(1) Sentiment Analysis

In [ ]:

```
'''
Q1: Open "tweet_stream_halloween_1000.json" and create a list of tweets, "Tweets".
'''
```

In [1]:

```
import json
from textblob import TextBlob

infile = open('tweet_stream_halloween_1000.json')
data = json.load(infile)
infile.close()

Tweets = []

for t in data:
    Tweets.append(t['text'])
```

In [ ]:

```
'''
Q2: Create two lists (e.g. "sub_list", "pol_list") that have the subjectivity scores an
'''
```

In [2]:

```
sub_list = []
pol_list = []

for t in Tweets:
    tb = TextBlob(t)
    sub_list.append(tb.sentiment.subjectivity)
    pol_list.append(tb.sentiment.polarity)
```

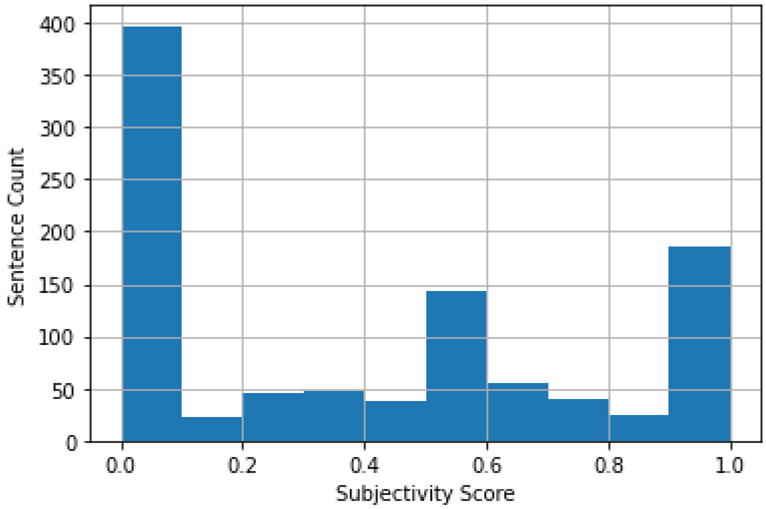
In [ ]:

```
'''
Q3: (1) Display a histogram of the subjectivy of 1K tweets.
    (2) Save the histogram as "subjectibity_1K_Tweets.pdf".
'''
```

In [3]:

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.hist(sub_list, bins = 10)
plt.xlabel('Subjectivity Score')
plt.ylabel('Sentence Count')
plt.grid(True)
plt.savefig('subjectibility_1K_Tweets.pdf')
```



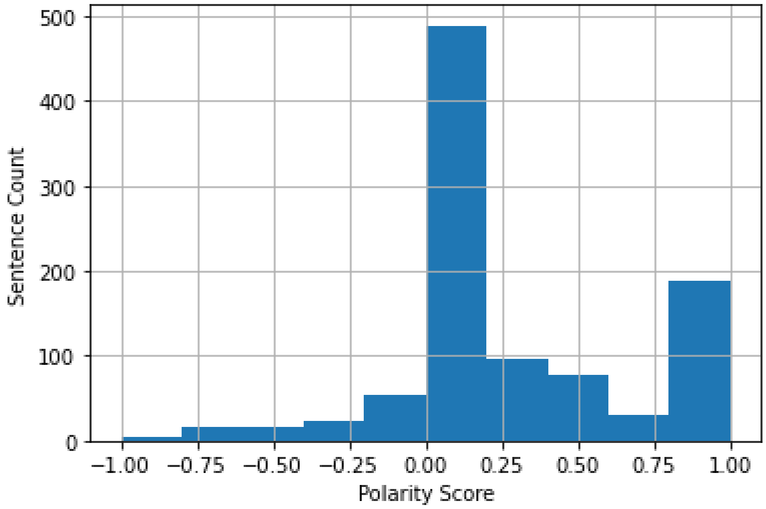
In [ ]:

```
'''
Q4. (1) Display a histogram of the polarity of 1K tweets.
    (2) Save the histogram as "polarity_1K_Tweets.pdf".
'''
```

In [4]:

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.hist(pol_list, bins = 10)
plt.xlabel('Polarity Score')
plt.ylabel('Sentence Count')
plt.grid(True)
plt.savefig('polarity_1K_Tweets.pdf')
plt.show()
```



