444,
('vb', 'find'): 0.011643677020277198,
('nns', 'reasons'): 0.0018548761063731125,
('vbg', 'respecting'): 0.00018540263271738457,
('nn', 'bigotry'): 1.4584487938628475e-05,
('nn', 'populace'): 2.916897587725695e-05,
('nns', 'passions'): 0.00024194036170084075,
('jj', 'genuine'): 0.0005214398692924061,
('nn', 'division'): 0.00041565790625091155,
('nns', 'abolitionists'): 6.048509042521019e-05,
('jj', 'clear-cut'): 0.00010428797385848122,
('jj', 'ultra-liberal'): 1.7381328976413536e-05,

```
('nn', 'minister'): 0.00028439751480325524,
('nn', 'pale'): 7.292243969314237e-06,
('nn', 'church'): 0.0014876177697401044,
('vbn', 'proclaimed'): 0.00011435105774728416,
('nn', 'repository'): 2.1876731907942712e-05,
('vbz', 'poses'): 0.00015008254539996998,
('jj', 'different'): 0.0047451028105608955,
('vbn', 'resolved'): 0.0006479893272346103,
('vbg', 'holding'): 0.0033372473889129225,
('rp', 'up'): 0.2839186691312384,
('nn', 'champion'): 0.00013126039144765629,
('nn', 'freedom'): 0.0008021468366245661,
('jj', 'theological'): 0.00041715189543392486,
('nns', 'theses'): 2.0161696808403396e-05,
('vbn', 'become'): 0.00411663807890223,
('nns', 'commonplaces'): 4.032339361680679e-05,
('jj', 'fundamental'): 0.0007473971459857821,
('nn', 'interrogation'): 1.4584487938628475e-05,
('vbd', 'phrased'): 4.245563386261357e-05,
('ql', 'very'): 0.08065131662638342,
('rb', 'much'): 0.0058685802900842274,
('rb', 'endlessly'): 0.0002128500623346611,
('vbn', 'rephrased'): 3.8117019249094724e-05,
('ppss', 'i'): 0.3718414466779688,
('rb', 'here'): 0.020555234591175843,
('vb', 'put'): 0.005788759882240085,
('rb', 'thus'): 0.007875452306382461,
('nn', 'point'): 0.0022460111425487853,
('do', 'do'): 0.9975349219391948,
('jj', 'tolerant'): 0.00013905063118113083,
('ppls', 'themselves'): 0.7731629392971247,
('vbn', 'obliged'): 0.0005717552887364208,
('vb', 'become'): 0.006946511858688101,
('jj', 'intolerant'): 1.7381328976413536e-05,
('jj', 'aware'): 0.0013731249891366694,
('vbn', 'reached'): 0.0021345530779493045,
('nn', 'end'): 0.00239185602193507,
('nn', 'patience'): 0.00013126039144765629,
('nn', 'dismay'): 2.916897587725695e-05,
('vb', 'learn'): 0.0026462902318811817,
('od', 'first'): 0.5092165898617511,
('nn', 'doubt'): 0.0006417174692996529,
('nn', 'era'): 0.00018230609923285593,
('ql', 'exquisitely'): 0.00012721027859051011,
('jj', 'cruel'): 0.00020857594771696243,
('vbg', 'enforcing'): 0.000247203510289461,
('nns', 'canons'): 0.00010080848404201697,
```

('nn', 'behavior'): 0.0006125484934223959,
('jj', 'gentle'): 0.00041715189543392486,
('np', 'channing'): 0.00016075102880658436,
('vbn', 'revered'): 0.0001905850962454736,
('nps', 'bostonians'): 0.0008771929824561404,
('jj', 'orthodox'): 0.0001911946187405489,
('vbd', 'wrote'): 0.006410800713254649,
('nn', 'friend'): 0.0008094390805938803,
('np', 'louisville'): 0.0001286008230452675,
('in', 'among'): 0.003079334347485364,
('ap', 'many'): 0.10528144544822794,
('nns', 'virtues'): 0.00030242545212605095,
('vb', 'abound'): 3.307862789851477e-05,
('nn', 'spirit'): 0.0011011288393664499,
('nn', 'yoke'): 2.1876731907942712e-05,
('vbd', 'crushed'): 8.491126772522713e-05,
