Casey Masamitsu | Week 13 | NLP

NLP

In [259…

```
#!pip install spacy
import pandas as pd
```

In [260…

```
# !python3 -m spacy download en_core_web_sm
```

In [261…

```
import spacy
nlp = spacy.load('en_core_web_sm')
```

In [262…

```
text = "Disney employs 38 lobbyists in Florida’s capital. Each election cycle,
```

In [263…

```
processed_text = nlp(text)
# (Omitted to shorten PDF) processed_text
```

Sentences

In [264…

```
# (Omitted to shorten PDF) n = 0
# (Omitted to shorten PDF) for sentence in processed_text.sents:
# (Omitted to shorten PDF)     print(n, sentence)
# (Omitted to shorten PDF)     n += 1
```

Words and Punctuation - Along with POS tagging

In [265…

```
# (Omitted to shorten PDF) n = 0
# (Omitted to shorten PDF) for sentence in processed_text.sents:
# (Omitted to shorten PDF)     for token in sentence:
# (Omitted to shorten PDF)         print(n, token, token.pos_, token.lemma_)
# (Omitted to shorten PDF)         n += 1
```

Entities

In [266…

```
# (Omitted to shorten PDF) for entity in processed_text.ents:
# (Omitted to shorten PDF)     print(entity, entity.label)
```

Noun Chunks

In [267…

```
# (Omitted to shorten PDF) for noun_chunk in processed_text.noun_chunks:
# (Omitted to shorten PDF)     print(noun_chunk)
```

Syntactic Dependency Parsing

In [268…

```
def pr_tree(word, level):
    if word.is_punct:
        return
    for child in word.lefts:
        pr_tree(child, level + 1)
    print('\t'* level + word.text + " - " + word.dep_)
    for child in word.rights:
        pr_tree(child, level + 1)
```

In [269…

```
# (Omitted to shorten PDF) for sentence in processed_text.sents:
# (Omitted to shorten PDF)     pr_tree(sentence.root, 0)
# (Omitted to shorten PDF)     print('-----------------------------------------
```

Word Vectorization

In [270…

```
# (Omitted to shorten PDF) proc_fruits = nlp('''I think green apples are delici
# (Omitted to shorten PDF)                           While pears have a strange textu
# (Omitted to shorten PDF)                           The bowls they sit in are ugly'

# (Omitted to shorten PDF) apples, pears, bowls = proc_fruits.sents
# fruit = proc_fruits.vocab["fruits"]
# print(apples.similarity(fruit))
```

In [271…

```
# (Omitted to shorten PDF) n = 0
# (Omitted to shorten PDF) for sent in proc_fruits.sents:
# (Omitted to shorten PDF)     for token in sent:
# (Omitted to shorten PDF)         if n < 3:
# (Omitted to shorten PDF)             print(token, token.vector)
# (Omitted to shorten PDF)         n += 1
```

Assignment

Find your favorite news source and grab the article text.

In [272…

```
florida = "Disney employs 38 lobbyists in Florida’s capital. Each election cycl
```

In [273…

```
import spacy
nlp = spacy.load('en_core_web_sm')
```

In [274…

```
florida = nlp(text)
# (Omitted to shorten PDF) florida
```

1. Show the most common words in the article.

In [275…

```
import pandas as pd
from collections import Counter
tokens = [token.text for token in florida if not token.is_punct if not token.is
freq = Counter(tokens)
freq.most_common(30)
```

Out[275]:

```
[('Disney', 40),
 ('Florida', 17),
 ('company', 15),
 ('state', 11),
 ('DeSantis', 11),
 ('World', 10),
 ('Reedy', 10),
 ('Creek', 10),
 ('special', 9),
 ('Mr.', 9),
 ('political', 6),
 ('tax', 6),
 ('law', 6),
 ('taxes', 6),
 ('theme', 5),
 ('park', 5),
 ('Orlando', 5),
 ('legislation', 5),
 ('million', 4),
 ('year', 4),
 ('district', 4),
 ('Republican', 4),
 ('said', 4),
 ('employees', 4),
 ('March', 4),
 ('voted', 3),
 ('revoke', 3),
 ('called', 3),
 ('Ron', 3),
 ('Gay', 3)]
```

