AIWIR Lab Week 2
Team-8

Team members :

1.	A Spoorthi Alva	PES2UG19CS001
2.	A R Manyatha	PES2UG19CS002
3.	Achyut Jagini.	PES2UG19CS013
4.	Amulya S Dinesh	PES2UG19CS035

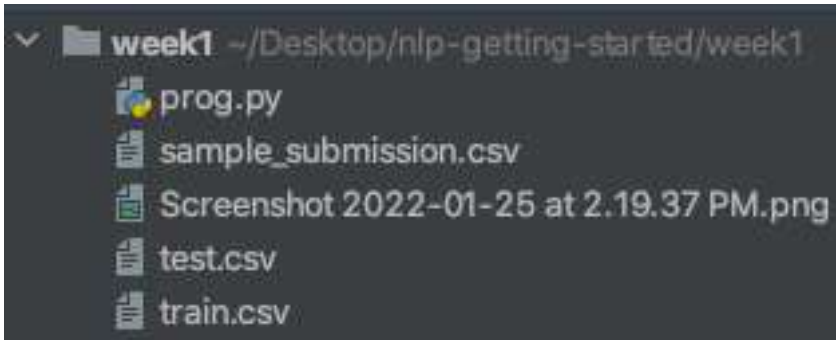
Dataset: https://www.kaggle.com/c/nlp-getting-started/data?select=train.csv

Tasks :

- ➔ Tokenize each Tweet into sentences
- ➔ Tokenize each tweet into words
- ➔ Remove stopwords in each tweet - NLTK library

Tool used : Pycharm

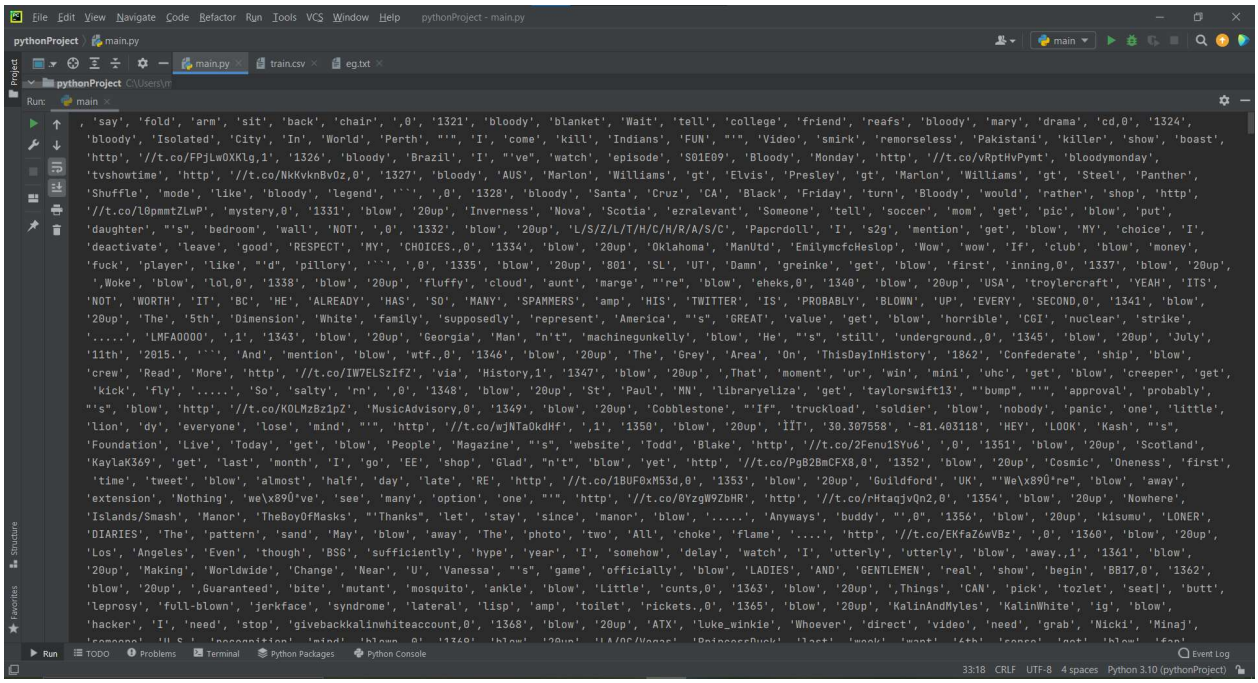
- ➔ It is installed on the computer.
- ➔ The dataset is downloaded and added to the folder where the main program exists
- ➔ Project is created in Pycharm by opening a folder containing the dataset.
- ➔ Required modules are installed.

