Google Colab Link: https://drive.google.com/file/d/1mLOzDdQhpagk1bBMMlHJjeYCIhBYaOMv/view?usp=sharing

LAB E5 POS Tagging and Lemmatization

1. Import the necessary libraries.

In [1]:
```
import sys
import nltk
#nltk.download("punkt")
#nltk.download('wordnet')
#nltk.download('maxent_ne_chunker')
#nltk.download('words')
#nltk.download('averaged_perceptron_tagger')
#nltk.download('tagsets')
from nltk import word_tokenize
from nltk import sent_tokenize
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
from nltk import pos_tag
import re
import spacy
import pandas as pd
import string
```

2. Given a text from a file: LabE5.txt

In [2]:
```
nlp = spacy.load("en_core_web_sm")
data_path = "C:\spark\MCA\Semester1\E3_NLP\input\lab_e5\LabE_5.txt"
text = ""
with open (data_path, "r") as f:
    text = f.read()
text = str(text.lower())
text= text.translate(str.maketrans('','',string.punctuation))
text = text.translate(str.maketrans('','','\t\n'))
```

a. For each word in the sentence, tag/ mark the corresponding POS using any POS tagger

In [3]:

```
doc = nlp(text)
list_token = []
col_list_token = ["Text", "POS", "Explanation", "Tag"]
for token in doc:
    list_token.append([token.text, token.pos_, spacy.explain(token.pos_),
token.tag_])
df_spacy_pos = pd.DataFrame(list_token, columns = col_list_token)
df_spacy_pos
```

Out[3]:

	Text	POS	Explanation	Tag
0	great	ADJ	adjective	JJ
1	buy	NOUN	noun	NN
2	always	ADV	adverb	RB
3	go	VERB	verb	VB
4	with	ADP	adposition	IN
...	...	...	...	...
12427	recommend	VERB	verb	VB
12428	the	DET	determiner	DT
12429	company	NOUN	noun	NN
12430	to	ADP	adposition	IN
12431	everyone	PRON	pronoun	NN

12432 rows × 4 columns

b. Reduce each words to lemma based on the POS tagger

In [4]:

```
wnl = WordNetLemmatizer()
lem_list = []
for i in df_spacy_pos.values:
    lem_list.append(wnl.lemmatize(i[0]))
df_spacy_pos["Lemma"] = lem_list
df_spacy_pos[:20]
```

Out[4]:

	Text	POS	Explanation	Tag	Lemma
0	great	ADJ	adjective	JJ	great
1	buy	NOUN	noun	NN	buy
2	always	ADV	adverb	RB	always
3	go	VERB	verb	VB	go
4	with	ADP	adposition	IN	with
5	white	ADJ	adjective	JJ	white
6	dark	ADJ	adjective	JJ	dark
7	colors	NOUN	noun	NNS	color

	Text	POS	Explanation	Tag	Lemma
8	melt	VERB	verb	VBP	melt
9	on	ADP	adposition	IN	on
10	the	DET	determiner	DT	the
11	florida	PROPN	proper noun	NNP	florida
12	sun	NOUN	noun	NN	sun
13	experience	NOUN	noun	NN	experience
14	over	ADP	adposition	IN	over
15		SPACE	space	_SP	
16	cordless	NOUN	noun	NN	cordless
17	is	AUX	auxiliary	VBZ	is

3. For same text in text file

In [5]:
```
taggedSentence = pos_tag(df_spacy_pos["Text"])
```

In [6]:
```
col_list_token = ["Text", "POS", "Explanation", "Tag", "Lemma"]
tags = nltk.data.load('help/tagsets/upenn_tagset.pickle')
list_token_nltk = []
for i in taggedSentence:
    list_token_nltk.append([i[0], i[1], tags[i[1]][0], i[1],
wnl.lemmatize(i[0])])
df_nltk_pos = pd.DataFrame(list_token_nltk, columns = col_list_token)
df_nltk_pos
```

Out[6]:

	Text	POS	Explanation	Tag	Lemma
0	great	JJ	adjective or numeral, ordinal	JJ	great
1	buy	NN	noun, common, singular or mass	NN	buy
2	always	RB	adverb	RB	always
3	go	VBP	verb, present tense, not 3rd person singular	VBP	go
4	with	IN	preposition or conjunction, subordinating	IN	with
...	...	...	...	...	...
12427	recommend	VB	verb, base form	VB	recommend
12428	the	DT	determiner	DT	the
12429	company	NN	noun, common, singular or mass	NN	company
12430	to	TO	"to" as preposition or infinitive marker	TO	to
12431	everyone	NN	noun, common, singular or mass	NN	everyone

12432 rows × 5 columns

a. Compare the tags generated spaCy and NLTK.

In [7]:

```
print(list(df_spacy_pos['Tag']) == list(df_nltk_pos['Tag']))
#comparing tag differences in first 50 toekns
for i in range(0,50):
    if(df_spacy_pos['Tag'][i] != df_nltk_pos['Tag'][i]):
        print(df_spacy_pos.iloc[i,0:4].values, "\t\t!=\t\t",
df_nltk_pos.iloc[i,0:4].values)
```

```
False
['go' 'VERB' 'verb' 'VB']               !=              ['go' 'VBP' 'verb, present ten
se, not 3rd person singular' 'VBP']
['dark' 'ADJ' 'adjective' 'JJ']                 !=              ['dark' 'NN' 'noun, co
mmon, singular or mass' 'NN']
['melt' 'VERB' 'verb' 'VBP']            !=              ['melt' 'VBD' 'verb, past tens
e' 'VBD']
['florida' 'PROPN' 'proper noun' 'NNP']                 !=              ['florida' 'JJ
' 'adjective or numeral, ordinal' 'JJ']
[' ' 'SPACE' 'space' '_SP']             !=              [' ' 'NNP' 'noun, proper, sing
ular' 'NNP']
['up' 'ADV' 'adverb' 'RB']              !=              ['up' 'RP' 'particle' 'RP']
['i' 'PRON' 'pronoun' 'PRP']            !=              ['i' 'VB' 'verb, base form' 'V
B']
['cordless' 'ADJ' 'adjective' 'JJ']             !=              ['cordless' 'NN' 'nou
n, common, singular or mass' 'NN']
['toowhen' 'ADV' 'adverb' 'RB']             !=              ['toowhen' 'NN' 'noun,
common, singular or mass' 'NN']
['blind' 'ADJ' 'adjective' 'JJ']                !=              ['blind' 'NN' 'noun, c
ommon, singular or mass' 'NN']
```

In [8]:

```
matched_count = 0
unmatched_count = 0
for i in range(len(df_spacy_pos)):
    if(df_spacy_pos['Tag'][i] != df_nltk_pos['Tag'][i]):
        unmatched_count = unmatched_count + 1
    else:
        matched_count = matched_count + 1
print("Number of matched tags = ", matched_count)
print("Number of unmatched tags = ", unmatched_count)
print("Spacy can generate both fine-grained(tag) and coarse-grained part of
speech(pos)")
```

```
Number of matched tags =  10565
Number of unmatched tags =  1867
Spacy can generate both fine-grained(tag) and coarse-grained part of speech(pos)
```

b. Use Penn Tree Bank Tagset

In [9]:

```
nltk.help.upenn_tagset()
```

```
$: 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
```

4. Give few product reviews from the file: LabE5.txt

a. Extract product descriptors of the concerned product.

In [10]:
```
df_nltk_pos
```

Out[10]:

	Text	POS	Explanation	Tag	Lemma
0	great	JJ	adjective or numeral, ordinal	JJ	great
1	buy	NN	noun, common, singular or mass	NN	buy
2	always	RB	adverb	RB	always
3	go	VBP	verb, present tense, not 3rd person singular	VBP	go
4	with	IN	preposition or conjunction, subordinating	IN	with
...	...	...	...	...	...
12427	recommend	VB	verb, base form	VB	recommend
12428	the	DT	determiner	DT	the

	Text	POS	Explanation	Tag	Lemma
12429	company	NN	noun, common, singular or mass	NN	company
12430	to	TO	"to" as preposition or infinitive marker	TO	to
12431	everyone	NN	noun, common, singular or mass	NN	everyone

In [11]:
```
print(df_nltk_pos[df_nltk_pos['Tag'].isin(["JJ", "JJR", "JJS"])]
["Text"].values[:100])
```

```
['great' 'white' 'florida' 'cool' 'tooawesome' 'easy' 'perfect' 'original'
 'frequent' 'top' 'free' 'free' 'fastvery' 'x' 'little' 'happy' 'able'
 'narrow' 'special' 'first' 'i' 'bottom' 'suppose' 'slat' 'full' 'last'
 'confirmationbought' 'high' 'nice' 'timely' 'old' 'french' 'several'
 'broken' 'cheap' 'quick' 'good' 'good' ' ' 'i' 'comfortable' 'fit'
 'adorable' 'vibrant' 'halloween' 'actual' 'burnt' 'super' 'soft' 'weird'
 'silky' 'soft' 'quilt' 'soft' 'inch' 'last' 'other' 'good' 'good'
 'different' 'other' 'soft' 'definite' 'such' 'sensitive' 'better' 'soft'
 'different' 'pillow' 'hot' 'workable' 'elastic' 'small' 'poor' 'unsown'
 'salmon' 'dirty' 'usedthis' 'correct' 'i' 'perfect' 'bedtite' 'deep'
 'fitting' 'described' 'several' 'perfect' 'many' 'overall' 'difficultthe'
 'soft' 'comfortable' 'orange' 'bright' 'holy' 'closed' 'peachy' 'neon'
 'last' 'i']
```

b. Here use spaCy : for getting descriptors

In [12]:
```
print(df_spacy_pos[df_spacy_pos['POS']=="ADJ"]["Text"].values[:100])
```

```
['great' 'white' 'dark' 'cool' 'tooawesome' 'easy' 'cordless' 'perfect'
 'blind' 'stuck' 'original' 'frequent' 'top' 'free' 'blind' 'free' 'blind'
 'little' 'happy' 'able' 'narrow' 'special' 'first' '2nd' 'bottom' 'front'
 'slat' 'full' 'last' 'high' 'cordless' 'nice' 'timely' 'easy' 'old'
 'french' 'several' 'slatsgreat' 'worth' 'cheap' 'quick' 'good' 'nice'
 'good' 'comfortable' 'fit' 'adorable' 'only' 'vibrant' 'actual' 'soft'
 'weird' 'soft' 'soft' 'thick' 'last' 'other' 'good' 'good' 'pink'
 'different' 'other' 'soft' 'definite' 'such' 'sensitive' 'scratchy'
 'better' 'soft' 'ill' 'different' 'stripe' 'hot' 'pink' 'workable'
 'shabby' 'small' 'poor' 'unsown' 'orange' 'ugly' 'dirty' 'usedthis'
 'correct' 'perfect' 'softweve' 'bedtite' 'deep' 'fitting' 'several'
 'highly' 'perfect' 'worth' 'many' 'overall' 'difficultthe' 'soft'
 'comfortable' 'madethere' 'orange']
```