2. Show the most common words under a part of speech. (i.e. NOUN: {'Bob':12, 'Alice':4,})

In [276…

```
pos = [token.pos_ for token in florida if not token.is_punct if not token.is_st
df = pd.DataFrame({"token": tokens, "type": pos})
df.head(25)
```

Out[276]:

	token	type
0	Disney	PROPN
1	employs	VERB
2	lobbyists	NOUN
3	Florida	PROPN
4	capital	NOUN
5	election	NOUN
6	cycle	NOUN
7	company	NOUN
8	gives	VERB
9	generous	ADJ
10	campaign	NOUN
11	contributions	NOUN
12	Florida	PROPN
13	candidates	NOUN
14	sides	NOUN
15	political	ADJ
16	aisle	NOUN
17	theme	NOUN
18	park	NOUN
19	megaresort	NOUN
20	near	ADP
21	Orlando	PROPN
22	attracts	VERB
23	million	NUM
24	visitors	NOUN

In [277…

```
df = df.groupby(["type", "token"]).size().reset_index(name = "counts")
df.head(50)
```

Out[277]:

	type	token	counts
0	ADJ	Real	1
1	ADJ	Republican	3
2	ADJ	able	1
3	ADJ	acute	1
4	ADJ	aggressive	1
5	ADJ	annual	1
6	ADJ	cheaper	1
7	ADJ	chief	1
8	ADJ	committed	1
9	ADJ	competitive	1
10	ADJ	congressional	1
11	ADJ	considerable	1
12	ADJ	critical	1
13	ADJ	domestic	1
14	ADJ	financial	1
15	ADJ	formal	1
16	ADJ	future	1
17	ADJ	generous	1
18	ADJ	great	1
19	ADJ	harsh	1
20	ADJ	huge	1
21	ADJ	largest	1
22	ADJ	latest	1
23	ADJ	local	2
24	ADJ	major	1
25	ADJ	medical	1
26	ADJ	municipal	2
27	ADJ	national	1
28	ADJ	new	1
29	ADJ	nonbinary	1
30	ADJ	older	1
31	ADJ	outdoor	1
32	ADJ	pandemic	2
33	ADJ	partisan	1
34	ADJ	political	6

	type	token	counts
35	ADJ	possible	2
36	ADJ	potential	1
37	ADJ	presidential	3
38	ADJ	primary	1
39	ADJ	private	1
40	ADJ	public	2
41	ADJ	real	1
42	ADJ	restrictive	1
43	ADJ	senior	2
44	ADJ	sexual	1
45	ADJ	similar	1
46	ADJ	social	1
47	ADJ	special	9
48	ADJ	swift	1
49	ADJ	symbolic	1

In [278…

```
com = df.sort_values(["type", "counts"], ascending=False).groupby("type").head(
com
```

Out[278]:

	type	token	counts
523	X	″Mr	1
490	VERB	said	4
404	VERB	called	3
450	VERB	including	3
484	VERB	revoke	3
...	...	...	...
26	ADJ	municipal	2
32	ADJ	pandemic	2
35	ADJ	possible	2
40	ADJ	public	2
43	ADJ	senior	2