('nn', 'individuality'): 2.916897587725695e-05,
('nn', 'judgment'): 0.00038648893037365457,
('nn', 'action'): 0.001874106700113759,
('nn', 'city'): 0.0017063850888195315,
('vbn', 'governed'): 0.00045740423098913663,
('ql', 'so'): 0.108000152652334309,
('rb', 'little'): 9.122145528628334e-05,
('nn', 'police'): 0.00048128810197473965,
('jj', 'mutual'): 0.0003476265795282707,
('nns', 'inspections'): 4.032339361680679e-05,
('vbn', 'called'): 0.007928340003811702,
('jj', 'public'): 0.004345332244103385,
('nn', 'sentiment'): 0.00014584487938628475,
('vb', 'stand'): 0.003175548278257418,
('rbr', 'more'): 0.1671415004748338,
('nn', 'awe'): 3.646121984657119e-05,
('dt', 'another'): 0.07524361662760577,
('ap', 'most'): 0.051308779244845955,
('nns', 'people'): 0.014939817335026915,
('ql', 'less'): 0.02086248568884366,
('jj', 'individual'): 0.0022595727669337596,
('vbz', 'runs'): 0.0019510730901996097,
('nns', 'masses'): 0.0003830722393596645,
('rb', 'often'): 0.010095174385015355,
('vbz', 'rules'): 0.00030016509079993996,
('nn', 'rod'): 2.916897587725695e-05,
('nn', 'iron'): 0.00021147507511011288,
('ql', 'even'): 0.013738710087775092,
('rb', 'poignantly'): 3.0407151762094443e-05,
('nn', 'insight'): 0.00014584487938628475,
('nn', 'genius'): 0.00014584487938628475,

```
('vbd', 'added'): 0.003141716905833404,
('vb', 'remember'): 0.004101749859415831,
('.', '!'): 0.02636838727422756,
('nn', 'letter'): 0.0009406994720415366,
('vb', 'think'): 0.012536799973537098,
('vb', 'say'): 0.014620753531143529,
('vbd', 'reached'): 0.004203107752398743,
('jj', 'such'): 0.015347713486173153,
('nn', 'candor'): 1.4584487938628475e-05,
('nns', 'sermons'): 4.032339361680679e-05,
('vbg', 'commencing'): 0.0004944070205796922,
('nn', 'mission'): 0.00043753463815885427,
('nn', 'world-at-large'): 7.292243969314237e-06,
('vbn', 'disguised'): 0.0003049361539927578,
('od', 'twenty-eighth'): 0.000576036866359447,
('jj-tl', 'congregational'): 0.0013550135501355014,
('vbd', 'bore'): 0.0005094676063513628,
('nn', 'resemblance'): 7.292243969314237e-05,
('nns', 'polities'): 2.0161696808403396e-05,
('vbn', 'descended'): 0.0001524680769963789,
('nns', 'founders'): 4.032339361680679e-05,
('nns', 'churches'): 0.0015524506542470614,
('vbd', 'made'): 0.01749172115139679,
('jj', 'explicit'): 0.0003650079085046843,
('nn', 'beginning'): 0.0005396260537292535,
('nn', 'conflict'): 0.00033544322258845493,
('jj', 'hunkerish'): 1.7381328976413536e-05,
('pn', 'something'): 0.17383669885864794,
('vbn', 'evaporated'): 3.8117019249094724e-05,
('jj', 'genteel'): 6.952531590565414e-05,
('jj', 'clerical'): 0.00015643196078772182,
('nn', 'decorum'): 1.4584487938628475e-05,
('cs', 'because'): 0.03244437747978504,
('vbd', 'spoke'): 0.0033115394412838585,
('rb', 'openly'): 0.0009122145528628334,
('vbn', 'prophesied'): 0.00011435105774728416,
('pn', 'someone'): 0.033801580333625986,
('md', 'might'): 0.05250425589104919,
('vbg', 'daring'): 0.0005562078981521537,
('nn', 'hyperbole'): 7.292243969314237e-06,
('vbd', 'vindicated'): 4.245563386261357e-05,
('np$', "channing's"): 0.0004344048653344917,
('jjr', 'further'): 0.03666102650874224,
('nn', 'prophecy'): 3.646121984657119e-05,
