NLTK\_Cheatsheet.docx

Table of Contents

[Break a text into sentences 1](#_Toc77052534)

[Filter out stop words 1](#_Toc77052535)

[POS Tag Meanings 2](#_Toc77052536)

[Tagging Parts of Speech 4](#_Toc77052537)

[Tokenize text by word 5](#_Toc77052538)

[Tweet Tokenizer 5](#_Toc77052539)

# Break a text into sentences

from nltk.tokenize import sent\_tokenize, word\_tokenize

import nltk

nltk.download('punkt')

example\_string = """

... Muad'Dib learned rapidly because his first training was in how to learn.

... And the first lesson of all was the basic trust that he could learn.

... It's shocking to find how many people do not believe they can learn,

... and how many more believe learning to be difficult."""

sent\_tokenize(example\_string)

Output:

["\nMuad'Dib learned rapidly because his first training was in how to learn.",

'And the first lesson of all was the basic trust that he could learn.',

"It's shocking to find how many people do not believe they can learn,\nand how many more believe learning to be difficult."]

# Filter out stop words

nltk.download("stopwords")

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

worf\_quote = "Sir, I protest. I am not a merry man!"

words\_in\_quote = word\_tokenize(worf\_quote)

stop\_words = set(stopwords.words("english"))

filtered\_list = [word for word in words\_in\_quote if word.casefold() not in stop\_words]

filtered\_list

Output:

['Sir', ',', 'protest', '.', 'merry', 'man', '!']

# Lemmatizing

Note: A lemma is a word that represents a whole group of words, and that group of words is called a lexeme.

For example, if you were to look up the word “blending” in a dictionary, then you’d need to look at the entry for “blend,” but you would find “blending” listed in that entry.

In this example, “blend” is the lemma, and “blending” is part of the lexeme. So when you lemmatize a word, you are reducing it to its lemma.

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

lemmatizer.lemmatize("scarves")

Output: ‘scarf’

## Lemmatizing a string

string\_for\_lemmatizing = "The friends of DeSoto love scarves."

words = word\_tokenize(string\_for\_lemmatizing)

words

Output: ['The', 'friends', 'of', 'DeSoto', 'love', 'scarves', '.']

lemmatized\_words = [lemmatizer.lemmatize(word) for word in words]

lemmatized\_words

Output: ['The', 'friend', 'of', 'DeSoto', 'love', 'scarf', '.']

# POS Tag Meanings

nltk.download('tagsets')

nltk.help.upenn\_tagset()

Output:

$: 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

# Tagging Parts of Speech

from nltk.tokenize import word\_tokenize

sagan\_quote = """

If you wish to make an apple pie from scratch,

you must first invent the universe."""

words\_in\_sagan\_quote = word\_tokenize(sagan\_quote)

import nltk

nltk.pos\_tag(words\_in\_sagan\_quote)

Output:

[('If', 'IN'),

('you', 'PRP'),

('wish', 'VBP'),

('to', 'TO'),

('make', 'VB'),

('an', 'DT'),

('apple', 'NN'),

('pie', 'NN'),

('from', 'IN'),

('scratch', 'NN'), ... ]

# Tokenize text by word

word\_tokenize(example\_string)

Output:

["Muad'Dib",

'learned',

'rapidly',

'because',

'his',

'first',

'training',

'was',

'in',

'how',

'to',

'learn',

'.',

…

# Tweet Tokenizer

from nltk.tokenize import TweetTokenizer # module for tokenizing strings

print()

print('\033[92m' + tweet2)

print('\033[94m')

# instantiate tokenizer class

tokenizer = TweetTokenizer(preserve\_case=False, strip\_handles=True,

reduce\_len=True)

# tokenize tweets

tweet\_tokens = tokenizer.tokenize(tweet2)

print()

print('Tokenized string:')

print(tweet\_tokens)

Output:

**My beautiful sunflowers on a sunny Friday morning off :) sunflowers favourites happy Friday off…**

**Tokenized string:**

**['my', 'beautiful', 'sunflowers', 'on', 'a', 'sunny', 'friday', 'morning', 'off', ':)', 'sunflowers', 'favourites', 'happy', 'friday', 'off', '…']**