62 rows × 3 columns

3. Find a subject/object relationship through the dependency parser in any sentence.

In [279…

```
import numpy as np
```

```
# First 10 sentences subject/object relationship
n = 1
for sentence in florida.sents:
    pr_tree(sentence.root, 0)
    print("SENTENCE", n, "~~~~~~~~~~~~")
    n += 1
    if n > 11:
        break
```

```
        Disney - nsubj
employs - ROOT
                38 - nummod
        lobbyists - dobj
                in - prep
                                Florida - poss
                                        ’s - case
                        capital - pobj
SENTENCE 1 ~~~~~~~~~~~~
                Each - det
                election - compound
        cycle - nsubj
                the - det
        company - nsubj
gives - ROOT
                generous - amod
                campaign - compound
        contributions - dobj
                to - prep
                                Florida - compound
                        candidates - pobj
                on - prep
                                both - det
                        sides - pobj
                                of - prep
                                                the - det
                                                political - amod
                                        aisle - pobj
SENTENCE 2 ~~~~~~~~~~~~
                Its - poss
                        theme - compound
                park - compound
        megaresort - nsubj
                near - prep
                        Orlando - pobj
attracts - ROOT
                        around - quantmod
                        50 - compound
                million - nummod
        visitors - dobj
                        a - det
                year - npadvmod
        powering - advcl
                        a - det
                                Central - compound
                        Florida - compound
                        tourism - compound
                economy - dobj
                                that - nsubj
                                annually - advmod
                        generates - relcl
                                        more - amod
                                        than - quantmod
                                        $ - quantmod
                                        5 - compound
                                billion - dobj
                                        in - prep
                                                                local - amod
                                                                        and -
cc
```

```
                                                                        state
- conj
                                                        tax - compound
                                                revenue - pobj
SENTENCE 3 ~~~~~~~~~~~~
                The - det
        upshot - nsubj
        Disney - nsubj
        usually - advmod
gets - ROOT
                whatever - dobj
                it - nsubj
        wants - ccomp
                in - prep
                        Florida - pobj
SENTENCE 4 ~~~~~~~~~~~~
                That - det
        era - nsubj
ended - ROOT
        on - prep
                Thursday - pobj
                                when - advmod
                                        the - det
                                        Florida - compound
                                House - nsubj
                        voted - relcl
                                        to - aux
                                revoke - xcomp
                                                        Disney - compound
                                                World - poss
                                                        ’s - case
                                        designation - dobj
                                        as - prep
                                                        a - det
                                                        special - amod
                                                        tax - compound
                                                district - pobj
                                                                a - det
                                                        privilege - appos
                                                                        that -
dobj
                                                                        Disney
- nsubj
                                                                        has -
aux
                                                                held - relcl
                                                                        for -
prep

55 - nummod

years - pobj
                                                effectively - advmod
                                        allowing - advcl
                                                                the - det
                                                        company - nsubj
                                                        to - aux
                                                        self - dep
                                                govern - ccomp
                                                                its - poss
```

```
                                                                        25,000
- nummod
                                                                acre - compoun
d
                                                                        theme
- compound
                                                                park - compoun
d
                                                        complex - dobj
SENTENCE 5 ~~~~~~~~~~~~
                The - det
                Florida - compound
        Senate - nsubj
voted - ROOT
        on - prep
                Wednesday - pobj
                to - aux
        eliminate - xcomp
                        the - det
                        special - amod
                zone - dobj
                                which - nsubjpass
                                is - auxpass
                        called - relcl
                                        the - det
                                        Reedy - compound
                                        Creek - compound
                                        Improvement - compound
                                District - oprd
SENTENCE 6 ~~~~~~~~~~~~
                Having - aux
        cleared - advcl
                        the - det
                way - dobj
                        to - prep
                                        this - det
                                outcome - pobj
                                        with - prep
                                                        a - det
                                                        formal - amod
                                                proclamation - pobj
                Gov. - compound
                Ron - compound
        DeSantis - nsubj
        will - aux
                almost - advmod
        certainly - advmod
make - ROOT
                        the - det
                measure - nsubj
        official - ccomp
        by - prep
                adding - pcomp
                                his - poss
                        signature - dobj
SENTENCE 7 ~~~~~~~~~~~~
        It - nsubj
        would - aux
take - ROOT
        effect - dobj
```

```
        in - prep
                June - pobj
                next - amod
        year - npadvmod
SENTENCE 8 ~~~~~~~~~~~~
                The - det
                swift - amod
        effort - nsubjpass
                        to - aux
                dissolve - acl
                                Reedy - compound
                        Creek - dobj
                        by - prep
                                        Florida - compound
                                Republicans - pobj
        has - aux
        been - auxpass
        widely - advmod
seen - ROOT
        as - prep
                        brazen - compound
                retaliation - pobj
                        after - prep
                                Disney - pobj
                                                Florida - poss
                                                        ’s - case
                                                largest - amod
                                                private - amod
                                        employer - appos
        paused - conj
                        political - amod
                donations - dobj
                in - prep
                                the - det
                        state - pobj
                and - cc
                condemned - conj
                                a - det
                                new - amod
                                education - compound
                        law - dobj
                                        that - dobj
                                        opponents - nsubj
                                call - relcl
                                                Do - aux
                                                n’t - neg
                                        Say - xcomp
                                                Gay - dobj
SENTENCE 9 ~~~~~~~~~~~~
        Among - prep
                        many - amod
                things - pobj
                the - det
        law - nsubj
prohibits - ROOT
        discussion - dobj
                about - prep
                                sexual - amod
                        orientation - pobj
                                and - cc
```

```
                                        gender - compound
                                identity - conj
        through - prep
                        the - det
                        third - amod
                grade - pobj
                        in - prep
                                        Florida - compound
                                classrooms - pobj
        and - cc
        limits - conj
                it - dobj
                for - prep
                                older - amod
                        students - pobj
        “If - punct
SENTENCE 10 ~~~~~~~~~~~~
                        Disney - nsubj
                wants - ccomp
                                to - aux
                        pick - xcomp
                                        a - det
                                fight - dobj
                they - nsubj
        chose - ccomp
                        the - det
                        wrong - amod
                guy - dobj
                Mr. - compound
        DeSantis - nsubj
                        a - det
                        potential - amod
                        Republican - amod
                        presidential - amod
                candidate - appos
                        in - prep
                                2024 - pobj
wrote - ROOT
        in - prep
                        a - det
                                fund - npadvmod
                        raising - amod
                email - pobj
        to - prep
                supporters - pobj
        on - prep
                Wednesday - pobj
SENTENCE 11 ~~~~~~~~~~~~
```

