Appendix

Appendix A: spaCy Part of Speech Tag List

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary verb
- CONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

Appendix B: NLTK Part of Speech Tag List

$: dollar

$ -$ --$ A$ C$ HK$ M$ NZ$ S$ U.S.$ US$

'': closing quotation mark

' ''

(: opening parenthesis

([{

): closing parenthesis

)] }

,: comma

,

--: dash

--

.: sentence terminator

. ! ?

:: colon or ellipsis

: ; ...

CC: conjunction, coordinating

& 'n and both but either et for less minus neither nor or plus so

therefore times v. versus vs. whether yet

CD: numeral, cardinal

mid-1890 nine-thirty forty-two one-tenth ten million 0.5 one forty-

seven 1987 twenty '79 zero two 78-degrees eighty-four IX '60s .025

fifteen 271,124 dozen quintillion DM2,000 ...

DT: determiner

all an another any both del each either every half la many much nary

neither no some such that the them these this those

EX: existential there

there

FW: foreign word

gemeinschaft hund ich jeux habeas Haementeria Herr K'ang-si vous lutihaw alai je jour objets salutaris fille quibusdam pas trop Monte terram fiche oui corporis ...

IN: preposition or conjunction, subordinating

astride among uppon whether out inside pro despite on by throughout below within for towards near behind atop around if like until below next into if beside ...

JJ: adjective or numeral, ordinal

third ill-mannered pre-war regrettable oiled calamitous first separable ectoplasmic battery-powered participatory fourth still-to-be-named multilingual multi-disciplinary ...

JJR: adjective, comparative

bleaker braver breezier briefer brighter brisker broader bumper busier calmer cheaper choosier cleaner clearer closer colder commoner costlier cozier creamier crunchier cuter ...

JJS: adjective, superlative

calmest cheapest choicest classiest cleanest clearest closest commonest corniest costliest crassest creepiest crudest cutest darkest deadliest dearest deepest densest dinkiest ...

LS: list item marker

A A. B B. C C. D E F First G H I J K One SP-44001 SP-44002 SP-44005 SP-44007 Second Third Three Two * a b c d first five four one six three two

MD: modal auxiliary

can cannot could couldn't dare may might must need ought shall should shouldn't will would

NN: noun, common, singular or mass

common-carrier cabbage knuckle-duster Casino afghan shed thermostat investment slide humour falloff slick wind hyena override subhumanity machinist ...

NNP: noun, proper, singular

Motown Venneboerger Czestochwa Ranzer Conchita Trumplane Christos Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin ODI Darryl CTCA Shannon A.K.C. Meltex Liverpool ...

NNPS: noun, proper, plural

Americans Americas Amharas Amityvilles Amusements Anarcho-Syndicalists Andalusians Andes Andruses Angels Animals Anthony Antilles Antiques Apache Apaches Apocrypha ...

NNS: noun, common, plural

undergraduates scotches bric-a-brac products bodyguards facets coasts divestitures storehouses designs clubs fragrances averages subjectivists apprehensions muses factory-jobs ...

PDT: pre-determiner

all both half many quite such sure this

POS: genitive marker

' 's

PRP: pronoun, personal

hers herself him himself hisself it itself me myself one oneself ours ourselves ownself self she thee theirs them themselves they thou thy us

PRP$: pronoun, possessive

her his mine my our ours their thy your

RB: adverb

occasionally unabatingly maddeningly adventurously professedly stirringly prominently technologically magisterially predominately swiftly fiscally pitilessly ...

RBR: adverb, comparative

further gloomier grander graver greater grimmer harder harsher

healthier heavier higher however larger later leaner lengthier less-

perfectly lesser lonelier longer louder lower more ...

RBS: adverb, superlative

best biggest bluntest earliest farthest first furthest hardest

heartiest highest largest least less most nearest second tightest worst

RP: particle

aboard about across along apart around aside at away back before behind

by crop down ever fast for forth from go high i.e. in into just later

low more off on open out over per pie raising start teeth that through

under unto up up-pp upon whole with you

SYM: symbol

% & ' '' ''.)). * + ,. < = > @ A[fj] U.S U.S.S.R * ** ***

TO: "to" as preposition or infinitive marker

to

UH: interjection

Goodbye Goody Gosh Wow Jeepers Jee-sus Hubba Hey Kee-reist Oops amen

huh howdy uh dammit whammo shucks heck anyways whodunnit honey golly

man baby diddle hush sonuvabitch ...

VB: verb, base form

ask assemble assess assign assume atone attention avoid bake balkanize

bank begin behold believe bend benefit bevel beware bless boil bomb

boost brace break bring broil brush build ...

VBD: verb, past tense

dipped pleaded swiped regummed soaked tidied convened halted registered

cushioned exacted snubbed strode aimed adopted belied figgered

speculated wore appreciated contemplated ...

VBG: verb, present participle or gerund

telegraphing stirring focusing angering judging stalling lactating

hankerin' alleging veering capping approaching traveling besieging

encrypting interrupting erasing wincing ...

VBN: verb, past participle

multihulled dilapidated aerosolized chaired languished panelized used

experimented flourished imitated reunifed factored condensed sheared

unsettled primed dubbed desired ...

VBP: verb, present tense, not 3rd person singular

predominate wrap resort sue twist spill cure lengthen brush terminate

appear tend stray glisten obtain comprise detest tease attract

emphasize mold postpone sever return wag ...

VBZ: verb, present tense, 3rd person singular

bases reconstructs marks mixes displeases seals carps weaves snatches

slumps stretches authorizes smolders pictures emerges stockpiles

seduces fizzes uses bolsters slaps speaks pleads ...

WDT: WH-determiner

that what whatever which whichever

WP: WH-pronoun

that what whatever whatsoever which who whom whosoever

WP$: WH-pronoun, possessive

whose

WRB: Wh-adverb

how however whence whenever where whereby whereever wherein whereof why

``: opening quotation mark

```
```

Appendix C: Stopword List

'm	almost	aren	beforehand
're	alone	aren'	begin
's	along	arent	beginning
've	already	arise	beginnings
a	also	around	begins
able	although	as	behind
about	always	aside	being
above	am	ask	believe
absolutely	amazing	asking	below
accordance	among	at	beside
according	amongst	auth	besides
accordingly	an	available	best
across	and	away	better
act	announce	awesome	between
actually	another	awful	beyond
added	any	awfully	big
adj	anybody	back	bigger
affected	anyhow	bad	biggest
affecting	anymore	basically	biol
affects	anyone	be	bit
after	anything	became	both
afterwards	anyway	because	brief
again	anyways	become	briefly
against	anywhere	becomes	but
ago	apparently	becoming	by
ah	approximately	been	ca
all	are	before	came

can
can'
can't
cannot
cause
causes
certain
certainly
clearly
co
com
come
comes
constantly
contain
containing
contains
could
couldn
couldn'
couldnt
currently
date
day
days
definitely
despite
did
didn
didn'
didn't
different
do
does
doesn
doesn'
doesn't
doing
don
don't
done
dont
down
downwards
due
during
each
easi
easier
easy
ed
edu
effect
eg
eight
eighty
either
else
elsewhere
end
ending
enough
entire
especially
et
et-al
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
excellent
except
extremely
fantastic
far
feel
feeling
few
ff
fifth
finally
fine
first
five
fix
followed
following
follows
for
forever
former
formerly
forth
found
four
free
from
further
furthermore
gave
get
gets
getting
give
given
gives
giving
glad
go
goes
going

gone
good
goodbye
got
gotten
great
greatest
had
half
happens
happy
hardly
has
hasn't
have
haven
haven'
haven't
having
he
hed
hello
hence
her
here
hereafter
hereby
herein
heres
hereupon
hers
herself
hes
hi
hid
him
himself
his
hither
home
hope
hour
hours
how
howbeit
however
huge
hundred
i
i'll
i've
id
ie
if
im
immediate
immediately
importance
important
in
inc
incredible
incredibly
indeed
index
information
initially
instead
into
invention
inward
is
isn
isn'
isn't
issue
issues
it
it'll
itd
its
itself
just
keep
keeps
kept
kg
kind
km
know
known
knows
largely
last
lately
later
latter
latterly
least
less
lest
let
lets
like
liked
likely
line
literally
little
ll
look
looking
looks
lot
lots
love

loved
ltd
luck
made
mainly
make
makes
many
matter
may
maybe
me
mean
means
meantime
meanwhile
merely
mg
might
million
minute
minutes
miss
ml
month
months
more
moreover
most
mostly
mr
mrs
much
mug
must
my
myself
n't
na
name
namely
nay
nd
near
nearly
necessarily
necessary
need
needs
neither
never
nevertheless
new
next
nine
ninety
no
nobody
non
none
nonetheless
noone
nor
normally
nos
not
noted
nothing
now
nowhere
obtain
obtained
obviously
of
off
often
oh
ok
okay
old
omitted
on
once
one
ones
only
onto
or
ord
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
owing
own
page
pages
part
particular
particularly
past
per
perfect
perhaps
placed
please
plus
poor
poorly

possible
possibly
potentially
pp
predominantly
present
pretty
previously
primarily
probably
problem
problems
promptly
proud
provides
put
que
quickly
quite
qv
ran
rather
rd
re
readily
really
recent
recently
ref
refs
regarding
regardless
regards
related
relatively
research
respectively
resulted
resulting
results
review
reviews
right
rock
rocks
run
s
said
same
saw
say
saying
says
sec
second
seconds
section
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sent
seriously
seven
several
shall
she
she'll
shed
shes
should
shouldn't
show
showed
shown
shows
significant
significantly
similar
similarly
since
six
slightly
so
some
somebody
somehow
someone
somethan
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specifically
specified
specify
specifying
still
stop
strongly
stuff
sub
substantially
successfully
such
suck
sucks
suddenly

sufficiently
suggest
sup
super
sure
t
take
taken
taking
tell
tends
terrible
th
than
thank
thanks
thanx
that
that'll
that've
thats
the
their
theirs
them
themselves
then
thence
there
there'll
there've
thereafter
thereby
thered
therefore
therein
thereof
theres
thereto
thereupon
these
they
they'll
they've
theyd
theyre
thing
things
think
this
those
thou
though
though
thousand
through
throughout
thru
thus
til
till
time
times
tip
to
together
too
took
totally
toward
towards
tried
tries
true
truly
try
trying
ts
twice
two
un
under
unfortunately
unless
unlike
unlikely
until
unto
up
upon
ups
us
use
used
useful
usefully
usefulness
uses
using
usually
various
ve
veri
very
via
viz
vol
vols
vs
want
wants
was
wasn
wasn'
wasnt
way

we
we'll
we've
wed
week
weeks
welcome
well
went
were
werent
what
what'll
whatever
whats
when
whence
whenever
where
whereafter
whereas
whereby
wherein
wheres
whereupon
wherever
whether
which
while
whim
whither
who
who'll
whod
whoever
whole
whom
whomever
whos
whose
why
widely
will
willing
wish
with
within
without
wont
words
world
worse
worst
would
wouldn
wouldn'
wouldnt
wow
wrong
www
year
years
yes
yet
you
you'll
you've
youd
your
youre
yours
yourself
yourselves
zero

Appendix D: spaCy Named Entity Tags

spaCy trains two types of models for Named Entity Recognition. Models trained on the OntoNotes 5 corpus have the following tags:

Type	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including ”%“.
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.

Models trained on Wikipedia data have the following scheme:

Type	Description
PER	Named person or family.
LOC	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
ORG	Named corporate, governmental, or other organizational entity.
MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art.