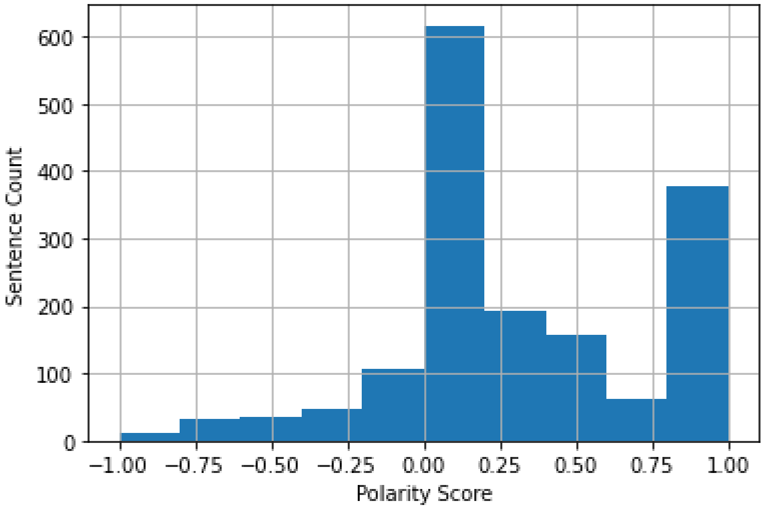
In [ ]:

```
'''
Q5: (1) Display a histogram of polarity of 1K tweets that removes objective tweets, whi
    (2) Save the histogram as "polarity_new_1K_Tweets.pdf".
'''
```

In [5]:

```
for t in Tweets:
    tb = TextBlob(t)
    if tb.sentiment.subjectivity != 0:
        sub_list.append(tb.sentiment.subjectivity)
        pol_list.append(tb.sentiment.polarity)

plt.hist(pol_list, bins = 10)
plt.xlabel('Polarity Score')
plt.ylabel('Sentence Count')
plt.grid(True)
plt.savefig('polarity_new_1K_Tweets.pdf')
```



(2) Topic Modeling (NMF)

In [ ]:

```
'''
Q6: Open "tweet_stream_easter_1000.json" and create a list of 1K tweets and the corpus_
'''
```

In [6]:

```
import json
import numpy as np
from pprint import pprint

infile = open('tweet_stream_easter_1000.json')
data = json.load(infile)
infile.close()

Tweets = []

for t in data:
    Tweets.append(t['text'])

corpus_contents = []

for t in data:
    corpus_contents.append(t['text'])
```

In [ ]:

```
'''
Q7. (1) Vectorize the 1K tweets with TfidfVectorizer.
    (2) Create a document-term matrix (i.e. "doc_term_matrix").
    (3) Create a list of unique words (i.e. "unique_words").
'''
```

In [7]:

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(stop_words = 'english', min_df = 2)

doc_term_matrix = vectorizer.fit_transform(corpus_contents)

print(doc_term_matrix)
```

```
  (0, 391)	0.35719258303447654
  (0, 1272)	0.33011729533980066
  (0, 770)	0.35719258303447654
  (0, 414)	0.199719394182866
  (0, 393)	0.075000311800004
  (0, 716)	0.2607823424827461
  (0, 1279)	0.33011729533980066
  (0, 1145)	0.366426678993408
  (0, 1045)	0.366426678993408
  (0, 1138)	0.366426678993408
  (0, 1008)	0.10269475798362519
  (1, 624)	0.117290415240031
  (1, 105)	0.2588127906220795
  (1, 25)	0.42162596479599157
  (1, 87)	0.2673940788487128
  (1, 236)	0.4454175341205642
  (1, 167)	0.4815374332321167
  (1, 444)	0.4815374332321167
  (1, 393)	0.08852949452095617
  (2, 318)	0.3210210225540688
  (2, 649)	0.28834699011063314
  (2, 253)	0.21573246803384047
  (2, 36)	0.3074600738373894
  (2, 8)	0.3074600738373894
  (2, 855)	0.28834699011063314
  :	:
  (995, 624)	0.12629385230966078
  (995, 393)	0.09532518819374594
  (996, 629)	0.44360237623448817
  (996, 839)	0.49239390911000896
  (996, 338)	0.5070700727351072
  (996, 177)	0.5250322453303855
  (996, 624)	0.13352518666405955
  (996, 393)	0.10078331854307437
  (997, 753)	0.719303976728933
  (997, 732)	0.6946954649786787
  (998, 874)	1.0
  (999, 160)	0.24928615238846075
  (999, 1260)	0.2343747921332161
  (999, 827)	0.2343747921332161
  (999, 1006)	0.2303901992953784
  (999, 533)	0.23877955516002985
  (999, 373)	0.2234062390232756
  (999, 346)	0.41960288157520076
  (999, 486)	0.21219103869704098
  (999, 1315)	0.4468124780465512
  (999, 207)	0.32470361854234475
  (999, 1104)	0.35935573791363107
  (999, 624)	0.069347902568998
  (999, 393)	0.052343021788762226
  (999, 1008)	0.07167108810244001
```

In [8]:

```python
unique_words = vectorizer.get_feature_names()
print(unique_words)
```

```
['000', '01637', '10', '10am', '10th', '11', '11th', '12', '12th', '13th', '14', '14th',
'15', '15th', '160', '17', '17th', '18', '18th', '19', '19th', '1st', '20', '200', '201
9', '20th', '21', '21st', '22nd', '247lovinglife', '25', '26', '28', '2xmki3xxre', '30',
'30mins', '30pm', '357avul9r0', '36', '40', '41kqou7pyk', '50', '5th', '67', '75', '7gkg
wmtz2h', '7th', '7x7', '84igjhetvq', '877555', '8pm', '8th', '9pm', '9quyjptbzl', '9th',
'aaamu_rys19', 'abilities', 'able', 'abroad', 'absolute', 'academy', 'accepting', 'acene
wquay', 'activities', 'actually', 'add', 'added', 'addition', 'adorable', 'adults', 'aen
focadtd', 'afternoon', 'afternoons', 'age', 'aged', 'ah', 'ain', 'air', 'al', 'aldiuk',
'alert', 'alex', 'allp1', 'alright', 'altrincham', 'altrinchambid', 'amazing', 'amp', 'a
ndrealeadsom', 'ang4vapv26', 'angryjoeshow', 'annoucement', 'announce', 'announced', 'an
nouncement', 'annual', 'answer', 'anthemsixx', 'antiqueclique', 'apostles', 'appaa_arkpr
iory', 'apply', 'approaching', 'appropriationisitc', 'apr', 'april', 'ar', 'archery', 'a
rea', 'argue', 'arranged', 'arrivals', 'arrives', 'art', 'asckennedy', 'asda', 'askchefd
ennis', 'asked', 'asking', 'atan', 'athens', 'attending', 'authorities', 'autos', 'avail
able', 'avery', 'away', 'awesome', 'aymdqivztl', 'babies', 'baby', 'bad', 'bag', 'baja
j', 'baking', 'bandcalledhappy', 'bangle', 'bankroll_chapo', 'basket', 'basketball', 'ba
skets', 'bath', 'bbloggers', 'beach', 'beaconoflight', 'beaded', 'beagle', 'beagle_boys
_', 'beautiful', 'beauty', 'begin', 'beginnings', 'believe', 'belleepoquec', 'bernardfoo
ng', 'best', 'bestprice', 'bet', 'better', 'bgea', 'biancasteinfeld', 'bible', 'big', 'b
iggest', 'bigsheepdayout', 'biking', 'billion', 'billy', 'birthday', 'bishop', 'bit', 'b
lack', 'blind', 'blossom', 'bnmarathon', 'board', 'boards', 'bonnet', 'book', 'booked',
'booking', 'books', 'bottle', 'bought', 'boulderlocavore', 'box', 'boy', 'bracelet', 'br
aids', 'break', 'breakfast', 'breaks', 'brexit', 'brighton', 'bring', 'british', 'briton
s', 'brokenshire', 'brunch', 'bu', 'buddies', 'buffet', 'build', 'building', 'bun', 'bun
dle', 'bunnies', 'bunny', 'burning', 'busy', 'buy', 'buying', 'buzzing', 'cadbury', 'cad
buryworld', 'cake', 'called', 'calling', 'camp', 'camps', 'campus', 'cancelled', 'candie
s', 'candle', 'candy', 'canvas', 'cape', 'card', 'cards', 'catholic', 'ccbasher', 'celeb
rate', 'celebrations', 'celent', 'cellent', 'centerpiece', 'centre', 'certificates', 'ch
allenge', 'challenging', 'chance', 'change', 'channel', 'chapel', 'check', 'chef', 'ches
hamutdfc', 'chick', 'chicken', 'chickenandfrog', 'chicks', 'child', 'childcare', 'childr
en', 'chocolate', 'chocolates', 'christ', 'christian', 'christians', 'christmas', 'churc
h', 'cinema', 'circ', 'circle', 'city', 'class', 'classic', 'click', 'closed', 'closer',
'closet', 'club', 'clubs', 'clue', 'clyde', 'cmd', 'code', 'cog_aic', 'collection', 'col
lects', 'college', 'color', 'colorful', 'come', 'comes', 'coming', 'comment', 'commissio
ns', 'commons', 'community', 'company', 'competition', 'complete', 'concentrating', 'con
cert', 'congratulations', 'connell', 'conosurwinetalk', 'considering', 'contouring', 'co
ntributing', 'cooking', 'cool', 'cooptywyn', 'copy', 'corner', 'correctly', 'cosigned',
'cosmetic', 'country', 'couple', 'coupons', 'course', 'courses', 'craft', 'crafts', 'cra
nmoreschool', 'crayola', 'crea', 'created', 'creating', 'creation', 'creative', 'creme',
'crib', 'cross', 'crossed', 'crying', 'crystals', 'culture', 'cunts', 'currents', 'cut
e', 'cutest', 'dagga', 'dance', 'daniellesmithtv', 'date', 'daughter', 'david', 'davidgo
ld93', 'day', 'days', 'deal', 'decided', 'decision', 'decor', 'decorate', 'decorating',
'decoration', 'decorations', 'decoupage', 'def', 'delicious', 'delighted', 'department',
'design', 'designed', 'details', 'deviled', 'did', 'different', 'dinner', 'dinosaur', 'd
is', 'dispensers', 'diy', 'dmnpys5nv7', 'does', 'dogs', 'doing', 'dollartree', 'don', 'd
oor', 'doses', 'download', 'drawn', 'dress', 'dresses', 'drive', 'drizzle', 'drizzled',
'drop', 'drug', 'duckyv72', 'duncanrighwb', 'duncanrigscndry', 'durableuk', 'ea', 'ear
l', 'early', 'eas', 'easte', 'easter', 'easter2019', 'easterbasket', 'easterbunny', 'eas
terchick', 'easterdecor', 'easterdecoration', 'easterdecorations', 'easteregg', 'eastere
gghunt', 'easterfoodandgames', 'eastergifts', 'eastern', 'eastersealsns', 'eastersunda
y', 'easy', 'eat', 'ebay', 'edition', 'education', 'educational', 'egg', 'eggmazing', 'e
ggs', 'eggwin', 'elegant', 'emoraleigh', 'emshelx', 'encouraging', 'end', 'enjoy', 'enjo
yed', 'enjoyment', 'ensure', 'enter', 'entered', 'entire', 'entirely', 'entries', 'epco
t', 'estate', 'etsy', 'eu0ob5h8d8', 'event', 'events', 'evidence', 'excited', 'exeterlaw
school', 'expected', 'experi', 'experience', 'explore', 'exploring', 'express', 'extra',
'eye', 'f8kusyd6a7', 'fabienlaine', 'fabulous', 'facebook', 'fair', 'fairy', 'faithful',
'family', 'fan', 'fanbases', 'fancy', 'fantastic', 'far', 'fashion', 'favorite', 'favour
ite', 'fb', 'fe', 'feel', 'feeling', 'felixstowe', 'festival', 'figured', 'filling', 'fi
lm', 'films', 'final', 'finally', 'fine', 'fingers', 'finished', 'fionagubelmann', 'firs
tdayofspring', 'fit', 'fleet', 'flights', 'floors', 'floral', 'florida', 'flower', 'flow
ers', 'fluffy', 'fluffychick', 'focused', 'folk', 'folkshire', 'follow', 'followvintag
e', 'food', 'fools', 'football', 'forget', 'form', 'forward', 'fr', 'fredblunt', 'free',
'french', 'fresh', 'friday', 'friend', 'friends', 'fss_uob', 'fuck', 'fuckin', 'fuckin
g', 'fun', 'fungalpeeps', 'fungi', 'funplexng', 'futsal', 'gadget', 'game', 'games', 'ga
rden', 'gardens', 'gaskellshouse', 'gearing', 'generous', 'getting', 'gift', 'gifts', 'g
in', 'girl', 'girls', 'giveaway', 'giving', 'glorious', 'glory', 'glut', 'gmt', 'god',
'goes', 'going', 'gold', 'gonna', 'good', 'goodfriday', 'goodies', 'gorgeous', 'got', 'g
otta', 'grab', 'grabbing', 'grandukholidays', 'gratitude', 'great', 'greece', 'green',
'greenwoodworker', 'greeting', 'group', 'grown', 'guess', 'guests', 'guide', 'gulliver',
'guy', 'guys', 'hair', 'half', 'halfway', 'halloween', 'hamper', 'hampers', 'hand', 'han
dmade', 'hands', 'hang', 'hanging', 'hangout', 'hantstopdaysout', 'happening', 'happen
s', 'happy', 'harlequin', 'harrisonpschool', 'hatch', 'hats', 'having', 'haydinmckenzi
e', 'hayesfinch', 'head', 'hear', 'heart', 'hello', 'help', 'helped', 'helpers', 'hemati
te', 'hi', 'hidden', 'high', 'hire', 'historical', 'ho', 'holder', 'holding', 'holiday',
'holidaying', 'holidays', 'holiest', 'holy', 'home', 'hoop', 'hop', 'hope', 'hopefully',
'hoppity', 'hoppy', 'host', 'hosting', 'hot', 'hotel', 'house', 'https', 'hundreds', 'hu
nt', 'ic5j0efmb9', 'icing', 'idea', 'ideas', 'im', 'immsnorfolk', 'important', 'includ
e', 'including', 'indicates', 'info', 'information', 'ing', 'initial', 'innovative', 'in
oculated', 'insert', 'inspiration', 'intended', 'interested', 'interfaith', 'invention',
'invited', 'inviting', 'involved', 'island', 'isn', 'italy', 'item', 'items', 'itsfamous
joe', 'itsindysev', 'james', 'je', 'jesus', 'jewelry', 'jewelrywiz', 'jlap64', 'join',
'josephdemauro1', 'jsismee', 'june', 'junior', 'just', 'kasoasmfanbase', 'kasson_wvu',
'keke', 'kellybullad', 'kelsey', 'keoshere', 'khloekardashian', 'ki', 'kid', 'kids', 'ki
mberlilarae', 'kind', 'kitchen', 'kitchensanc2ary', 'kiwitraveleu', 'kiwitraveleuhotel
s', 'knitted_niceties', 'knittedniceties', 'know', 'kz8j2ikshj', 'labradorite', 'lady',
'land', 'landed', 'large', 'later', 'latest', 'launched', 'launches', 'launching', 'lav
a', 'lawrence', 'lbz2', 'leads', 'learn', 'leave', 'leestrobel', 'left', 'lego', 'lemo
n', 'lent', 'lesson', 'let', 'life', 'lights', 'like', 'likes', 'limited', 'lindenrdoa
k', 'link', 'lions', 'list', 'literally', 'little', 'live', 'liverpool', 'living', 'll',
'loads', 'local', 'logo', 'lol', 'lolitastarr95', 'london', 'long', 'look', 'looking',
'looks', 'lord', 'lot', 'lots', 'loudly', 'love', 'loved', 'lovely', 'low', 'lucky', 'lu
nch', 'luxury', 'magic', 'magodo', 'make', 'makes', 'making', 'man', 'manchester', 'marc
h', 'marwa', 'massive', 'materials', 'maxsportshangout', 'mean', 'means', 'mediaatbright
on', 'meet', 'members', 'memorial', 'men', 'mens', 'menu', 'message', 'messi', 'messy',
'metanoia', 'mgrant76308', 'mi', 'microbiology', 'mini', 'minute', 'miss', 'misssdeaki
n', 'mistressd25', 'mmr', 'mom', 'mommyblogger', 'momtograndma', 'monday', 'money', 'mon
th', 'morning', 'morrisons', 'mosaic', 'mother', 'mothersday', 'movie', 'moviessssss',
'moving', 'mr', 'muddy', 'multi', 'museum', 'music', 'nature', 'nc', 'near', 'nearly',
'necklace', 'need', 'nejfrvma5f', 'nest', 'netball', 'new', 'newquay', 'newrelease', 'ne
ws', 'newsanddeli', 'newsletter', 'nice', 'nigeria', 'night', 'nin', 'non', 'normal', 'n
orthants', 'note', 'novelties', 'nursery', 'o1rr8onlpk', 'offer', 'office', 'officiall
y', 'oh', 'ok', 'ola', 'old', 'olympics', 'omg', 'ones', 'online', 'open', 'opened', 'op
ening', 'opportunity', 'orange', 'order', 'ordered', 'orders', 'ou', 'outdoors', 'oxpast
urehotel', 'paces', 'pacismultiplex', 'page', 'paper', 'parade', 'paradigm', 'paramour',
'parents', 'paris', 'parishpriest', 'park', 'parliament', 'party', 'paschal', 'pass', 'p
assionately', 'passover', 'pastel', 'patfurstenberg', 'pattern', 'pay', 'peascommunity',
'peeled', 'peeps', 'pendant', 'people', 'peperhade', 'perf', 'perfect', 'personalised',
'personalized', 'peston', 'pez', 'phone', 'photo', 'photos', 'piagg', 'pieces', 'pink',
'pippin143', 'pitcher', 'place', 'places', 'planning', 'plans', 'plastic', 'plate', 'pla
y', 'players', 'playing', 'plenty', 'plot', 'plus', 'plush', 'pm', 'point', 'police', 'p
olitical', 'pool', 'pop', 'poshmark', 'poshmarkapp', 'positive', 'post', 'poster', 'post
ers', 'pouch', 'prayer', 'pre', 'premium', 'prepare', 'presents', 'pretty', 'preview',
'price', 'primary', 'princess', 'printable', 'prioritize', 'prize', 'prizes', 'problemat
ic', 'product', 'professing', 'profile', 'program', 'programme', 'project', 'promo', 'pr
osecco', 'prototypi', 'proud', 'provide', 'provision', 'pta', 'pun', 'pupils', 'purchas
e', 'purenogluten', 'purse', 'qu', 'question', 'quick', 'quickly', 'quiet', 'quite', 'ra
bbit', 'rabbits', 'raffle', 'raise', 'raleigh', 'range', 'rare', 'read', 'reading', 'rea
dy', 'real', 'realized', 'really', 'rec', 'recently', 'recess', 'recipe', 'recomm', 're
d', 'redwoodphotos', 'register', 'released', 'reliefs', 'religion', 'remain', 'remembe
r', 'reminded', 'reminder', 'renewal', 'reservations', 'reserve', 'residents', 'respec
t', 'resur', 'resurr', 'resurrection', 'retired', 'return', 'retweets', 'rev', 'revelaio
```

```
ns', 'rfl9yxq0il', 'ridiculous', 'right', 'rip', 'road', 'rolled', 'romans', 'ronaldo',
'room', 'roses', 'rt', 'rugby', 'ruined', 'rule', 'run', 'running', 'ruthnotman', 'safcf
ol', 'said', 'sainsburys', 'sale', 'samkellymusic', 'sammie', 'santa', 'santaclausvsthee
asterbunny', 'saturday', 'save', 'saved', 'saw', 'say', 'saying', 'sazmeister88', 'sbeth
caplin', 'scdiunbkwh', 'school', 'schools', 'scrapped', 'screen', 'screenshot', 'screens
tudies', 'seale15eastcoast', 'seals', 'search', 'seaside', 'season', 'seat', 'second',
'secret', 'seder', 'seders', 'selling', 'send', 'sending', 'sentiment', 'series', 'serio
usly', 'service', 'serving', 'session', 'set', 'sets', 'sha', 'shaped', 'share', 'share
d', 'shazam', 'sheep', 'sheila', 'shift', 'shifted', 'shire', 'shit', 'shoe', 'shop', 's
hoplocal', 'shopmycloset', 'shops', 'shows', 'shun', 'sign', 'signed', 'silver', 'simpl
y', 'site', 'size', 'skills', 'slfacademy', 'sm2daworld', 'small', 'soccer', 'social',
'sold', 'soldering', 'som', 'songs', 'soon', 'sorts', 'spaces', 'special', 'spending',
'spent', 'sport', 'sporting', 'sports', 'spot', 'spread', 'spring', 'springy', 'square',
'st', 'staff', 'staniford', 'start', 'start360org', 'started', 'starting', 'starts', 'st
erling', 'stitch', 'stjohnfarnworth', 'stoaters', 'stone', 'stop', 'story', 'strike', 's
trip', 'strong', 'students', 'stuff', 'stuffed', 'stunning', 'style', 'su', 'successfu
l', 'sugar', 'sullivan', 'summer', 'sun', 'sunday', 'sunrise', 'super', 'superpowers',
'support', 'sure', 'sweet', 'switch', 'symbol', 'table', 'tadi', 'tags', 'takeover', 'ta
king', 'talking', 'tastes', 'tatehxppy', 'te', 'teaching', 'team', 'tech', 'techcampuk',
'teens', 'tell', 'tennis', 'term', 'tesco', 'testimony', 'tgdsquad', 'thank', 'thanks',
'thatsnotmy', 'the_christian_movies_', 'theatre', 'themed', 'theradr', 'thing', 'thing
s', 'think', 'thinking', 'tho', 'thought', 'thoughtful', 'thriftythistle', 'throwback',
'thursday', 'ti', 'ticket', 'tickets', 'tiger', 'tightly', 'til', 'till', 'time', 'time
s', 'timetable', 'tlop8fl6yt', 'today', 'toddler', 'toiletry', 'told', 'ton', 'tonight',
'took', 'tooth', 'topped', 'toppsta', 'toronto', 'touch', 'tourism', 'town', 'toy', 'tpd
esigns1', 'tracycampanell', 'traditional', 'traditions', 'trail', 'trailer', 'trampalin
a', 'travel', 'treat', 'treats', 'tree', 'trials', 'trick', 'trip', 'try', 'tuesday', 't
une', 'turnout', 'tweens', 'tweet', 'twitter', 'uglydolls', 'undecided', 'uni', 'uniqu
e', 'unitary', 'unless', 'uobmedia', 'upcoming', 'update', 'urhvg43two', 'use', 'using',
'v4qivynpuc', 'vacccine', 'valley', 'varied', 've', 'vegan', 'video', 'view', 'vintage',
'visit', 'visitblackpool', 'visiting', 'volume', 'volunteers', 'voucher', 'vpjffigwzb',
'wae', 'wait', 'waiting', 'wall', 'wanna', 'want', 'wanted', 'war', 'warning', 'watch',
'watched', 'watching', 'water', 'watercolor', 'wax', 'way', 'waynehotel', 'wdyt', 'wean
s', 'wear', 'wednesday', 'week', 'weekend', 'weeks', 'welcome', 'went', 'whatsonbrum',
'wheels', 'whimsyandwicked', 'whpiauaknj', 'wild', 'willing', 'win', 'winds', 'wine', 'w
inner', 'winners', 'winning', 'winter', 'wit', 'wnukrt', 'woman', 'womaninbiz', 'women',
'womeninbiz', 'won', 'wonder', 'wondered', 'wonderful', 'wondering', 'wood', 'word', 'wo
rds', 'work', 'works', 'world', 'worship', 'wow', 'wreath', 'wunderground', 'x5gx9nvrc
h', 'ya', 'yatetownfc', 'yeah', 'year', 'years', 'yellow', 'yes', 'yesterday', 'yl5albog
qc', 'yo', 'young', 'youth', 'youtube', 'yr4', 'yrs', 'zingy', 'zipper']
```

In [ ]:

```
'''
Q8: (1) Perform NMF decomposition using a document-term matrix with TfidfVectorizer.
    (2) Set the number of topics to 7.
    (3) Create a document-topic matrix (i.e. "doc_top_matrix") and a topic-term matrix
'''
```

In [9]:

```
from sklearn import decomposition

num_topics = 7

clf = decomposition.NMF(n_components = num_topics)

doc_top_matrix = clf.fit_transform(doc_term_matrix)

print(doc_top_matrix)
```

```
[[0.05037002 0.00360844 0.0043616  ... 0.         0.01844328 0.        ]
 [0.04742787 0.         0.         ... 0.         0.05042501 0.06726873]
 [0.01616372 0.00182809 0.00388993 ... 0.00075566 0.09142365 0.        ]
 ...
 [0.00434349 0.         0.         ... 0.         0.         0.        ]
 [0.00341369 0.0003687  0.         ... 0.         0.         0.        ]
 [0.01100788 0.00066708 0.         ... 0.26117198 0.         0.        ]]
C:\Users\Blake\anaconda3\lib\site-packages\sklearn\decomposition\_nmf.py:312: FutureWarn
ing: The 'init' value, when 'init=None' and n_components is less than n_samples and n_fe
atures, will be changed from 'nndsvd' to 'nndsvda' in 1.1 (renaming of 0.26).
  warnings.warn(("The 'init' value, when 'init=None' and "
```

In [10]:

```
top_term_matrix = clf.components_

print(top_term_matrix)
```

```
[[4.92025794e-03 9.73287149e-04 5.10370323e-02 ... 9.73287149e-04
  1.06346265e-03 1.43078060e-03]
 [0.00000000e+00 4.18793924e-03 0.00000000e+00 ... 4.18793924e-03
  6.47488478e-04 1.10722297e-04]
 [4.87472834e-04 5.13604817e-04 0.00000000e+00 ... 5.13604817e-04
  7.51528481e-04 2.97642846e-04]
 ...
 [0.00000000e+00 1.44461940e-05 0.00000000e+00 ... 1.44461940e-05
  5.68394388e-05 0.00000000e+00]
 [2.31659818e-02 1.44826080e-03 2.79801805e-02 ... 1.44826080e-03
  2.85970475e-03 2.61888186e-04]
 [0.00000000e+00 6.53785687e-03 1.72377503e-02 ... 6.53785687e-03
  0.00000000e+00 0.00000000e+00]]
```

In [ ]:

```
'''
Q9: Display the constructed topics with the top 6 keywords.
'''
```

In [15]:

```
import numpy as np
from pprint import pprint

num_top_words = 6

np.argsort(topic_1)[-num_top_words:][::-1]

print(unique_words[624], unique_words[393], unique_words[1008], unique_words[414], uniq
```

```
https easter rt egg eggs make
```

In [16]:

```
topic_words = []

for topic in clf.components_:

    word_idx = np.argsort(topic)[-num_top_words:]

    temp_lst = []
    for idx in word_idx[::-1]:
        temp_lst.append(unique_words[idx])

    topic_words.append(temp_lst)

pprint(topic_words)
```

```
[['https', 'easter', 'rt', 'basket', 'make', 'great'],
 ['grown', 'shazam', 'loudly', 'argue', 'superpowers', 'passionately'],
 ['getting',
  'passover',
  'theradr',
  'problematic',
  'appropriationisitc',
  'seders'],
 ['ve', 'got', 'kasson_wvu', 'different', 'inoculated', 'called'],
 ['bunny', 'themed', 'available', '84igjhetvq', 'undecided', 'labradorite'],
 ['april', 'just', 'win', 'follow', 'break', 'tweet'],
 ['coming', 'amp', 'football', 'poster', 'trials', 'education'],
 ['wreath', 'spring', 'decor', 'floral', 'bunny', 'door'],
 ['womaninbiz', 'wnukrt', 'tpdesigns1', 'order', 'days', 'holder'],
 ['egg', 'easter', 'eggs', 'hunt', 'love', 'question']]
```

In [ ]:

```
'''
Q10: Perform topic modeling with "tweet_stream_easter_1000.json". Set the number of top
'''
```

In [13]:

```
import json
import numpy as np
from pprint import pprint
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import decomposition
import numpy as np

infile = open('tweet_stream_easter_1000.json')
data = json.load(infile)
infile.close()

corpus_contents = []

for t in data:
    corpus_contents.append(t['text'])

vectorizer = TfidfVectorizer(stop_words = 'english', min_df = 2)
doc_term_matrix = vectorizer.fit_transform(corpus_contents)

unique_words = vectorizer.get_feature_names()

num_topics = 10

clf = decomposition.NMF(n_components = num_topics)
doc_top_matrix = clf.fit_transform(doc_term_matrix)
top_term_matrix = clf.components_

topic_words = []
num_top_words = 10

for topic in clf.components_:
    word_idx = np.argsort(topic)[-num_top_words: ]
    temp_lst = []
    for idx in word_idx[::-1]:
        temp_lst.append(unique_words[idx])
    topic_words.append(temp_lst)

pprint(topic_words)
```

```
[['https',
  'easter',
  'rt',
  'basket',
  'make',
  'great',
  'sunday',
  'good',
  'holiday',
  'ideas'],
 ['grown',
  'shazam',
  'loudly',
  'argue',
  'superpowers',
  'passionately',
  'old',
  'men',
  'boy',
  '14'],
 ['getting',
  'passover',
  'theradr',
  'problematic',
  'appropriationisitc',
  'seders',
  'closer',
  'reminder',
  'christian',
  'rt'],
 ['ve',
  'got',
  'kasson_wvu',
  'different',
  'inoculated',
  'called',
  'microbiology',
  'fungalpeeps',
  'fungi',
  'successful'],
 ['bunny',
  'themed',
  'available',
  '84igjhetvq',
  'undecided',
  'labradorite',
  'update',
  'date',
  'haydinmckenzie',
  'rt'],
 ['april',
  'just',
  'win',
  'follow',
  'break',
  'tweet',
  '18th',
  'winner',
  'luxury',
  'drawn'],
 ['coming',
  'amp',
  'football',
  'poster',
  'trials',
  'education',
  'details',
  'academy',
  'easter',
  'https'],
 ['wreath',
  'spring',
  'decor',
  'floral',
  'bunny',
  'door',
  'wall',
  'nursery',
  'room',
  'biancasteinfeld'],
 ['womaninbiz',
  'wnukrt',
  'tpdesigns1',
  'order',
  'days',
  'holder',
  'boards',
  '7gkgwmtz2h',
  'breakfast',
  'chance'],
 ['egg',
  'easter',
  'eggs',
  'hunt',
  'love',
  'question',
  'chocolate',
  'giveaway',
  'like',
  'rt']]
C:\Users\Blake\anaconda3\lib\site-packages\sklearn\decomposition\_nmf.py:312: FutureWarn
ing: The 'init' value, when 'init=None' and n_components is less than n_samples and n_fe
atures, will be changed from 'nndsvd' to 'nndsvda' in 1.1 (renaming of 0.26).
  warnings.warn(("The 'init' value, when 'init=None' and "
```