4. Show the most common Entities and their types.

In [280…

```
entities = pd.DataFrame({"entities": [entity for entity in florida.ents],
                        "entity_type": [entity.label_ for entity in florida.ent
entities.head(50)
```

Out[280]:

	entities	entity_type
0	(Disney)	ORG
1	(38)	CARDINAL
2	(Florida)	GPE
3	(Florida)	GPE
4	(Orlando)	GPE
5	(around, 50, million)	CARDINAL
6	(Central, Florida)	LOC
7	(more, than, $, 5, billion)	MONEY
8	(Disney)	ORG
9	(Florida)	GPE
10	(That, era, ended, on)	DATE
11	(Thursday)	DATE
12	(the, Florida, House)	ORG
13	(Disney, World, ’s)	ORG
14	(Disney)	ORG
15	(55, years)	DATE
16	(25,000, -, acre)	QUANTITY
17	(The, Florida, Senate)	ORG
18	(Wednesday)	DATE
19	(the, Reedy, Creek, Improvement, District)	ORG
20	(Ron, DeSantis)	PERSON
21	(June, next, year)	DATE
22	(Reedy, Creek)	GPE
23	(Florida)	GPE
24	(Republicans)	NORP
25	(Disney)	ORG
26	(Florida)	GPE
27	(Do, n’t, Say, Gay)	WORK_OF_ART
28	(third)	ORDINAL
29	(Florida)	GPE
30	(Disney)	ORG
31	(DeSantis)	PERSON
32	(Republican)	NORP
33	(2024)	DATE
34	(Wednesday)	DATE

	entities	entity_type
35	(California)	GPE
36	(Disney)	ORG
37	(Florida)	GPE
38	(Florida)	GPE
39	(Ron, DeSantis)	PERSON
40	(Republican)	NORP
41	(the, Conservative, Political, Action, Confere...	ORG
42	(Orlando)	GPE
43	(February)	DATE
44	(Ron, DeSantis)	PERSON
45	(Republican)	NORP
46	(the, Conservative, Political, Action, Confere...	ORG
47	(Orlando)	GPE
48	(February)	DATE
49	(Octavio, Jones, /, ReutersDisney)	PERSON

In [281…

```
com_entities = entities.groupby(["entities", "entity_type"]).size().reset_index
com_entities = com_entities.sort_values("counts", ascending = False).head(50)
com_entities
```