('vbd', 'committed'): 0.00016982253545045427,
('nn', 'infraction'): 7.292243969314237e-06,
('nn', 'taste'): 0.0003500277105270834,
```

```
('rb', 'promptly'): 0.0006689573387660778,
('vb', 'discover'): 0.001290066488042076,
('wql', 'how'): 0.8076923076923077,
('ap', 'little'): 0.042274727820245554,
('nn', 'mercy'): 0.00010209141557039932,
('vbn', 'disposed'): 0.0004955212502382314,
('vb', 'allow'): 0.00215011081340346,
('nns', 'libertarians'): 4.032339361680679e-05,
('vbd', 'appeared'): 0.004372930287849197,
('nns', 'libertines'): 4.032339361680679e-05,
('vbn', 'institutionalized'): 0.00011435105774728416,
('vbd', 'proved'): 0.0019529591576802241,
('rb', 'fundamentally'): 0.0002128500623346611,
('nn', 'institution'): 0.00024793629495668406,
('rb', 'only'): 0.03685346793565847,
('vbn', 'defined'): 0.0012959786544692205,
('nns', 'limits'): 0.0006451742978689087,
('nn', 'license'): 0.00024793629495668406,
('vbg', 'reminding'): 0.00030900438786230763,
('nns', 'factors'): 0.0018750378031815158,
('nn', 'situation'): 0.0012688504506606774,
('bem', 'am'): 1.0,
('jj', 'sure'): 0.0034936471242591207,
('jj', 'fresh'): 0.001199311699372534,
('nn', 'realization'): 0.00015313712335559898,
('wql', 'however'): 0.19230769230769232,
('jj', 'painful'): 0.0003302452505518572,
('nn', 'battle'): 0.0004667036140361112,
('nns', 'neighbors'): 0.0007056593882941188,
('vbn', 'fought'): 0.0008004574042309891,
('jj', 'earnest'): 0.00026071993464620307,
('vbd', 'arraigned'): 8.491126772522713e-05,
('nns', 'citizens'): 0.0014919655638218512,
('nn', 'language'): 0.0007000554210541668,
('nn', 'courtesy'): 4.3753463815885424e-05,
('vb', 'respond'): 0.0005954153021732659,
('ap', 'least'): 0.0339356034283067,
('nn', 'resentment'): 0.00012396814747834203,
('rb', 'otherwise'): 0.002402164989205461,
('nn', 'lawyer'): 0.00029898200274188375,
('nn', 'doctor'): 0.0005396260537292535,
('nn', 'science'): 0.000736516640900738,
('nns', 'letters'): 0.001996007984031936,
('vbn', 'told'): 0.004383457213645893,
('nns', 'cherubim'): 2.0161696808403396e-05,
('nns', 'seraphim'): 2.0161696808403396e-05,
('cd', 'three'): 0.044378454178008744,
```

```
('nns', 'archangels'): 2.0161696808403396e-05,
('vbd', 'stood'): 0.007302369024369534,
('in', 'before'): 0.0033374821370949154,
('jj', 'golden'): 0.00039977056645751135,
('nn', 'throne'): 1.4584487938628475e-05,
('nn', 'merchant'): 7.292243969314237e-05,
('rb', 'continually'): 0.0006993644905281722,
('vbd', 'cried'): 0.0010189352127027256,
("'", "'"): 1.0,
('jj', 'holy'): 0.00026071993464620307,
('jj-tl', 'almighty'): 0.0005420054200542005,
('nn-tl', 'dollar'): 0.00016622340425531914,
('vb', 'recollect'): 3.307862789851477e-05,
('jj', 'sensitive'): 0.0009559730937027445,
('nns', 'emotions'): 0.0007459827819109256,
('jj', 'old'): 0.009299011002381242,
('jj-tl', 'puritan'): 0.001084010840108401,
('nn', 'stock'): 0.0009115304961642797,
('jj', 'recent'): 0.0027810126362261657,
('nns', 'tides'): 8.064678723361358e-05,
('nn', 'immigration'): 5.104570778519966e-05,
('vbn', 'astonished'): 0.0001905850962454736,
('jj', 'thin'): 0.0013905063181130829,
('nns', 'lips'): 0.0012903485957378173,
('vbn', 'compressed'): 0.00022870211549456832,
('jj', 'white'): 0.00399770566645751135,
('nn', 'line'): 0.0018886911880523875,
('nn', 'rage'): 8.750692763177085e-05,
('vbd', 'snarled'): 0.000297189437038295,
('vb', 'talk'): 0.0032417055340544475,
('nps', 'catholics'): 0.02543859649122807,
('vbg', 'voting'): 0.0013596193065941536,
('nn', 'bishop'): 3.646121984657119e-05,
('vbz', 'tells'): 0.00435239381659913,
('vb', 'reproach'): 3.307862789851477e-05,
('ppss', 'you'): 0.1686445467021878,
('ap', 'same'): 0.07180912670836229,
('ql', 'great'): 0.0016537336216766315,
('nns', 'bishops'): 0.00010080848404201697,
('hvn', 'had'): 0.9953917050691244,
('nn', 'cross'): 0.0001166759035090278,
('nn', 'bosom'): 5.833379517545139e-05,
('nn', 'mitre'): 7.292243969314237e-06,
('nn', 'head'): 0.0026543768048303824,
('vb', 'appeal'): 0.00026462902318811815,
('nn', 'authority'): 0.0005323338097599393,
('nn-tl', 'pope'): 0.0018284574468085107,
```

```
('np', 'rome'): 0.0018647119341563786,
('nn', 'pope'): 3.646121984657119e-05,
('ql', 'much'): 0.026205317389645082,
('in', 'nearer'): 1.8439127829253676e-05,
('nr', 'home'): 0.1918194640338505,
('vbn', 'controlled'): 0.0012578616352201257,
('ap', 'few'): 0.0620801482511003,
('nns', 'capitalists'): 4.032339361680679e-05,
('nns', 'lawyers'): 0.0004233956329764713,
('ap', 'other'): 0.17674310864025944,
('nns', 'managers'): 0.00044355732978487467,
('vbd', 'told'): 0.010911097902691687,
('nns', 'editors'): 0.00030242545212605095,
('nns', 'preachers'): 4.032339361680679e-05,
('nn', 'war'): 0.002027243823469358,
('vbd', 'meant'): 0.0025473380317568143,
('nn', 'business'): 0.00234081031414987,
('vbn', 'repeated'): 0.0013722126929674098,
('nn', 'care'): 0.000583379517545139,
('vb', 'know'): 0.01981409811121035,
('vbd', 'intended'): 0.0004245563386261357,
('vbn', 'caught'): 0.0014484467314655994,
('nn', 'malaria'): 2.1876731907942712e-05,
('jjr', 'worse'): 0.024816694867456288,
('vbd', 'deceived'): 4.245563386261357e-05,
('vbd', 'implied'): 0.00021227816931306784,
('jj', 'bad'): 0.0020857594771696245,
('qlp', 'indeed'): 0.04741379310344827,
('nn', 'duty'): 0.0003719044424350261,
('vbd', 'said'): 0.06640061136112763,
('nns', 'parents'): 0.0016129357446722716,
('vbn', 'given'): 0.013074137602439488,
('nn', 'rule'): 0.00036461219846571185,
('in', 'beyond'): 0.0013921541511086525,
('nn', 'love'): 0.0010646676195198786,
('wps', 'that'): 0.4237960339943343,
('vbd', 'suffused'): 0.00016982253545045427,
('nn', 'being'): 0.0001750138552635417,
('nn', 'sense'): 0.0019834903596534724,
('nn', 'humor'): 0.00029898200274188375,
('rb', 'largely'): 0.0015203575881047222,
('vbn', 'supplied'): 0.0008766914427291786,
('vb', 'perform'): 0.0007608084416658397,
('vbd', 'cost'): 0.00038210070476352214,
('jj', 'acute'): 0.00020857594771696243,
('nn', 'pain'): 0.000583379517545139,
('vbd', 'exhausted'): 0.0001273669015878407,
```

```
('nn', 'age'): 0.001443864305924219,
('cd', 'fifty'): 0.00544419698094531,
('vb', 'weep'): 0.0003307862789851477,
('vbd', 'wept'): 0.00021227816931306784,
('ql', 'astonishingly'): 0.0005088411143620405,
('jjt', 'slightest'): 0.013303769401330377,
('nn', 'provocation'): 3.646121984657119e-05,
('nn', 'psychology'): 9.479917160108509e-05,
('rb', 'entirely'): 0.001581171891628911,
('jj', 'compatible'): 0.00020857594771696243,
('jj', 'remorseless'): 1.7381328976413536e-05,
('nn', 'readiness'): 0.00010209141557039932,
('vb', 'massacre'): 3.307862789851477e-05,
('nns', 'opponents'): 0.0002621020585092441,
('cs', 'if'): 0.0985887198031239,
('ppo', 'me'): 0.103674932775661995,
('nn', 'pleasure'): 0.0003719044424350261,
('jj', 'hard'): 0.0020857594771696245,
('nns', 'things'): 0.00647190467549749,
('vb', 'shut'): 0.0004631007905792068,
('rb', 'forever'): 0.0010034360081491167,
('nn', 'vein'): 0.00016042936732491324,
('vb', 'continue'): 0.003175548278257418,
('nn', 'truth'): 0.0008240235685325088,
('md', 'must'): 0.08010034943105457,
('vb', 'speak'): 0.003208626906155933,
('jjr', 'hotter'): 0.003948110547095319,
('nn', 'fire'): 0.0010865443514278214,
('rb', 'copiously'): 3.0407151762094443e-05,
('vbd', 'shed'): 0.0001273669015878407,
('vbd', 'yielded'): 0.000297189437038295,
('vbg', 'living'): 0.008590321982572152,
('nn', 'sacrifice'): 0.00013126039144765629,
('vbn', 'impersonalized'): 3.8117019249094724e-05,
('vbn', 'disturbed'): 0.0007242233657327997,
('nn', 'havoc'): 2.1876731907942712e-05,
('vbd', 'worked'): 0.0030992612719707904,
('nns$', "people's"): 0.05726872246696035,
('nns', 'consciences'): 2.0161696808403396e-05,
('nn', 'endeavor'): 3.646121984657119e-05,
('vb', 'capture'): 0.0003969435347821772,
('jj', 'faint'): 0.00041715189543392486,
('jj', 'strenuous'): 0.00017381328976413536,
('nn', 'fight'): 0.00037919668640434037,
('vbn', 'muffled'): 0.0003811701924909472,
('nn', 'indifference'): 0.00010209141557039932,
('ap', 'very'): 0.009497336113041464,
```

```
('cd', '1848'): 0.00016497566608925183,
('vbd', 'seemed'): 0.011845121847669185,
('jj', 'central'): 0.001616463594806459,
('nn', 'hope'): 0.00061254849342239594,
('nn', 'survival'): 0.0001750138552635417,
('jj', 'literal'): 0.00020857594771696243,
('jj', 'factual'): 0.00012166930283489475,
('nn', 'historicity'): 7.292243969314237e-06,
('nns', 'miracles'): 0.00016129357446722717,
('vbn', 'reported'): 0.0023632551934438728,
('cd-tl', 'four'): 0.011335012594458438,
('nps', 'gospels'): 0.0035087719298245615,
('jj', 'idle'): 0.00017381328976413536,
('vb', 'ask'): 0.0038040422083291986,
('wrb', 'why'): 0.08376447400837644,
('ql', 'no'): 0.015519653988042234,
('rbr', 'longer'): 0.10731244064577398,
('pn', 'somebody'): 0.019754170324846356,
('vbg', 'professing'): 0.00012360175514492306,
('jjt', 'deepest'): 0.013303769401330377,
('nn', 'piety'): 2.916897587725695e-05,
('vbz', 'decides'): 0.0018009905447996398,
('rb', 'anew'): 0.00012162860704837777,
('nn', 'importance'): 0.0007438088848700522,
('cs', 'whether'): 0.012957661594093717,
('np', 'christ'): 0.0024434156378600825,
('vbd', 'transformed'): 4.245563386261357e-05,
('nn', 'water'): 0.00282209841612461,
('nn', 'wine'): 0.00044482688212816847,
('cd', 'eleven'): 0.003052049822651159,
('rb', 'a.m.'): 0.0011250646151974943,
('od', 'third'): 0.09101382488479262,
('np', 'august'): 0.0014789094650205761,
('rb', 'a.d.'): 0.00027366436585885,
('cd', '32'): 0.0008248783304462591,
('nn', 'answer'): 0.0007219321529621094,
('in', 'as'): 0.0010418107223528327,
('vbn', 'alarmed'): 0.0003049361539927578,
('vb', 'give'): 0.011676755648175713,
('nn', 'credit'): 0.0003719044424350261,
('vbn', 'taken'): 0.009224318658280923,
('jj', 'right'): 0.0034067404793770534,
('nn', 'side'): 0.0024720070055975264,
('jj', 'unnecessary'): 0.00024333860566978955,
('nn', 'controversy'): 0.00018230609923285593,
('vb', 'salute'): 3.307862789851477e-05,
('vb', 'pass'): 0.0018193245344183123,
```

```
('rp', 'on'): 0.0933456561922366,
('vbg', 'forgetting'): 0.0004326061430072307,
('jj', 'entire'): 0.0023812420697686544,
('nn', 'episode'): 8.021468366245662e-05,
('nn', 'leisure'): 8.021468366245662e-05,
('nn', 'skill'): 0.00024793629495668406,
('vb', 'comprehend'): 0.00016539313949257386,
('vbg', 'working'): 0.00846672022742723,
('nn', 'mind'): 0.00191056791996033,
('nn', 'heart'): 0.001042790887611936,
('nn', 'graduate'): 0.0001166759035090278,
('vb', 'muster'): 6.615725579702954e-05,
('nn', 'audacity'): 2.1876731907942712e-05,
('vb', 'contradict'): 0.000132231451159405908,
('ql', 'most'): 0.07327312046813382,
('jj', 'formidable'): 0.00026071993464620307,
('nn', 'instructor'): 5.83379517545139e-05,
('jj', 'majesterial'): 1.7381328976413536e-05,
('np', 'andrews'): 6.430041152263375e-05,
('np', 'norton'): 0.000257201646090535,
('np', 'jesus'): 0.0015753600823045267,
('cs', 'like'): 0.04565315654663252,
('jj', 'religious'): 0.002502911372603549,
('nns', 'teachers'): 0.0012298635053126072,
('vb', 'see'): 0.022923489133670736,
('nn', 'miracle'): 9.479917160108509e-05,
('vbz', 'proves'): 0.0024013207263995197,
('nn', 'doctrine'): 0.00026981300268646268,
('nn', 'part'): 0.003135664906805122,
('nn', 'missionary'): 6.563019572382814e-05,
('nn', 'movement'): 0.0008094390805938803,
('vbn', 'looked'): 0.0013340956737183153,
('rb', 'upon'): 0.0011858789187216834,
('jj', 'deficient'): 5.214398692924061e-05,
('jj', 'essential'): 0.001268837015278188,
('nn', 'life'): 0.004448268821281685,
('nn', 'education'): 0.0011011288393664499,
('nns', 'children'): 0.006129155829754632,
('rb', 'too'): 0.007723416547571989,
('rb', 'once'): 0.01122023900021285,
('rb', 'hardly'): 0.0025542007480159334,
('nn', 'sideshow'): 7.292243969314237e-06,
('nn', 'congregation'): 0.00029898200274188375,
('vb', 'assume'): 0.0019516390460123715,
('jj', 'full'): 0.003528409782211948,
('nn', 'oversight'): 7.292243969314237e-06,
('nn', 'school'): 0.0026835457807076394,
```

```
('vbn', 'thought'): 0.0024776062511911567,
('vbg', 'failing'): 0.0008652122860144614,
('jj', 'serious'): 0.0017033702396885267,
('nn', 'weakness'): 0.0002916897587725695,
('jj', 'ecumenical'): 0.00039977056645751135,
('rb', 'generally'): 0.0036488582114513335,
('vbn', 'regarded'): 0.0014865637507146942,
('nn', 'responsibility'): 0.0007000554210541668,
('jj', 'national'): 0.0034415031373298804,
('nns', 'leaders'): 0.0019153611967983226,
('dt', 'each'): 0.09929690391020106,
('jj', 'interdenominational'): 3.476265795282707e-05,
('nns', 'executives'): 0.00014113187765882378,
('vb', 'even'): 0.0007277298137673249,
('vb', 'dismiss'): 0.00016539313949257386,
('vbz', 'concerns'): 0.0018009905447996398,
('np-tl', 'rover'): 0.0002756339581036384,
('nns-tl', 'boys'): 0.0034482758620689655,
('vbn', 'dubbed'): 7.623403849818945e-05,
('vb', 'like'): 0.006417253812311866,
('vb', 'go'): 0.018424795739472727,
('jj', 'international'): 0.0018945648584290755,
('nns', 'assemblies'): 0.00012097018085042038,
('nn', 'expense'): 0.0003135664906805122,
('nns', 'accounts'): 0.0005242041170184882,
('vb', 'permit'): 0.002017796301809401,
('ql', 'as'): 0.12670143747614807,
('rb', 'long'): 0.0052908444066044336,
('nn', 'view'): 0.001079252107458507,
('vbz', 'prevails'): 0.0009004952723998199,
('md', 'will'): 0.1705940327927605,
('jj', 'lame'): 3.476265795282707e-05,
('vb', 'halt'): 0.00019847176739108886,
('ap', 'next'): 0.030692610609219365,
('nn', 'stage'): 0.0011302978152437067,
('rb', 'ahead'): 0.002888679417398972,
('vbg', 'making'): 0.012669179902354614,
('rb', 'thoroughly'): 0.0007601787940523611,
('jj', 'local'): 0.004432238888985452,
('nn', 'community'): 0.0014584487938628474,
('nn', 'progress'): 0.0007000554210541668,
('vb', 'take'): 0.018358638483675697,
('nn', 'place'): 0.0030919114429892366,
('ql', 'far'): 0.014756392316499172,
('in', 'through'): 0.007357212003872217,
('nn', 'summit'): 7.292243969314237e-05,
('nn', 'conference'): 0.00042295015022022575,
```

```
       ('jj-tl', 'national'): 0.03929539295392954,
       ('nn-tl', 'council'): 0.005568484042553191,
       ('nn-tl', 'world'): 0.007313829787234043,
       ('nns', 'offices'): 0.0006250126010605053,
       ('jj', 'denominational'): 0.00012166930283489475,
       ('nns', 'boards'): 0.0008266295691445392,
       ('vbz', 'happens'): 0.005402971634398919,
       ('nns', 'communities'): 0.0006653359946773121,
       ('wrb', 'where'): 0.20768662232076865,
       ('vb', 'live'): 0.004763322417386127,
       ('rb', 'together'): 0.0074193450299951044,
       ('jj', 'front'): 0.00118193037039661205,
       ('nn', 'advance'): 0.00026252078289531257,
       ('vbg', 'witnessing'): 0.00037080526543476915,
       ('vbg', 'worshiping'): 0.00018540263271738457,
       ('nns', 'traditions'): 0.0003830722393596645,
       ('cs', 'until'): 0.015418612827080508,
       ('nns', 'terms'): 0.0028427992499848787,
       ('vbz', 'makes'): 0.022662464355395467,
       …})
```

0.5 Test the model

will pick few sentences from test set and predict parts of speech tags using Viterbi Algorithm

```python
[76]: from prettytable import PrettyTable
```

```python
[80]: table = PrettyTable()
      table.field_names = ['word','pred_tag','exp_tag']
      sent, expected_tags = zip(*test[1])
      for word,pred_tag,exp_tag in zip(sent,pos_tagger.viterbi(sent),expected_tags):
          table.add_row([word,pred_tag,exp_tag])
      table
```

```
[80]: +--------------+----------+---------+
      |     word     | pred_tag | exp_tag |
      +--------------+----------+---------+
      | confucianism |    np    |    np   |
      |     had      |   hvd    |   hvd   |
      |     its      |   pp$    |   pp$   |
      |     own      |    jj    |    jj   |
      |    magic     |    nn    |    nn   |
      |      in      |    in    |    in   |
      |     the      |    at    |    at   |
      |     idea     |    nn    |    nn   |
      |     that     |    cs    |    cs   |
      |    virtue    |    nn    |    nn   |
```

```
|    had    |   hvd   |   hvd   |
|   power   |   nn    |   nn    |
|    .      |   .     |   .     |
+-----------+---------+---------+
```

[81]:
```python
table = PrettyTable()
table.field_names = ['word','pred_tag','exp_tag']
sent, expected_tags = zip(*test[2])
for word,pred_tag,exp_tag in zip(sent,pos_tagger.viterbi(sent),expected_tags):
    table.add_row([word,pred_tag,exp_tag])
table
```

[81]:
```
+-----------+----------+---------+
|    word   | pred_tag | exp_tag |
+-----------+----------+---------+
|     ``    |    ``    |    ``   |
|     a     |    at    |    at   |
| stake-out |    nn    |    nn   |
|     ,     |    ,     |    ,    |
|     i     |   ppss   |   ppss  |
|  suppose  |    vb    |    vb   |
|     .     |    .     |    .    |
+-----------+----------+---------+
```

[82]:
```python
table = PrettyTable()
table.field_names = ['word','pred_tag','exp_tag']
sent, expected_tags = zip(*test[3])
for word,pred_tag,exp_tag in zip(sent,pos_tagger.viterbi(sent),expected_tags):
    table.add_row([word,pred_tag,exp_tag])
table
```

[82]:
```
+-------------+----------+---------+
|     word    | pred_tag | exp_tag |
+-------------+----------+---------+
|      ``     |    ``    |    ``   |
|     they    |   ppss   |   ppss  |
|    won't    |   md*    |   md*   |
|     talk    |    vb    |    vb   |
|    about    |    rb    |    in   |
|     who     |   wps    |   wps   |
|     gave    |   vbd    |   vbd   |
|     the     |    at    |    at   |
| information |    nn    |    nn   |
|      .      |    .     |    .    |
+-------------+----------+---------+
```

```
[83]: table = PrettyTable()
      table.field_names = ['word','pred_tag','exp_tag']
      sent, expected_tags = zip(*test[4])
      for word,pred_tag,exp_tag in zip(sent,pos_tagger.viterbi(sent),expected_tags):
          table.add_row([word,pred_tag,exp_tag])
      table
```

```
[83]: +-------------+----------+---------+
      |     word    | pred_tag | exp_tag |
      +-------------+----------+---------+
      |    these    |    dts   |   dts   |
      |    curves   |    nns   |   nns   |
      |     were    |    bed   |   bed   |
      |    derived  |    vbn   |   vbn   |
      |      by     |    in    |    in   |
      |      an     |    at    |    at   |
      |   analysis  |    nn    |    nn   |
      |      of     |    in    |    in   |
      |  extensive  |    jj    |    jj   |
      |   skywave   |    nn    |    nn   |
      | measurement |    nn    |    nn   |
      |     data    |    nns   |    nn   |
      |      .      |     .    |    .    |
      +-------------+----------+---------+
```

as we can observe almost all of the tags in a sentence are being predicted correctly by Viterbi POS tagger

0.6   Evaluation

```
[61]: # accuracy = number of correctly predicted tags / total number of encountered␣
      ↪tags
      pos_tagger.test_acc(test)
```

```
[61]: 0.7310082664574911
```

Our Viterbi Parts of Speech tagger is 73% accurate for predicting tags for the words in Brown data set