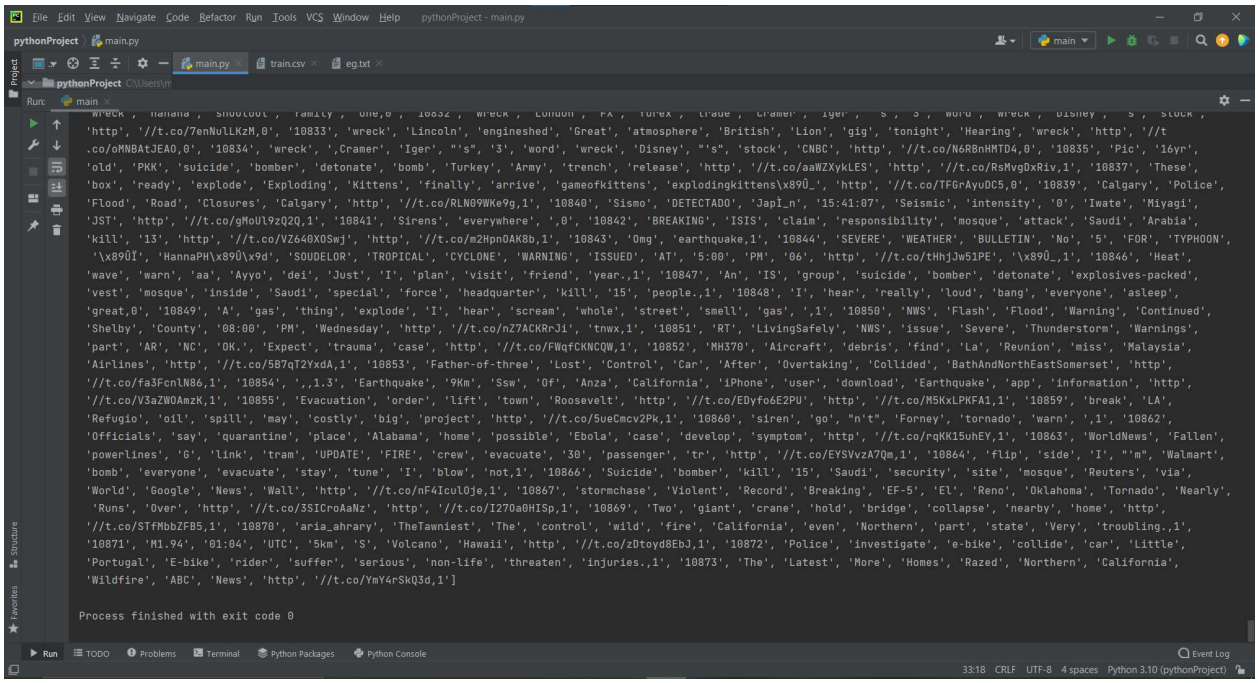


Code :

```
import nltk
nltk.download('punkt')
nltk.download('omw-1.4')
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
stop_words=set(stopwords.words('english'))
#using stemming-TEXT NORMALIZATION
file='train.csv'
filer=open(file,"r",encoding="utf-8")
text=filer.read()
text=text.replace("\n"," ")
#tokenizing words
word_tokens = word_tokenize(text)

filtered_sentence = []
symbols=[':','/',',','"',"'",'(',')','@','?',';','//','#','!','&','$','%','*','...
','.','..','-','[',']','{','}']
for w in word_tokens:
    if w not in stop_words and w not in symbols:
        filtered_sentence.append(w)
print("The text after stemming" )
Stem_words = []
ps = PorterStemmer()


for w in filtered_sentence:
    rootWord = ps.stem(w)
    Stem_words.append(rootWord)
print(filtered_sentence)
print(Stem_words)


lemma_word = []


wordnet_lemmatizer = WordNetLemmatizer()
for w in filtered_sentence:
    word1 = wordnet_lemmatizer.lemmatize(w, pos = "n")
    word2 = wordnet_lemmatizer.lemmatize(word1, pos = "v")
    word3 = wordnet_lemmatizer.lemmatize(word2, pos = ("a"))
    lemma_word.append(word3)
print("The text after lemmatization")
print(lemma_word)
```

Output :


File Edit View Navigate Code Refactor Run Tools VCS Window Help
pythonProject - main.py
pythonProject
main.py
train.csv
eg.txt
Run:
main
'say', 'fold', 'arm', 'sit', 'back', 'chair', ',0', '1321', 'bloody', 'blanket', 'Wait', 'tell', 'college', 'friend', 'reafs', 'bloody', 'mary', 'drama', 'cd,0', '1324',
'bloody', 'Isolated', 'City', 'In', 'World', 'Perth', "'", 'I', 'come', 'kill', 'Indians', 'FUN', "'", 'Video', 'smirk', 'remorseless', 'Pakistani', 'killer', 'show', 'boast',
'http', '//t.co/FPjLwOXKlg,1', '1326', 'bloody', 'Brazil', 'I', "'ve", 'watch', 'episode', 'S01E09', 'Bloody', 'Monday', 'http', '//t.co/vRptHvPymt', 'bloodymonday',
'tvshowtime', 'http', '//t.co/NkKvknBvOz,0', '1327', 'bloody', 'AUS', 'Marlon', 'Williams', 'gt', 'Elvis', 'Presley', 'gt', 'Marlon', 'Williams', 'gt', 'Steel', 'Panther',
'Shuffle', 'mode', 'like', 'bloody', 'legend', '``', ',0', '1328', 'bloody', 'Santa', 'Cruz', 'CA', 'Black', 'Friday', 'turn', 'Bloody', 'would', 'rather', 'shop', 'http',
'//t.co/l0pmmtZLwP', 'mystery,0', '1331', 'blow', '20up', 'Inverness', 'Nova', 'Scotia', 'ezralevant', 'Someone', 'tell', 'soccer', 'mom', 'get', 'pic', 'blow', 'put',
'daughter', "'s", 'bedroom', 'wall', 'NOT', ',0', '1332', 'blow', '20up', 'L/S/Z/L/T/H/C/H/R/A/S/C', 'Papcrdoll', 'I', 's2g', 'mention', 'get', 'blow', 'MY', 'choice', 'I',
'deactivate', 'leave', 'good', 'RESPECT', 'MY', 'CHOICES.,0', '1334', 'blow', '20up', 'Oklahoma', 'ManUtd', 'EmilymcfcHeslop', 'Wow', 'wow', 'If', 'club', 'blow', 'money',
'fuck', 'player', 'like', "'d", 'pillory', '``', ',0', '1335', 'blow', '20up', '801', 'SL', 'UT', 'Damn', 'greinke', 'get', 'blow', 'first', 'inning,0', '1337', 'blow', '20up',
',Woke', 'blow', 'lol,0', '1338', 'blow', '20up', 'fluffy', 'cloud', 'aunt', 'marge', "'re", 'blow', 'eheks,0', '1340', 'blow', '20up', 'USA', 'troylercraft', 'YEAH', 'ITS',
'NOT', 'WORTH', 'IT', 'BC', 'HE', 'ALREADY', 'HAS', 'SO', 'MANY', 'SPAMMERS', 'amp', 'HIS', 'TWITTER', 'IS', 'PROBABLY', 'BLOWN', 'UP', 'EVERY', 'SECOND,0', '1341', 'blow',
'20up', 'The', '5th', 'Dimension', 'White', 'family', 'supposedly', 'represent', 'America', "'s", 'GREAT', 'value', 'get', 'blow', 'horrible', 'CGI', 'nuclear', 'strike',
'.....', 'LMFAOOOO', ',1', '1343', 'blow', '20up', 'Georgia', 'Man', "n't", 'machinegunkelly', 'blow', 'He', "'s", 'still', 'underground.,0', '1345', 'blow', '20up', 'July',
'11th', '2015.', '``', 'And', 'mention', 'blow', 'wtf.,0', '1346', 'blow', '20up', 'The', 'Grey', 'Area', 'On', 'ThisDayInHistory', '1862', 'Confederate', 'ship', 'blow',
'crew', 'Read', 'More', 'http', '//t.co/IW7ELSzIfZ', 'via', 'History,1', '1347', 'blow', '20up', ',That', 'moment', 'ur', 'win', 'mini', 'uhc', 'get', 'blow', 'creeper', 'get',
'kick', 'fly', '....', 'So', 'salty', 'rn', ',0', '1348', 'blow', '20up', 'St', 'Paul', 'MN', 'libraryeliza', 'get', 'taylorswift13', "'bump", "'", 'approval', 'probably',
"'s", 'blow', 'http', '//t.co/KOLMzBz1pZ', 'MusicAdvisory,0', '1349', 'blow', '20up', 'Cobblestone', "'If", 'truckload', 'soldier', 'blow', 'nobody', 'panic', 'one', 'little',
'lion', 'dy', 'everyone', 'lose', 'mind', "'", 'http', '//t.co/wjNTaOkdHf', ',1', '1350', 'blow', '20up', 'ÌÏT', '30.307558', '-81.403118', 'HEY', 'LOOK', 'Kash', "'s",
'Foundation', 'Live', 'Today', 'get', 'blow', 'People', 'Magazine', "'s", 'website', 'Todd', 'Blake', 'http', '//t.co/2Fenu1SYu6', ',0', '1351', 'blow', '20up', 'Scotland',
'KaylaK369', 'get', 'last', 'month', 'I', 'go', 'EE', 'shop', 'Glad', "n't", 'blow', 'yet', 'http', '//t.co/PgB2BmCFX8,0', '1352', 'blow', '20up', 'Cosmic', 'Oneness', 'first',
'time', 'tweet', 'blow', 'almost', 'half', 'day', 'late', 'RE', 'http', '//t.co/1BUF0xM53d,0', '1353', 'blow', '20up', 'Guildford', 'UK', "'We\x89Û"re", 'blow', 'away',
'extension', 'Nothing', 'we\x89Û"ve', 'see', 'many', 'option', 'one', "'", 'http', '//t.co/0YzgW9ZbHR', 'http', '//t.co/rHtaqjvQn2,0', '1354', 'blow', '20up', 'Nowhere',
'Islands/Smash', 'Manor', 'TheBoyOfMasks', "'Thanks", 'let', 'stay', 'since', 'manor', 'blow', '.....', 'Anyways', 'buddy', "',0", '1356', 'blow', '20up', 'kisumu', 'LONER',
'DIARIES', 'The', 'pattern', 'sand', 'May', 'blow', 'away', 'The', 'photo', 'two', 'All', 'choke', 'flame', '....', 'http', '//t.co/EKfaZ6wVBz', ',0', '1360', 'blow', '20up',
'Los', 'Angeles', 'Even', 'though', 'BSG', 'sufficiently', 'hype', 'year', 'I', 'somehow', 'delay', 'watch', 'I', 'utterly', 'utterly', 'blow', 'away.,1', '1361', 'blow',
'20up', 'Making', 'Worldwide', 'Change', 'Near', 'U', 'Vanessa', "'s", 'game', 'officially', 'blow', 'LADIES', 'AND', 'GENTLEMEN', 'real', 'show', 'begin', 'BB17,0', '1362',
'blow', '20up', ',Guaranteed', 'bite', 'mutant', 'mosquito', 'ankle', 'blow', 'Little', 'cunts,0', '1363', 'blow', '20up', ',Things', 'CAN', 'pick', 'tozlet', 'seat|', 'butt',
'leprosy', 'full-blown', 'jerkface', 'syndrome', 'lateral', 'lisp', 'amp', 'toilet', 'rickets.,0', '1365', 'blow', '20up', 'KalinAndMyles', 'KalinWhite', 'ig', 'blow',
'hacker', 'I', 'need', 'stop', 'givebackkalinwhiteaccount,0', '1368', 'blow', '20up', 'ATX', 'luke_winkie', 'Whoever', 'direct', 'video', 'need', 'grab', 'Nicki', 'Minaj',
Run
TODO
Problems
Terminal
Python Packages
Python Console
Event Log
33:18
CRLF
UTF-8
4 spaces
Python 3.10 (pythonProject)


File Edit View Navigate Code Refactor Run Tools VCS Window Help
pythonProject - main.py
pythonProject
main.py
train.csv
eg.txt
Run:
main
'http', '//t.co/7enNulLKzM,0', '10833', 'wreck', 'Lincoln', 'engineshed', 'Great', 'atmosphere', 'British', 'Lion', 'gig', 'tonight', 'Hearing', 'wreck', 'http', '//t
.co/oMNBAtJEAO,0', '10834', 'wreck', 'Cramer', 'Iger', "'s", '3', 'word', 'wreck', 'Disney', "'s", 'stock', 'CNBC', 'http', '//t.co/N6RBnHMTD4,0', '10835', 'Pic', '16yr',
'old', 'PKK', 'suicide', 'bomber', 'detonate', 'bomb', 'Turkey', 'Army', 'trench', 'release', 'http', '//t.co/aaWZXykLES', 'http', '//t.co/RsMvgDxRiv,1', '10837', 'These',
'box', 'ready', 'explode', 'Exploding', 'Kittens', 'finally', 'arrive', 'gameofkittens', 'explodingkittens\x89Û_', 'http', '//t.co/TFGrAyuDC5,0', '10839', 'Calgary', 'Police',
'Flood', 'Road', 'Closures', 'Calgary', 'http', '//t.co/RLN09WKe9g,1', '10840', 'Sismo', 'DETECTADO', 'JapÍ_n', '15:41:07', 'Seismic', 'intensity', '0', 'Iwate', 'Miyagi',
'JST', 'http', '//t.co/gMoUl9zQ2Q,1', '10841', 'Sirens', 'everywhere', ',0', '10842', 'BREAKING', 'ISIS', 'claim', 'responsibility', 'mosque', 'attack', 'Saudi', 'Arabia',
'kill', '13', 'http', '//t.co/VZ640XOSwj', 'http', '//t.co/m2HpnOAK8b,1', '10843', 'Omg', 'earthquake,1', '10844', 'SEVERE', 'WEATHER', 'BULLETIN', 'No', '5', 'FOR', 'TYPHOON',
'\x89ÛÏ', 'HannaPH\x89Û\x9d', 'SOUDELOR', 'TROPICAL', 'CYCLONE', 'WARNING', 'ISSUED', 'AT', '5:00', 'PM', '06', 'http', '//t.co/tHhjJw51PE', '\x89Û_,1', '10846', 'Heat',
'wave', 'warn', 'aa', 'Ayyo', 'dei', 'Just', 'I', 'plan', 'visit', 'friend', 'year.,1', '10847', 'An', 'IS', 'group', 'suicide', 'bomber', 'detonate', 'explosives-packed',
'vest', 'mosque', 'inside', 'Saudi', 'special', 'force', 'headquarter', 'kill', '15', 'people.,1', '10848', 'I', 'hear', 'really', 'loud', 'bang', 'everyone', 'asleep',
'great,0', '10849', 'A', 'gas', 'thing', 'explode', 'I', 'hear', 'scream', 'whole', 'street', 'smell', 'gas', ',1', '10850', 'NWS', 'Flash', 'Flood', 'Warning', 'Continued',
'Shelby', 'County', '08:00', 'PM', 'Wednesday', 'http', '//t.co/nZ7ACKRrJi', 'tnwx,1', '10851', 'RT', 'LivingSafely', 'NWS', 'issue', 'Severe', 'Thunderstorm', 'Warnings',
'part', 'AR', 'NC', 'OK.', 'Expect', 'trauma', 'case', 'http', '//t.co/FWqfCKNCQW,1', '10852', 'MH370', 'Aircraft', 'debris', 'find', 'La', 'Reunion', 'miss', 'Malaysia',
'Airlines', 'http', '//t.co/5B7qT2YxdA,1', '10853', 'Father-of-three', 'Lost', 'Control', 'Car', 'After', 'Overtaking', 'Collided', 'BathAndNorthEastSomerset', 'http',
'//t.co/fa3FcnlN86,1', '10854', ',,1.3', 'Earthquake', '9Km', 'Ssw', 'Of', 'Anza', 'California', 'iPhone', 'user', 'download', 'Earthquake', 'app', 'information', 'http',
'//t.co/V3aZWOAmzK,1', '10855', 'Evacuation', 'order', 'lift', 'town', 'Roosevelt', 'http', '//t.co/EDyfo6E2PU', 'http', '//t.co/M5KxLPKFA1,1', '10859', 'break', 'LA',
'Refugio', 'oil', 'spill', 'may', 'costly', 'big', 'project', 'http', '//t.co/5ueCmcv2Pk,1', '10860', 'siren', 'go', "n't", 'Forney', 'tornado', 'warn', ',1', '10862',
'Officials', 'say', 'quarantine', 'place', 'Alabama', 'home', 'possible', 'Ebola', 'case', 'develop', 'symptom', 'http', '//t.co/rqKK15uhEY,1', '10863', 'WorldNews', 'Fallen',
'powerlines', 'G', 'link', 'tram', 'UPDATE', 'FIRE', 'crew', 'evacuate', '30', 'passenger', 'tr', 'http', '//t.co/EYSVvzA7Qm,1', '10864', 'flip', 'side', 'I', "'m", 'Walmart',
'bomb', 'everyone', 'evacuate', 'stay', 'tune', 'I', 'blow', 'not,1', '10866', 'Suicide', 'bomber', 'kill', '15', 'Saudi', 'security', 'site', 'mosque', 'Reuters', 'via',
'World', 'Google', 'News', 'Wall', 'http', '//t.co/nF4IculOje,1', '10867', 'stormchase', 'Violent', 'Record', 'Breaking', 'EF-5', 'El', 'Reno', 'Oklahoma', 'Tornado', 'Nearly',
'Runs', 'Over', 'http', '//t.co/3SICroAaNz', 'http', '//t.co/I27OaOHISp,1', '10869', 'Two', 'giant', 'crane', 'hold', 'bridge', 'collapse', 'nearby', 'home', 'http',
'//t.co/STfMbbZFB5,1', '10870', 'aria_ahrary', 'TheTawniest', 'The', 'control', 'wild', 'fire', 'California', 'even', 'Northern', 'part', 'state', 'Very', 'troubling.,1',
'10871', 'M1.94', '01:04', 'UTC', '5km', 'S', 'Volcano', 'Hawaii', 'http', '//t.co/zDtoyd8EbJ,1', '10872', 'Police', 'investigate', 'e-bike', 'collide', 'car', 'Little',
'Portugal', 'E-bike', 'rider', 'suffer', 'serious', 'non-life', 'threaten', 'injuries.,1', '10873', 'The', 'Latest', 'More', 'Homes', 'Razed', 'Northern', 'California',
'Wildfire', 'ABC', 'News', 'http', '//t.co/YmY4rSkQ3d,1']
Process finished with exit code 0
Run
TODO
Problems
Terminal
Python Packages
Python Console
Event Log
33:18
CRLF
UTF-8
4 spaces
Python 3.10 (pythonProject)