Out[281]:

	entities	entity_type	counts
0	(Disney)	ORG	1
118	(annual)	DATE	1
110	(two)	CARDINAL	1
111	(Orange)	GPE	1
112	(Osceola)	GPE	1
113	(Reedy, Creek)	PERSON	1
114	(Orange, County, ’s)	GPE	1
115	(Scott, Randolph)	PERSON	1
116	(as, much, as, 20, percent)	PERCENT	1
117	(Reedy, Creek)	PERSON	1
119	($, 355, million)	MONEY	1
108	(Disney)	ORG	1
120	($, 977, million)	MONEY	1
121	(Disney)	ORG	1
122	(Disney)	ORG	1
123	(DeSantis)	PERSON	1
124	(2020)	DATE	1
125	(Disney)	ORG	1
126	(Florida)	GPE	1
127	(Disney, World)	ORG	1
109	(Disney, World)	ORG	1
107	(more, than, $, 780, million)	MONEY	1
1	(38)	CARDINAL	1
96	(Daytona, International, Speedway)	ORG	1
88	(Wells, Fargo)	ORG	1
89	(Thursday)	DATE	1
90	(2.3, percent)	PERCENT	1
91	(Florida)	GPE	1
92	(hundreds)	CARDINAL	1
93	(One)	CARDINAL	1
94	(Villages)	ORG	1
95	(Orlando)	GPE	1
97	(Disney)	ORG	1
106	(Disney, World)	ORG	1
98	(six)	CARDINAL	1

	entities	entity_type	counts
99	(220, -, acre)	QUANTITY	1
100	(18)	CARDINAL	1
101	(Disney)	ORG	1
102	(24,000)	CARDINAL	1
103	(Disney, World)	ORG	1
104	(St., Louis)	GPE	1
105	(2021)	DATE	1
128	(March, 2020)	DATE	1
129	(July)	DATE	1
130	(Disneyland)	FAC	1
161	(March, 28)	DATE	1
153	(Disney)	ORG	1
154	(Chapek)	PERSON	1
155	(DeSantis)	PERSON	1
156	(Florida)	GPE	1

5. Find Entites and their dependency (hint: entity.root.head)

In [282…

```
n = 0
for entity in florida.ents:
    print(entity, "->", entity.root.head)
    n += 1
```

```
Disney -> employs
38 -> lobbyists
Florida -> capital
Florida -> candidates
Orlando -> near
around 50 million -> visitors
Central Florida -> economy
more than $5 billion -> generates
Disney -> gets
Florida -> in
That era ended on -> ended
Thursday -> on
the Florida House -> voted
Disney World’s -> designation
Disney -> held
55 years -> for
25,000-acre -> complex
The Florida Senate -> voted
Wednesday -> on
the Reedy Creek Improvement District -> called
Ron DeSantis -> make
June next year -> take
Reedy Creek -> dissolve
Florida -> Republicans
Republicans -> by
Disney -> after
Florida -> employer
Don’t Say Gay -> call
third -> grade
Florida -> classrooms
Disney -> wants
DeSantis -> wrote
Republican -> candidate
2024 -> in
Wednesday -> on
California -> in
Disney -> gotten
Florida -> of
Florida -> ruled
Ron DeSantis -> DeSantis
Republican -> nomination
the Conservative Political Action Conference -> at
Orlando -> in
February -> in
Ron DeSantis -> DeSantis
Republican -> nomination
the Conservative Political Action Conference -> at
Orlando -> in
February -> in
Octavio Jones/ReutersDisney -> declined
The Reedy Creek Improvement District -> saves
1967 -> in
Disney -> entice
20 miles -> south
Orlando -> of
millions of dollars -> saves
annually -> saves
Disney -> spent
decades -> spent
the Florida Senate -> voted
```

```
Disney World’s -> status
The End of Social Distancing: -> End
Disneyland -> at
Disney+.A Documentary -> of
one -> of
Disney -> founders
Reedy Creek -> provide
Disney -> provide
A few years ago -> had
Disney -> wanted
Hollywood Studios -> park
Disney World -> at
Reedy Creek -> issue
the 1990s -> in
Disney -> needed
Anaheim -> in
Calif. -> Anaheim
California Adventure -> park
Anaheim -> persuade
Disney -> to
Reedy Creek -> gives
Disney -> gives
Reedy Creek -> levies
Disney -> on
Disney World -> generates
Reedy Creek -> through
Disney -> on
Steven Cahall -> said
Wells Fargo -> analyst
Thursday -> on
2.3 percent -> down
Florida -> has
hundreds -> has
One -> covers
Villages -> covers
Orlando -> of
Daytona International Speedway -> covers
Disney -> for
six -> parks
220-acre -> basketball
18 -> hotels
Disney -> owned
24,000 -> rooms
Disney World -> has
St. Louis -> of
2021 -> In
Disney World -> paid
more than $780 million -> paid
Disney -> disclosure
Disney World -> straddles
two -> counties
Orange -> counties
Osceola -> Orange
Reedy Creek -> by
Orange County’s -> collector
Scott Randolph -> collector
as much as 20 percent -> climb
Reedy Creek -> has
annual -> budget
$355 million -> of
```

```
$977 million -> carries
Disney -> apply
Disney -> been
DeSantis -> with
2020 -> In
Disney -> benefited
Florida -> reopen
Disney World -> closed
March 2020 -> in
July -> in
Disneyland -> reopen
California -> in
last April -> until
Florida -> with
Last year -> threatened
Georgia -> politicians
Delta Air Lines -> on
Texas -> lawmakers
Citigroup -> bar
DeSantis -> between
Disney -> DeSantis
March 9 -> on
the Parents Rights -> legislation
Education -> in
Don’t Say Gay -> bill
More than 150 -> companies
Marriott -> including
American Airlines -> Marriott
a Human Rights Campaign -> letter
Disney -> avoided
Bob Chapek -> executive
March 7 -> on
days later -> did
Disney -> for
Chapek -> did
DeSantis -> called
Florida -> in
Chapek -> said
DeSantis -> rile
Disney -> spokesman
DeSantis -> signed
March 28 -> on
Disney -> renewed
Disney -> said
Disney World’s -> district
The Florida Legislature -> convened
this week -> convened
DeSantis -> issued
Tuesday -> on
Republican -> controlled
1968 -> before
Disney -> with
```

6. Find the most similar words in the article

In [283…

```
noun_chunks = [(token1.text, token2.text, token1.similarity(token2)) for token2
noun_chunks = sorted([item for item in noun_chunks if item[-1] != 1], key=lambd
df = pd.DataFrame(noun_chunks)
```

```
/var/folders/tm/ygp25lv10ss4cp_scjbh05r40000gp/T/ipykernel_10195/3498856822.p
y:1: UserWarning: [W007] The model you're using has no word vectors loaded, so
the result of the Span.similarity method will be based on the tagger, parser a
nd NER, which may not give useful similarity judgements. This may happen if yo
u're using one of the small models, e.g. `en_core_web_sm`, which don't ship wi
th word vectors and only use context-sensitive tensors. You can always add you
r own word vectors, or use one of the larger models instead if available.
  noun_chunks = [(token1.text, token2.text, token1.similarity(token2)) for tok
en2 in florida.noun_chunks for token1 in florida.noun_chunks]
```

In [284…

```
df.columns = ["First Chunk", "Second Chunk", "Similarity"]
df.head(20)
```

Out[284]:

	First Chunk	Second Chunk	Similarity
0	Disney World’s special tax district	Disney World’s special tax status	0.986590
1	Disney World’s special tax status	Disney World’s special tax district	0.986590
2	The Florida Legislature	The Florida Senate	0.933412
3	The Florida Senate	The Florida Legislature	0.933412
4	Florida	California	0.925607
5	California	Florida	0.925607
6	He	They	0.920532
7	They	He	0.920532
8	Reedy Creek	Disney World	0.918186
9	Disney World	Reedy Creek	0.918186
10	the law	the company	0.917102
11	the company	the law	0.917102
12	The Reedy Creek Improvement District	the Reedy Creek Improvement District	0.911015
13	the Reedy Creek Improvement District	The Reedy Creek Improvement District	0.911015
14	The company	The designation	0.903494
15	The designation	The company	0.903494
16	Gov. Ron DeSantis	Ron DeSantis	0.901868
17	Ron DeSantis	Gov. Ron DeSantis	0.901868
18	July	June	0.895180
19	June	July	0.895180

In [ ]: