

▼ Stance Detection for the Fake News Challenge

Identifying Textual Relationships with Deep Neural Nets

Check the problem context [here](#).

Download files required for the project from [here](#).

▼ Step1: Load the given dataset

1. Mount the google drive
2. Import Glove embeddings
3. Import the test and train datasets

▼ Mount the google drive to access required project files

Run the below commands

```
from google.colab import drive
```

```
drive.mount('/content/drive/')
```



Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=9473189

Enter your authorization code:

.....

Mounted at /content/drive/

▼ Path for Project files on google drive

Note: You need to change this path according where you have kept the files in google drive.

```
project_path = "/content/drive/fake news challenge/"
```

▼ Loading the Glove Embeddings

```
from zipfile import ZipFile  
with ZipFile(project_path+'glove.6B.zip', 'r') as z:  
    z.extractall()
```

▼ Load the dataset [5 Marks]

1. Using [read_csv\(\)](#) in pandas load the given train datasets files `train_bodies.csv` and `train_stances.csv`.
2. Using [merge](#) command in pandas merge the two datasets based on the Body ID.

Note: Save the final merged dataset in a dataframe with name `dataset`.

```
import pandas as pd
```

```
train_bodies = pd.read_csv('/content/drive/fake news challenge/train_bodies.csv')
train_stances = pd.read_csv('/content/drive/fake news challenge/train_stances.csv')
```

```
dataset = pd.merge(train_bodies , train_stances , on = 'Body ID')
```

```
dataset.tail()
```



	Body ID	articleBody
49967	2532 ANN ARBOR, Mich. – A pizza delivery man in Mic...	Pizza delivery man gets tipped mc
49968	2532 ANN ARBOR, Mich. – A pizza delivery man in Mic...	Pizza delivery man
49969	2532 ANN ARBOR, Mich. – A pizza delivery man in Mic...	Luckiest Pizza Delivery Guy Ever (
49970	2532 ANN ARBOR, Mich. – A pizza delivery man in Mic...	Ann Arbor pizza delivery driver :
49971	2532 ANN ARBOR, Mich. – A pizza delivery man in Mic...	Ann Arbor pizza delivery driver :

Check1:

You should see the below output if you run `dataset.head()` command as given b

```
dataset.head()
```



	Body ID	articleBody
0	0	A small meteorite crashed into a wooded area i... Soldier shot, Parliament locked down a
1	0	A small meteorite crashed into a wooded area i... Tourist dubbed 'Spider Man' after spid
2	0	A small meteorite crashed into a wooded area i... Luke Somers 'killed in failed rescue a
3	0	A small meteorite crashed into a wooded area i... BREAKING: Soldier shot at War Memorial
4	0	A small meteorite crashed into a wooded area i... Giant 8ft 9in catfish weighing 19 stone

▼ Step2: Data Pre-processing and setting some hyper parameters needed

Run the code given below to set the required parameters.

1. MAX_SENTS = Maximum no.of sentences to consider in an article.
2. MAX_SENT_LENGTH = Maximum no.of words to consider in a sentence.
3. MAX_NB_WORDS = Maximum no.of words in the total vocabualry.
4. MAX_SENTS_HEADING = Maximum no.of sentences to consider in a heading of an article.

```
MAX_NB_WORDS = 20000
MAX_SENTS = 20
MAX_SENTS_HEADING = 1
MAX_SENT_LENGTH = 20
VALIDATION_SPLIT = 0.2
```

▼ Download the Punkt from nltk using the commands given below. This is for sente

For more info on how to use it, read [this](#).

```
import nltk
nltk.download('punkt')
```



```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

Tokenizing the text and loading the pre-trained Glove word embeddings [5 marks]

Keras provides [Tokenizer API](#) for preparing text. Read it before going any further.

Import the Tokenizer from keras preprocessing text

```
from keras.preprocessing.text import Tokenizer
```



Using TensorFlow backend.

Initialize the Tokenizer class with maximum vocabulary count as MAX_NB_WORDS initialized

```
tk = Tokenizer(MAX_NB_WORDS)
```

Now, using fit_on_texts() from Tokenizer class, lets encode the data

Note: We need to fit articleBody and Headline also to cover all the words.

```
codetext = dataset['articleBody'] + dataset['Headline']  
tk.fit_on_texts(codetext)
```

fit_on_texts() gives the following attributes in the output as given [here](#).

- **word_counts:** dictionary mapping words (str) to the number of times they appeared on during fit called.
- **word_docs:** dictionary mapping words (str) to the number of documents/texts they appeared or fit_on_texts was called.
- **word_index:** dictionary mapping words (str) to their rank/index (int). Only set after fit_on_texts was called.
- **document_count:** int. Number of documents (texts/sequences) the tokenizer was trained on. Or fit_on_sequences was called.

```
tk.word_counts
```



```
OrderedDict([('a', 956997),
             ('small', 10498),
             ('meteorite', 11621),
             ('crashed', 414),
             ('into', 51208),
             ('wooded', 1108),
             ('area', 14264),
             ('in', 816158),
             ('nicaragua's', 1052),
             ('capital', 13734),
             ('of', 883546),
             ('managua', 3656),
             ('overnight', 1498),
             ('the', 2258230),
             ('government', 47284),
             ('said', 268912),
             ('sunday', 10832),
             ('residents', 6856),
             ('reported', 43852),
             ('hearing', 5028),
             ('mysterious', 1706),
             ('boom', 2918),
             ('that', 468860),
             ('left', 23286),
             ('16', 4344),
             ('foot', 4882),
             ('deep', 2436),
             ('crater', 7366),
             ('near', 25876),
             ('city's', 1452),
             ('airport', 11184),
             ('associated', 9164),
             ('press', 18783),
             ('reports', 48671),
             ('spokeswoman', 4920),
             ('rosario', 982),
             ('murillo', 2024),
             ('committee', 4834),
             ('formed', 1300),
             ('by', 209484),
             ('to', 999445),
             ('study', 2864),
             ('event', 9724),
             ('determined', 2468),
             ('it', 251804),
             ('was', 360153),
             ('relatively', 1406),
             ('appears', 14184),
             ('have', 220829),
             ('come', 17153),
             ('off', 31122),
             ('an', 182403),
             ('asteroid', 5400),
             ('passing', 2508),
             ('close', 9906),
             ('earth', 6664),
             ('house', 23593),
             ...])
```

```
('sized', 1148),
('2014', 21928),
('rc', 1806),
('which', 81342),
('measured', 456),
('60', 2974),
('feet', 4230),
('diameter', 142),
('skimmed', 72),
('this', 139660),
('weekend', 7022),
('abc', 1016),
('news', 59265),
('nicaragua', 3270),
('will', 89301),
('ask', 4810),
('international', 13726),
('experts', 6669),
('help', 13346),
('local', 17494),
('scientists', 2888),
('understanding', 864),
('what', 56873),
('happened', 10314),
('had', 138500),
('radius', 396),
('39', 2198),
('and', 767174),
('depth', 670),
('humberto', 1276),
('saballos', 692),
('volcanologist', 664),
('with', 259832),
('nicaraguan', 4822),
('institute', 4014),
('territorial', 1428),
('studies', 3074),
('who', 141398),
('on', 359247),
('he', 263116),
('is', 377173),
('still', 26658),
('not', 164901),
('clear', 9446),
('if', 60717),
('disintegrated', 1110),
('or', 83914),
('buried', 2768),
('garcia', 594),
('astronomy', 574),
('center', 8356),
('at', 195759),
('national', 15268),
('autonomous', 1244),
('university', 6511),
('could', 57258),
('be', 187634),
('related', 6586),
```

```

('forecast', 934),
('pass', 2204),
('planet', 1038),
('saturday', 9642),
('night', 16696),
('we', 97230),
('more', 72155),
('because', 34188),
('ice', 746),
('rock', 2536),
('wilfried', 788),
('strauch', 1286),
('adviser', 1022),
('very', 22680),
('strange', 3694),
('no', 69053),
('one', 85557),
('streak', 562),
('light', 3836),
('anyone', 10047),
('has', 200056),
('photo', 18089),
('something', 17228),
('loud', 2850),
('but', 141940),
('they', 122392),
("didn't", 7356),
('see', 25252),
('anything', 5804),
('sky', 2366),
('i', 97909),
('sitting', 2396),
('my', 36022),
('porch', 580),
('saw', 8782),
('nothing', 7808),
('then', 40876),
('all', 68414),
('sudden', 1940),
('heard', 18670),
('large', 9992),
('blast', 2650),
('thought', 14022),
('bomb', 2366),
('felt', 5866),
('expansive', 864),
('wave', 1784),
('jorge', 910),
('santamaria', 746),
('told', 70926),
('site', 17930),
("managua's", 1028),
('air', 26151),
('force', 10832),
('base', 2610),
('only', 32418),
('journalists', 5674),
('from', 105250)

```

```

('from', 133333),
('state', 76933),
('media', 42334),
('were', 97382),
('allowed', 2810),
('visit', 4360),
('last', 47992),
('week', 24982),
('hinted', 320),
('as', 213222),
('ebola', 25012),
('fears', 3410),
('spread', 6048),
('across', 15164),
('america', 7484),
('today', 13811),
('get', 26892),
('confirmation', 3976),
('daily', 14732),
('caller', 456),
('passenger', 1615),
('dulles', 644),
('outside', 10708),
('washington', 12838),
('d', 1234),
('c', 3298),
('apparently', 10790),
('taking', 10532),
('any', 42940),
('chances', 1126),
('female', 6891),
('dressed', 3766),
('hazmat', 576),
('suit', 1392),
('complete', 2066),
('full', 10370),
('body', 10056),
('gown', 252),
('mask', 926),
('gloves', 1086),
('spotted', 3074),
('wednesday', 16842),
('waiting', 3010),
('for', 295417),
('flight', 2642),
('source', 20727),
('particularly', 3008),
('liked', 632),
('jcpenny', 2),
('bag', 926),
('maybe', 3190),
('that's', 6297),
('new', 70021),
('business', 6486),
('line', 5880),
('bankrupt', 60),
('retailer', 332),
('side', 5394),

```



```
('note', 3732),
('try', 5014),
('halloween', 174),
('stores', 3104),
('you', 62422),
('need', 7164),
('haz', 2),
('mat', 2),
('hurry', 18),
('newser', 312),
('-', 22526),
('wonder', 2824),
('how', 29013),
('long', 17430),
('quarter', 2612),
('pounder', 1068),
('cheese', 5256),
('can', 40212),
('two', 46233),
('australians', 400),
('say', 30858),
('bought', 3948),
('few', 13216),
('mcdonald's', 1402),
('burgers', 424),
('friends', 14958),
('back', 28536),
('1995', 262),
('when', 71430),
('teens', 1700),
('never', 17860),
('showed', 8870),
('up', 75785),
('so', 49238),
('kid's', 112),
('burger', 3864),
('went', 22420),
('uneaten-and', 112),
('stayed', 1072),
('way', 20708),
('australia's', 196),
('network', 6060),
('we're', 3062),
('pretty', 5520),
('sure', 9234),
('it's', 22508),
('oldest', 1426),
('world', 26226),
('says', 39352),
('men', 11612),
('casey', 566),
('dean', 1058),
('holding', 7948),
('onto', 776),
('their', 72568),
('friend', 10416),
('started', 9798),
('joke', 1940),
```

```

('adds', 2028),
('months', 13374),
('became', 6136),
('years', 30036),
('now', 40018),
('20', 10330),
('later', 19468),
('looks', 4690),
('same', 16282),
('did', 25140),
('day', 24228),
('perfectly', 1474),
('preserved', 528),
('its', 61066),
('original', 5237),
('wrapping', 250),
('his', 227368),
('buying', 1680),
('mate', 394),
('eduard', 556),
('nitz', 550),
('even', 34654),
('took', 18894),
('australian', 2344),
('tv', 13008),
('show', 18020),
('project', 4034),
('mold', 206),
('free', 10650),
('specimen', 668),
('9', 12608),
('pair', 2410),
('offered', 4074),
('take', 16524),
('bite', 1022),
('charity', 1970),
('dissuaded', 112),
('show's', 112),
('hosts', 552),
('they've', 1856),
('also', 71848),
('facebook', 16652),
('page', 10682),
('called', 16628),
('year', 55848),
('old', 32506),
('likes', 1116),
('than', 41632),
('kanye', 1236),
('west', 11634),
('4', 8810),
('044', 112),
('writing', 3320),
('they're', 4666),
('selling', 2950),
('itunes', 1094),
('song', 1996),
('1', 15404)

```

```
( 1 , 15474),
('69', 420),
('giving', 3028),
('proceeds', 536),
('beyond', 1724),
('blue', 1936),
('helps', 1210),
('battle', 3340),
('anxiety', 496),
('depression', 236),
('ago', 16838),
('man', 59271),
('sold', 5396),
('bottle', 1120),
('mcjordan', 112),
('sauce', 116),
('10', 18877),
('000', 26910),
("here's", 2556),
('why', 11060),
('mickey', 112),
("d's", 112),
('food', 10128),
('seemingly', 1770),
('decays', 112),
('posting', 1392),
('photos', 7872),
('gun', 4712),
('toting', 2),
('child', 9578),
('online', 15568),
('isis', 82290),
('supporters', 2112),
('announced', 12172),
('group's', 2868),
('youngest', 1866),
('soldier', 7116),
('died', 13750),
('combat', 1590),
('twitter', 26165),
('accounts', 4118),
('linked', 5144),
('islamic', 59282),
('iraq', 37793),
('al', 54431),
('sham', 424),
('claimed', 18150),
('“got', 8),
('martyred"', 2),
('father', 15844),
('while', 43176),
('fighting', 11348),
('terrorist', 10505),
('group', 52210),
('syria', 47731),
('posted', 18652),
('smiling', 1832),
('boy', 10558),
```

```
('military', 24558),
('fatigues', 1050),
('weapons', 13702),
('times', 20276),
('are', 152155),
('almost', 8502),
('british', 23524),
('roughly', 2452),
('first', 49665),
('emerged', 4150),
('june', 5764),
('charlie', 2146),
('cooper', 1096),
('researcher', 1082),
('monitors', 1096),
('social', 15698),
('london', 14002),
('based', 15538),
('quilliam', 482),
('counter', 3768),
('extremism', 692),
('think', 16868),
('tank', 1954),
('past', 14118),
('mr', 24622),
('noticed', 2024),
('hashtag', 1492),
('“shibal', 2),
('albaghdadi”', 2),
('-', 46876),
('translates', 696),
('“the', 13964),
('cub', 796),
('baghdadi”', 2),
('fighters', 24814),
('commonly', 1376),
('refer', 474),
('themselves', 5250),
('lions', 120),
('soldiers', 8798),
('cubs', 70),
('abu', 9682),
('bahr', 5770),
('baghdadi', 10991),
('isis's', 352),
('self', 2412),
('proclaimed', 1330),
('caliph', 438),
('ubaidah', 942),
('fighter', 5682),
('got', 12638),
('martyred', 178),
('us', 45553),
('airstrikes', 14214),
('2', 16247),
('weeks', 12276),
('http', 3466),
('t', 4942),
```

```

('co', 9754),
('n1puexi7ir', 2),
('istimes1', 2),
('october', 6920),
('07', 142),
('tweeting', 432),
('young', 11848),
('dead', 16308),
('sept', 5068),
('26', 5272),
('often', 6500),
('produces', 552),
('fabricated', 734),
('make', 21148),
('seem', 4132),
('“more', 208),
('brutal', 4840),
('”', 142237),
('appear', 5368),
('coming', 11780),
('directly', 2490),
('account', 9956),
('killed', 35490),
('u', 51640),
('s', 51676),
('strike', 5946),
('though', 16790),
('been', 150084),
('confirmed', 24354),
('identified', 11570),
('“it', 5660),
('seems', 7220),
('like', 41816),
('legitimate', 560),
('thing', 8942),
('telephone', 780),
('interview', 8722),
('thursday', 12222),
('adding', 5538),
('known', 21700),
('use', 14926),
('“i', 13162),
('would', 72617),
('vouch', 70),
('united', 14030),
('nations', 2794),
('report', 39827),
('month', 21536),
('children', 10412),
('conduct', 964),
('patrols', 212),
('arrest', 5846),
('guard', 5750),
('prisoners', 1976),
('carry', 2244),
('forced', 4964),
('give', 8277),
('“it', 5660)

```

```
( '01000', 5228),  
( 'injured', 4596),  
( 'allegedly', 12714),  
( 'listed', 1350),  
( 'un', 17892),  
( '12', 16802),  
( '13', 7416),  
( 'frequently', 1342),  
( 'used', 18464),  
( 'propaganda', 2842),  
( 'showing', 8536),  
( '"uniform', 2),  
( 'parading', 70),  
( 'alongside', 4196),  
( 'adults', 226),  
( 'being', 48836),  
( '"everyone', 62),  
( 'nato', 572),  
( 'headquarters', 3104),  
( 'worried', 1778),  
( 'about', 88388),  
( 'shelly', 90),  
( 'whitman', 70),  
( 'executive', 3798),  
( 'director', 10528),  
( 'halifax', 70),  
( 'romeo', 126),  
( 'dallaire', 70),  
( 'initiative', 218),  
( '"if', 2726),  
( 'canada', 3690),  
( 'send', 1803),  
( 'boots', 554),  
( 'ground', 8536),  
( 'going', 22954),  
( 'face', 10194),  
( 'she', 80204),  
( 'currently', 7846),  
( 'six', 18448),  
( 'canadian', 13553),  
( 'mission', 3064),  
( 'limited', 2706),  
( 'strikes', 10826),  
( 'ban', 1018),  
( 'deploying', 384),  
( 'troops', 4392),  
( 'least', 20842),  
( '25', 6266),  
( 'suspected', 7858),  
( 'boko', 27449),  
( 'haram', 26452),  
( 'insurgents', 2909),  
( 'clashes', 2026),  
( 'between', 16072),  
( 'islamist', 7048),  
( 'militants', 33454),  
( 'northeast', 1586),  
( 'nigeria', 9560),
```

```
('five', 11818),
('civilians', 4340),
('elsewhere', 1876),
('region', 6020),
('monday', 19582),
('ceasefire', 8564),
('agreement', 3046),
('nigerian', 12286),
('expected', 9016),
('lead', 4752),
('liberation', 308),
('200', 7668),
('schoolgirls', 7532),
('kidnapped', 15168),
('talks', 7462),
('due', 8968),
('continue', 6770),
('neighbouring', 1184),
('chad', 4140),
('truce', 3520),
('there', 62897),
('attacks', 11740),
('over', 53662),
('blamed', 1796),
('security', 32190),
('sources', 18072),
('several', 19408),
('dozen', 2456),
('people', 66173),
('since', 32656),
('announcement', 4126),
('spokesman', 14686),
('may', 38383),
('work', 15900),
('criminal', 4500),
('gangs', 1144),
('lawless', 362),
('army', 9074),
('officer', 16192),
('requested', 1112),
('anonymity', 3326),
('tried', 4558),
('enter', 2186),
('town', 19038),
('damboa', 1212),
('late', 11480),
('through', 26426),
('alagarno', 248),
('hideout', 408),
('fought', 2216),
('them', 44504),
('“our', 500),
('gunned', 588),
('down', 22896),
('entered', 2152),
('unleashed', 384),
('terror', 11594),
('iust', 48498).
```

```

\ , - - - , - - - - ,
('picking', 1214),
('ruins', 388),
('armoured', 1046),
('vehicle', 6348),
('some', 52992),
('arms', 6118),
('recovered', 1908),
('garrison', 204),
('border', 29592),
('cameroon', 1318),
('fierce', 544),
('forces', 21822),
('sacked', 862),
('july', 5192),
('driven', 1320),
('out', 69486),
('offensive', 1152),
('member', 6346),
('pro', 6378),
('civilian', 1716),
('joint', 2630),
('task', 1252),
('vigilantes', 591),
('mohammed', 7319),
('haruna', 1078),
('“two', 220),
('our', 30788),
('members', 14054),
('came', 14044),
('biu', 204),
('morning', 10590),
('engaged', 886),
('yesterday', 5064),
('lasted', 642),
('till', 296),
('midnight', 1278),
('separately', 1064),
('maiduguri', 592),
('resident', 3246),
('andrew', 826),
('tada', 612),
('gava', 408),
('hilly', 204),
('gwoza', 234),
('far', 10106),
('brother', 7938),
('lucky', 2014),
('escaped', 786),
('top', 7706),
('mountain', 2806),
('“my', 1324),
('other', 42546),
('relatives', 2218),
('women', 11427),
('reuters', 7579),
('after', 106322),
('speaking', 6244),
\ , - - - , - - - - ,

```



```
('phone', 1544),  
('they', 3064),  
('scouting', 720),  
('sighted', 312),  
('pursued', 318),  
('slaughtered', 394),  
('much', 16848),  
('fake', 10954),  
('stuff', 1552),  
('internet', 11816),  
('given', 11754),  
('we've', 2048),  
('grown', 1658),  
('tired', 772),  
('debunking', 788),  
('fights', 442),  
('dhl', 90),  
('ads', 1430),  
('amazing', 2110),  
('viral', 6596),  
('video', 81321),  
('nope', 340),  
('jimmy', 636),  
('kimmel', 90),  
('stunt', 922),  
('rather', 4426),  
('each', 12392),  
('every', 7914),  
('undeservedly', 90),  
('story', 38306),  
('crosses', 104),  
('rounding', 90),  
('quick', 2484),  
('once', 8956),  
('friday', 11918),  
('debunk', 598),  
('misleading', 360),  
('headlines', 1258),  
('bad', 4706),  
('probably', 8408),  
('shouldn't', 538),  
('share', 5000),  
('ready', 3828),  
('here's', 2272),  
('nobel', 540),  
('doesn't', 5398),  
('regret', 90),  
('president', 20566),  
('obama', 23256),  
('peace', 3606),  
('prize', 1076),  
('according', 52121),  
('invented', 396),  
('news', 136),  
('release', 90),  
('uploaded', 4610),  
('scribd', 90),  
('prominent', 950),
```

```
( 'blogger', 1608),
( 'anti', 4614),
( 'war', 19363),
( 'activist', 1138),
( 'david', 10913),
( 'swanson', 90),
( 'can't', 4542),
( 'retract', 90),
( 'obama's', 3316),
( '2009', 3920),
( 'wishes', 804),
( 'remains', 6566),
( 'obligation', 158),
( 'disassociate', 90),
( 'itself', 4464),
( 'actions', 1810),
( 'taken', 20104),
( 'laureates', 180),
( 'frustrate', 90),
( 'advancing', 1446),
( 'sic', 660),
( 'fraternity', 96),
( 'statement', 23472),
( 'reads', 2422),
( 'tip', 1978),
( 'organizations', 3082),
( 'rarely', 762),
( 'glaring', 92),
( 'grammatical', 162),
( 'errors', 162),
( 'major', 7364),
( 'statements', 4550),
( 'therefore', 1312),
( 'joins', 124),
( 'public', 20322),
( 'expressing', 1000),
( 'regrets', 90),
( 'recipient', 90),
( 'picked', 2608),
( 'number', 16136),
( 'right', 16220),
( 'wing', 1346),
( 'outlets', 6770),
( 'including', 25896),
( 'before', 41528),
( 'problem', 6888),
( 'course', 5320),
( 'isn't', 4464),
( 'true', 11228),
( 'dear', 132),
( '"regrets"', 90),
( 'bogus', 1468),
( 'spoke', 7076),
( 'head', 15608),
( 'such', 17902),
( '"pumpkin', 184),
( 'spice', 8242),
( 'flavored', 2084).
```

```

, 'tve77ov5jz', 184),
('3', 8032),
('expects', 1070),
('“record', 90),
('shattering', 208),
('snowfall"', 270),
('many', 25556),
('others', 8454),
('comes', 10956),
('unfunny', 90),
('“satirists"', 90),
('empirenews', 90),
('net', 1422),
('increasingly', 1230),
('favorite', 644),
('generally', 630),
('unfavorable', 90),
('field', 1984),
('accompanying', 312),
('map', 1002),

```

```
('meteorologists', 90),  
('“above', 90),  
('normal', 3078),  
('huge', 4446),  
('swaths', 822),  
('states', 13864),  
('entire', 3944),  
('mid', 4742),  
('atlantic', 608),  
('coast', 484),  
('everything', 6412),  
('else', 4710),  
('empire', 1608),  
('moreover', 622),  
('difficult', 4274),  
('predict', 850),  
('seasonal', 90),  
('weather', 2249),  
('patterns', 212),  
('oceanic', 200),  
('atmospheric', 200),  
('administration', 4640),  
('bust', 540),  
('theirs', 210),  
('until', 13478),  
('november', 7690),  
('climate', 370),  
('prediction', 480),  
('college', 4638),  
('park', 4178),  
('md', 90),  
('baltimore', 804),  
('sun', 4842),  
('virtually', 972),  
('snow', 3864),  
('look', 9850),  
('three', 30488),  
('inothetnews', 90),  
('read', 8330),  
('meteorologist', 1528),  
('having', 8230),  
('“record', 90),  
('winter', 1386),  
('lord', 688),  
('jesus', 1324),  
('“male', 180),  
('birth', 6784),  
('control”', 180),  
('here', 11413),  
('2017', 228),  
('gleefully', 164),  
('injection', 90),  
('vasalgel', 270),  
('hitting', 1834),  
('world’s', 1438),  
('bedrooms', 90),  
('within', 7020),  
('alas', 414),
```

```
( 'suggest', 3284),
( 'proven', 692),
( 'humans', 594),
( 'already', 11703),
( 'approved', 600),
( 'drug', 6888),
( 'neither', 2470),
( 'those', 24546),
( 'things', 8702),
( 'release', 22021),
( 'call', 8386),
( 'donations', 702),
( 'foundation', 1120),
( 'behind', 8378),
( 'julia', 752),
( 'belluz', 90),
( 'explained', 3480),
( 'vox', 262),
( "highlighting", 90),
( 'promising', 510),
( 'again', 12726),
( 'unpublished', 90),
( 'anecdotal', 334),
( 'findings', 248),
( 'testing', 2330),
( 'baboons', 90),
( 'bottom', 1030),
( 'great', 6190),
( 'someone's', 224),
( 'working', 12988),
( 'male', 4420),
( 'control', 15084),
( 'highly', 2736),
( 'unlikely', 2380),
( 'particular', 2112),
( 'product', 4726),
( 'reach', 4600),
( 'market', 5918),
( '5', 11944),
( 'shooting', 22738),
( 'california', 4110),
( 'marine', 3003),
( 'corps', 528),
( 'anonymous', 2206),
( 'sparked', 2050),
( 'panic', 938),
( 'north', 40458),
( 'san', 3424),
( 'diego', 404),
( 'shooter', 3146),
( 'active', 2930),
( 'installation', 106),
( "all", 1366),
( 'marines', 1070),
( 'instructed', 262),
( 'lock', 1106),
( 'inside', 16244),
( 'rooms', 490)
```

```
\n    ('stupidly', 90),\n    ('followed', 5594),\n    ('prank', 1658),\n    ('post', 24360),\n    ('“updates”', 90),\n    ('situation', 7224),\n    ('shared', 5938),\n    ('hundreds', 4938),\n    ('police', 46945),\n    ('insider', 1882),
```

tk.word_docs



```
defaultdict(int,  
    {'an': 71648,  
     'international': 10213,  
     'nicaragua': 2303,  
     'press': 13606,  
     '2014': 12584,  
     'murillo': 1322,  
     'diameter': 142,  
     'a': 102511,  
     'reports': 31591,  
     'rock': 1997,  
     'all': 39674,  
     "city's": 992,  
     'streak': 494,  
     'foot': 3681,  
     'had': 58617,  
     'has': 77894,  
     'if': 36212,  
     'near': 16294,  
     'institute': 3272,  
     'hearing': 3940,  
     'something': 13543,  
     'anything': 5532,  
     'small': 7352,  
     'night': 11548,  
     'meteorite': 3745,  
     'i': 31314,  
     'told': 41129,  
     'site': 11484,  
     'nicaraguan': 2188,  
     'weekend': 6097,  
     'understanding': 864,  
     'saballos': 692,  
     'local': 11399,  
     'was': 81702,  
     'appears': 11772,  
     'expansive': 864,  
     "nicaragua's": 969,  
     'by': 75289,  
     'which': 46990,  
     'said': 68120,  
     'with': 82303,  
     'anyone': 8245,  
     'more': 45224,  
     'measured': 404,  
     'nothing': 7247,  
     'deep': 2166,  
     'sudden': 1884,  
     'this': 60874,  
     'sitting': 2324,  
     'depth': 670,  
     'were': 44996,  
     'then': 27198,  
     'abc': 758,  
     'rc': 1064,  
     'my': 17215,  
     'passing': 2273,  
     ..
```

'house': 11636,
'studies': 2184,
'base': 2256,
'sky': 1800,
'they': 53827,
'that': 91408,
'because': 26210,
'spokeswoman': 4476,
'humberto': 880,
'adviser': 1022,
'clear': 8536,
'government': 22614,
'determined': 2468,
'sunday': 7692,
'mysterious': 1697,
'he': 65508,
'wooded': 902,
'left': 18161,
'buried': 2538,
'not': 68941,
'strange': 2809,
'will': 40959,
'scientists': 2633,
'event': 6314,
'boom': 1283,
'didn't': 6294,
'and': 97270,
'large': 7855,
'media': 27440,
'formed': 1276,
'be': 71300,
'light': 3218,
'air': 12930,
'come': 14626,
'saturday': 6494,
'strauch': 844,
'saw': 7670,
'airport': 5741,
'associated': 8070,
'territorial': 1030,
'state': 32097,
'see': 18759,
'could': 32847,
'reported': 28772,
'into': 33335,
'what': 35568,
'happened': 8418,
'radius': 396,
'on': 92288,
'thought': 11708,
'visit': 3382,
'garcia': 594,
'jorge': 910,
'force': 7998,
'off': 23672,
'ask': 4210,
'very': 18140,
'porch': 580,


```
'sized': 1147,  
'still': 19620,  
'the': 104386,  
'committee': 3486,  
'relatively': 1406,  
'who': 61135,  
'planet': 904,  
'no': 41922,  
'skimmed': 72,  
'60': 2708,  
'close': 8500,  
'is': 89884,  
'capital': 9349,  
'disintegrated': 970,  
'at': 72927,  
'allowed': 2746,  
'blast': 2047,  
'to': 106834,  
'rosario': 982,  
'astronomy': 504,  
'study': 1932,  
'bomb': 2222,  
'we': 40361,  
'have': 75098,  
'earth': 3034,  
'crater': 2673,  
'experts': 5427,  
'ice': 726,  
'managua': 2068,  
'or': 42324,  
'39': 2090,  
'journalists': 3914,  
'overnight': 1498,  
'national': 11655,  
'asteroid': 1884,  
'loud': 2550,  
'but': 65802,  
'area': 10558,  
'santamaria': 746,  
'from': 73544,  
'feet': 3068,  
'help': 11297,  
'pass': 2200,  
'news': 33936,  
'university': 4150,  
'autonomous': 1034,  
'related': 6100,  
'wilfried': 788,  
'photo': 11118,  
'it': 78163,  
'one': 48580,  
'only': 23430,  
'center': 6655,  
'of': 103416,  
'felt': 5526,  
'in': 108177,  
'16': 3543,  
'heard': 12005
```

```
managua's": 819,  
'residents': 5410,  
'volcanologist': 664,  
'forecast': 700,  
'crashed': 414,  
'wave': 1714,  
'get': 20132,  
'dressed': 3174,  
'across': 11519,  
'full': 8973,  
'wednesday': 9376,  
'last': 32427,  
'note': 2868,  
'you': 30025,  
'daily': 10194,  
'haz': 2,  
'that's': 5020,  
'hurry': 18,  
'any': 28061,  
'ebola': 7782,  
'spotted': 2679,  
'outside': 8094,  
'c': 2787,  
'female': 4949,  
'mask': 790,  
'd': 1132,  
'apparently': 9128,  
'america': 6634,  
'waiting': 2914,  
'body': 8192,  
'jcpenny': 2,  
'try': 4465,  
'fears': 2838,  
'hinted': 230,  
'line': 4942,  
'for': 86185,  
'bankrupt': 60,  
'need': 6224,  
'hazmat': 534,  
'source': 11577,  
'new': 35165,  
'spread': 5227,  
'chances': 1082,  
'halloween': 146,  
'particularly': 2974,  
'taking': 9670,  
'caller': 454,  
'gown': 252,  
'stores': 1183,  
'complete': 2062,  
'confirmation': 3622,  
'today': 10571,  
'maybe': 2790,  
'business': 5073,  
'as': 72878,  
'passenger': 1047,  
'flight': 2284,
```

```
'gloves': 936,  
'mat': 2,  
'bag': 724,  
'washington': 8872,  
'dulles': 456,  
'liked': 632,  
'week': 19590,  
'suit': 1274,  
'retailer': 332,  
'side': 4446,  
'up': 45745,  
'selling': 2647,  
'dean': 440,  
'holding': 6764,  
'went': 17255,  
'preserved': 444,  
'it's': 14126,  
'even': 24579,  
'sauce': 116,  
'can': 28328,  
'tv': 7881,  
'started': 7432,  
'few': 11222,  
'adds': 1948,  
'offered': 3940,  
'1': 11058,  
'man': 28831,  
'kid's': 112,  
'sold': 4330,  
'page': 7596,  
'also': 44406,  
'project': 3200,  
'casey': 448,  
'say': 23128,  
'australia's': 196,  
'they've': 1706,  
'writing': 2906,  
'mold': 206,  
'they're': 3423,  
'why': 8239,  
'free': 7631,  
'1995': 256,  
'network': 5019,  
'day': 17360,  
'helps': 814,  
'pair': 2014,  
'never': 12830,  
'mcdonald's': 687,  
'their': 38459,  
'now': 29600,  
'beyond': 1724,  
'burger': 680,  
'burgers': 334,  
'old': 23981,  
'pounder': 592,  
'onto': 692,  
'itunes': 616,  
'quarter': 1736,
```

```
'teens': 1399,  
'perfectly': 1474,  
'same': 12490,  
'australians': 288,  
'pretty': 4374,  
'kanye': 702,  
'long': 14336,  
'its': 32723,  
'bottle': 946,  
'000': 14820,  
'back': 20856,  
'-': 12390,  
'friends': 9326,  
'bite': 997,  
'bought': 2635,  
'way': 16781,  
'year': 35276,  
'facebook': 10004,  
'ago': 14276,  
'two': 29181,  
'showed': 7577,  
'friend': 6321,  
'battle': 3259,  
'did': 19483,  
'69': 420,  
'4': 6105,  
'west': 8605,  
'likes': 828,  
'uneaten-and': 112,  
'nitz': 424,  
'd's': 112,  
'stayed': 892,  
'9': 9590,  
'mickey': 112,  
'years': 19887,  
'blue': 1442,  
'we're': 2964,  
'here's': 2430,  
'sure': 7606,  
'original': 4300,  
'buying': 1522,  
'his': 59539,  
'oldest': 1140,  
'men': 8617,  
'cheese': 3141,  
'later': 16059,  
'called': 13982,  
'show's': 112,  
'charity': 1376,  
'joke': 1836,  
'australian': 1930,  
'dissuaded': 112,  
'giving': 2823,  
'20': 7662,  
'so': 31835,  
'than': 28614,  
'took': 15373,  
'song': 1472
```

```
song': 1472,  
'months': 10826,  
'seemingly': 1579,  
'show': 12373,  
'depression': 236,  
'looks': 3882,  
'hosts': 536,  
'wonder': 2152,  
'specimen': 490,  
'newser': 212,  
'became': 4841,  
'wrapping': 250,  
'mcjordan': 112,  
'world': 19618,  
'mate': 258,  
'take': 14033,  
'eduard': 440,  
'how': 20113,  
'when': 41580,  
'anxiety': 496,  
'044': 112,  
'10': 13065,  
'food': 6959,  
'says': 24524,  
'proceeds': 520,  
'decays': 112,  
'cooper': 708,  
'group': 23623,  
'themselves': 4856,  
'caliph': 320,  
'fighter': 4802,  
'seem': 3594,  
'are': 60674,  
'posting': 1244,  
'first': 30722,  
'proclaimed': 1258,  
'": 34991,  
'almost': 7468,  
'fatigues': 760,  
't': 4128,  
'forced': 4244,  
'frequently': 1272,  
'fabricated': 674,  
'announced': 9360,  
'listed': 1134,  
'airstrikes': 6934,  
'u': 20141,  
'being': 34547,  
'linked': 4198,  
'brutal': 4156,  
'headquarters': 2993,  
'isis's': 348,  
'canadian': 5374,  
'noticed': 1932,  
'albaghdadi": 2,  
'young': 8874,  
'"i': 9014,  
'known': 16701,
```

```
'currently': 6994,  
'12': 9482,  
'soldier': 4529,  
'account': 6922,  
'going': 16405,  
'thing': 6936,  
'killed': 19362,  
'dallaire': 70,  
'“if': 2630,  
'like': 29488,  
'tweeting': 432,  
'emerged': 3316,  
'executive': 2988,  
'send': 1761,  
'abu': 5188,  
'carry': 2204,  
'co': 7584,  
'often': 5398,  
'nations': 2722,  
'strike': 3361,  
'“it': 4834,  
'would': 38675,  
'deploying': 384,  
'legitimate': 560,  
'fighters': 14177,  
'produces': 552,  
'photos': 6006,  
'weeks': 9191,  
'face': 6998,  
'http': 2805,  
'military': 12514,  
'ground': 7728,  
'month': 16839,  
'while': 30176,  
'think': 12761,  
'children': 6343,  
'times': 15078,  
'arrest': 3579,  
'canada': 2621,  
'al': 16956,  
'adding': 4952,  
'martyred”': 2,  
'baghdadi”': 2,  
'mission': 2394,  
'posted': 12264,  
'n1puexi7ir': 2,  
'commonly': 1338,  
'tank': 1672,  
'group’s': 2488,  
'parading': 70,  
'“shibal': 2,  
'social': 12567,  
'sham': 358,  
'interview': 7624,  
'give': 7302,  
'soldiers': 5351,  
'iraq': 17473,  
'istimes1': 2.
```

```
'died': 9242,  
'been': 66511,  
'counter': 2820,  
'youngest': 1099,  
'though': 13596,  
'used': 13235,  
'mr': 7872,  
'got': 10781,  
'based': 13618,  
'weapons': 6527,  
'terrorist': 7023,  
'identified': 8483,  
'charlie': 1590,  
'un': 7100,  
'propaganda': 2364,  
'claimed': 13175,  
'martyred': 178,  
'ubaidah': 360,  
'combat': 1476,  
'cubs': 70,  
'07': 142,  
'“everyone': 62,  
'june': 5012,  
'coming': 8407,  
'adults': 222,  
'“got': 8,  
'self': 2327,  
'patrols': 212,  
'bakr': 3449,  
'boots': 546,  
'romeo': 126,  
'13': 5113,  
'telephone': 750,  
'october': 5218,  
'shelly': 90,  
'limited': 2515,  
'london': 7345,  
'guard': 3729,  
'blood': 4520,  
'roughly': 2246,  
'boy': 5413,  
'director': 8146,  
'us': 24731,  
'baghdadi': 3836,  
'thursday': 8266,  
'past': 11855,  
'“the': 10546,  
'“more': 208,  
'nato': 566,  
'vouch': 70,  
'six': 12942,  
'hashtag': 1090,  
'lions': 120,  
'she': 23599,  
'make': 16715,  
'“uniform': 2,  
'twitter': 16982,  
'station': 2
```

```
totaling : 2,  
'researcher': 1082,  
'strikes': 6452,  
'2': 10237,  
'whitman': 70,  
'translates': 696,  
'dead': 10028,  
'child': 5584,  
'-': 19492,  
'confirmed': 20012,  
'showing': 7510,  
'conduct': 908,  
'prisoners': 1612,  
'united': 9879,  
'father': 7613,  
'cub': 541,  
'26': 3895,  
'injured': 3857,  
'british': 9529,  
'gun': 3281,  
'worried': 1690,  
'alongside': 3205,  
'ban': 781,  
'monitors': 1094,  
'extremism': 692,  
'initiative': 174,  
'seems': 6523,  
'about': 47818,  
's': 20488,  
'supporters': 1666,  
'report': 26654,  
'accounts': 3858,  
'directly': 2356,  
'use': 10038,  
'allegedly': 10493,  
'syria': 17882,  
'quilliam': 482,  
'islamic': 23187,  
'refer': 404,  
'sept': 3068,  
'smiling': 894,  
'appear': 4474,  
'online': 11961,  
'troops': 3612,  
'isis': 29789,  
'halifax': 70,  
'fighting': 7580,  
'hilly': 204,  
'tada': 204,  
'agreement': 2192,  
'terror': 7838,  
'out': 44917,  
'schoolgirls': 3968,  
'200': 6114,  
'alagarno': 248,  
'fought': 2148,  
'army': 6492,  
'pro': 4579,
```



```
'reuters': 5737,  
'mohammed': 4335,  
'scouting': 462,  
'kidnapped': 9743,  
'them': 30514,  
'monday': 12527,  
'midnight': 1206,  
'till': 296,  
'gangs': 1104,  
'member': 5893,  
'militants': 15613,  
'recovered': 1627,  
'border': 11477,  
'“two': 220,  
'late': 9919,  
'criminal': 3413,  
'work': 10430,  
'tried': 3831,  
'suspected': 5295,  
'july': 4474,  
'continue': 5502,  
'lasted': 642,  
'region': 4550,  
'spokesman': 11890,  
'slaughtered': 394,  
'may': 25307,  
'five': 7681,  
'armoured': 756,  
'town': 10016,  
'northeast': 1340,  
'women': 7360,  
'expected': 7363,  
'damboa': 278,  
'mountain': 2154,  
'sacked': 738,  
'announcement': 3689,  
'“my': 1182,  
'pursued': 264,  
'after': 55665,  
'ruins': 388,  
'blamed': 1424,  
'cameroon': 764,  
'members': 11157,  
'gwoza': 234,  
'truce': 2585,  
'clashes': 1416,  
'boko': 6939,  
'gava': 204,  
'there': 40870,  
'25': 5440,  
'dozen': 2430,  
'nigerian': 3792,  
'escaped': 786,  
'relatives': 1670,  
'lead': 4421,  
'over': 35875,  
'yesterday': 3828,  
'brother': 5019.
```

```
.....  
'morning': 7417,  
'sighted': 312,  
'came': 12472,  
'hideout': 408,  
'elsewhere': 1688,  
'security': 16089,  
'phone': 5291,  
'speaking': 5932,  
'unleashed': 384,  
'engaged': 840,  
'offensive': 1116,  
'other': 30842,  
'sources': 11815,  
'civilian': 1550,  
'just': 32484,  
'haruna': 1046,  
'people': 36245,  
'islamist': 4943,  
'fierce': 544,  
'top': 6826,  
'officer': 6737,  
'maiduguri': 562,  
'chad': 2376,  
'andrew': 778,  
'driven': 1320,  
'haram': 7066,  
'resident': 2658,  
'through': 17891,  
'garrison': 204,  
'some': 34203,  
'attacks': 7072,  
'between': 12897,  
'vehicle': 4114,  
'lucky': 1352,  
'ceasefire': 3894,  
'since': 21982,  
'nigeria': 4317,  
'our': 500,  
'enter': 2130,  
'least': 15475,  
'several': 15231,  
'joint': 2608,  
'vigilantes': 589,  
'far': 8693,  
'insurgents': 1540,  
'our': 19754,  
'gunned': 558,  
'neighbouring': 1130,  
'picking': 1214,  
'they': 2764,  
'down': 18701,  
'civilians': 3668,  
'forces': 12033,  
'requested': 1010,  
'entered': 2098,  
'anonymity': 3082,  
'arms': 3562,  
'...': 7005
```

```
'due': 7805,  
'liberation': 308,  
'lawless': 362,  
'biu': 204,  
'separately': 1064,  
'task': 1236,  
'talks': 4737,  
'mid': 4368,  
'remains': 4960,  
'sparked': 2050,  
'cause': 4349,  
'article': 7954,  
'vasalgel': 90,  
'anonymous': 2056,  
'marines': 626,  
'fallen': 1976,  
'atmospheric': 200,  
'el': 414,  
'attached': 1516,  
'otherwise': 1004,  
'“it’s’: 2716,  
'panic': 851,  
'scam': 454,  
'sodas': 90,  
'amazingly': 264,  
'world’s’: 1436,  
'birth': 4430,  
'prize': 804,  
'thereby': 140,  
'wing': 1136,  
'prediction': 316,  
'swaths': 766,  
'bust': 420,  
'time': 34925,  
'seasonal': 90,  
'actually': 9290,  
'shared': 4849,  
'fact': 10863,  
'below': 4986,  
'circulated': 2662,  
'notable': 480,  
'telling': 4084,  
'donations': 698,  
'sad': 2066,  
'ever': 8550,  
'foundation': 1112,  
'nope': 340,  
'“but': 1624,  
'here’s’: 1460,  
'character': 1538,  
'tip': 1685,  
'glaring': 92,  
'huge': 4050,  
'shooting': 10505,  
'sheikh': 546,  
'circles': 216,  
'credulous': 90,  
'boyfriend': 1228,
```

```
'md': 90,  
'meteorologist': 1073,  
'promising': 510,  
'peppa': 90,  
'talk': 4446,  
'increasingly': 1194,  
'release": 90,  
'haram': 90,  
'remote': 1330,  
'generally': 574,  
'war': 12350,  
'spoke': 5960,  
'proven': 692,  
'ye': 90,  
'days': 15447,  
'don't': 7209,  
'muslims': 3618,  
'considering': 3144,  
'edition": 90,  
'stupidly': 90,  
'snow': 1806,  
'bedroom': 1608,  
'conspiracy': 847,  
'served': 2828,  
'net': 1396,  
'jesus': 860,  
'snowfall": 90,  
'singer's': 182,  
'saturation': 184,  
'alliterative': 90,  
'durex': 1682,  
'lord': 504,  
'much': 13368,  
'beverage': 232,  
'headlines': 1238,  
'hitting': 1834,  
'...': 3459,  
'once': 7851,  
'laced': 246,  
'colombia': 138,  
'donald': 144,  
'many': 19202,  
'peppa': 90,  
'office': 8368,  
'matter': 3608,  
'others': 7348,  
'invented': 396,  
'tve77ov5jz': 184,  
'australia': 1169,  
'regrets': 90,  
'male': 90,  
'normal': 2995,  
'record': 90,  
'map': 676,  
'second': 10164,  
'virus': 3187,  
'field': 1922,  
'rights': 6154.
```

```

    'taken': 16243,
    'moreover': 622,
    'possible': 8706,
    'corps': 390,
    'lock': 1106,
    'states': 10486,
    'activist': 1138,
    'california': 3463,
    'julia': 668,
    'else': 4238,
    'video': 33187,
    'widespread': 906,
    '"record': 90,
    'number': 12614,
    'pig": 90,
    'here': 9282,
    'rounding': 90,
    'problem': 5971,
    'commenter': 438,
    'meteorologists': 90,
    'situation': 5846,
    '"limited': 90,
    'youtube': 8066,
    'campaign': 6116,
    'har': 90,
    'anybody': 726,
    'we've': 1920,
    'according': 35259,
    'advancing': 1446,
    'quite': 4166,
    'published': 5510,
    'dewey': 128,
    '"regret": 90,
    'networks': 1145,
    'hoax': 6552,
    'atlantic': 599,
    'existence': 2028,
    'park': 3180,
    'such': 13588,
    'dhl': 90,
    'olden': 90,
    'meantime': 888,
    'scribd': 90,
    'belluz': 90,
    'perhaps': 4388,
    'civil': 3491,
    'most': 18723,
    'image': 8638,
    'huzlers': 278,
    'bottom': 1030,
    'probably': 7244,
    'viral': 5084,
    'sic': 598,
    'go': 11607,
    'beer': 564,
    'beverages': 90,
    'errors': 162,
    '':

```

```
'soon': 7414,  
'bull': 572,  
'phenomenon': 340,  
'caitlin': 1696,  
'section': 1190,  
'statement': 14759,  
'anecdotal': 334,  
'lifetime': 232,  
'living': 5671,  
'faux': 124,  
'course': 4702,  
'idea': 3964,  
'gleefully': 164,  
'car': 6827,  
'misleading': 360,  
'feel': 3878,  
'photoshopped': 548,  
'contained': 960,  
'difficult': 3978,  
'actions': 1652,  
'pumpkin': 184,  
'red': 3681,  
'every': 6833,  
'jimmy': 636,  
'oreos': 222,  
'amid': 2999,  
'hundreds': 3882,  
'do': 23465,  
'bedrooms': 90,  
'approved': 600,  
'comes': 9778,  
'word': 3416,  
'david': 7014,  
'diego': 404,  
'images': 5003,  
'punctured': 90,  
'enterprising': 102,  
'accompanying': 312,  
'virtually': 972,  
'rooms': 348,  
'rather': 4042,  
'lie': 668,  
'desert': 1718,  
'prominent': 950,  
'crosses': 104,  
'reached': 6174,  
'expressing': 870,  
'tired': 592,  
'pendleton': 90,  
'dispatcher': 144,  
'police': 17043,  
'accident': 382,  
'30': 90,  
'3': 6403,  
'therefore': 1299,  
'fights': 442,  
'point': 8026,  
'this': 2834,
```

```

'alleged': 10953,
'peace': 3074,
'laureates': 90,
'shakira': 90,
'unfunny': 90,
'ads': 886,
'ready': 3364,
'outraged': 711,
'"regrets"': 90,
'president': 13884,
'"updates"': 90,
'"saw"': 90,
'empirenews': 90,
'san': 1832,
'whole': 2639,
'colombia's': 90,
'testing': 2070,
'male': 2709,
'coca': 162,
'unlikely': 2138,
'next': 16277,
'surely': 570,
'market': 4352,
'unfavorable': 90,
'within': 6212,
'insider': 1492,
'"exclusive"': 136,
'explains': 1357,
'bad': 4402,
'winter': 1249,
'obama': 10056,
'"highlighting"': 90,
'having': 7280,
'unpublished': 90,
'name': 15813,
'organize': 312,
'condom': 1638,
'ancient': 436,
'"satirists"': 90,
'outlets': 5711,
'call': 6834,
'college': 2648,
'spice': 2171,
'where': 24187,
'stupid': 409,
'nobel': 270,
'irrigation': 90,
'obama's': 2165,
'traces': 90,
'"we"': 7932,
'fake': 6081,
'courses': 256,
'read': 7509,
'stay': 2878,
'before': 28463,
'service': 9867,
'expects': 743,
'"in"': 90

```

```

    'joke': 50,
    'reads': 1770,
    'incredible': 1421,
    'kimmel': 90,
    'shattering': 118,
    'including': 19110,
    'instructed': 262,
    'trump': 142,
    'swanson': 90,
    'straight': 1009,
    'head': 12763,
    'edition': 3419.

```

```
word_index = tk.word_index
```

```
tk.document_count
```



149916

Now, tokenize the sentences using `nltk sent_tokenize()` and encode the sentences with the ids we got from `word_index`.
Initialise 2 lists with names `texts` and `articles`.

```
texts = [] to store text of article as it is.
```

```
articles = [] split the above text into a list of sentences.
```

```
texts = []
articles = []
```

```
texts = dataset['articleBody']
```

```

from nltk.tokenize import sent_tokenize
for items in texts.iteritems():
    articles.append(sent_tokenize(items[1]))

```

▼ Check 2:

first element of `texts` and `articles` should be as given below.

```
texts[0]
```



```
'A small meteorite crashed into a wooded area in Nicaragua\'s capital of Managua overnight'
```

```
articles[0]
```



```
articles[0]
```



```
["A small meteorite crashed into a wooded area in Nicaragua's capital of Managua overnig
"Residents reported hearing a mysterious boom that left a 16-foot deep crater near the
'Government spokeswoman Rosario Murillo said a committee formed by the government to st
'House-sized asteroid 2014 RC, which measured 60 feet in diameter, skimmed the Earth th
'Murillo said Nicaragua will ask international experts to help local scientists in unde
'The crater left by the meteorite had a radius of 39 feet and a depth of 16 feet, said
'He said it is still not clear if the meteorite disintegrated or was buried.',
'Humberto Garcia, of the Astronomy Center at the National Autonomous University of Nica
'"We have to study it more because it could be ice or rock," he said.',
'Wilfried Strauch, an adviser to the Institute of Territorial Studies, said it was "ver
'Ve have to ask if anyone has a photo or something."',
"Local residents reported hearing a loud boom Saturday night, but said they didn't see
'"I was sitting on my porch and I saw nothing, then all of a sudden I heard a large bla
'Ve thought it was a bomb because we felt an expansive wave," Jorge Santamaria told The
"The site of the crater is near Managua's international airport and an air force base."
'Only journalists from state media were allowed to visit it.']
```

Now iterate through each article and each sentence to encode using `t.word_index` [5 marks]

Here, to get words from sentence you can use `text_to_word_sequence` from keras preprocessing text

1. Import `text_to_word_sequence`
2. Initialize a variable of shape (no.of articles, MAX_SENTS, MAX_SENT_LENGTH) with name `data` numpy [np.zeros](#) to initialize with all zeros)and then update it while iterating through the words a

```
from keras.preprocessing.text import text_to_word_sequence
import numpy as np
```

```
data = np.zeros((len(articles), MAX_SENTS, MAX_SENT_LENGTH)).astype(int)
```

```
data.shape
```



```
(49972, 20, 20)
```

```
for pos1, item in enumerate(articles):
    for pos2, sent in enumerate(item):
        if(sent is not None) & (pos2 < MAX_SENTS):
            temp = text_to_word_sequence(sent)
            for pos3, word in enumerate(temp):
                if (pos3 < MAX_SENT_LENGTH):
                    data[pos1][pos2][pos3] = word_index.get(word, 0)
```

▼ Check 3:

Accessing first element in data should give something like given below.

```
data[0, :, :]
```



— — — — —


```
[ 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0]])
```

Repeat the same process for the Headings as well. Use variable `texts_heading` and `articles_heading` accordingly. [5 marks]

```
texts_headings = []
articles_headings = []
texts_headings = dataset['Headline']
for items in texts_headings.iteritems():
    articles_headings.append(sent_tokenize(items[1]))
```

```
data_heading = np.zeros((len(articles_headings), MAX_SENTS, MAX_SENT_LENGTH)).astype(int)
```

```
data_heading.shape
```

 (49972, 20, 20)

```
for pos1, item in enumerate(articles_headings):
    for pos2, sent in enumerate(item):
        if(sent is not None) & (pos2 < MAX_SENTS):
            temp = text_to_word_sequence(sent)
            for pos3, word in enumerate(temp):
                if (pos3 < MAX_SENT_LENGTH):
                    data_heading[pos1][pos2][pos3] = word_index.get(word, 0)
```

```
data_heading[0, :, :]
```



— — — — —

```
[ 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0]])
```


- ▼ Now the features are ready, lets make the labels ready for the model to process.

Convert labels into one-hot vectors

You can use [get_dummies](#) in pandas to create one-hot vectors.

```
labels = pd.get_dummies(dataset['Stance'])
```

```
labels.sample(5)
```




	agree	disagree	discuss	unrelated
46694	0	0	1	0
10997	1	0	0	0
19366	0	0	0	1
4570	0	0	0	1
4468	0	0	0	1

- ▼ Check 4:

The shape of data and labels should match the given below numbers.

```
print('Shape of data tensor:', data.shape)
print('Shape of label tensor:', labels.shape)
```



```
Shape of data tensor: (49972, 20, 20)
Shape of label tensor: (49972, 4)
```

- ▼ Shuffle the data

```
## get numbers upto no.of articles
indices = np.arange(data.shape[0])
## shuffle the numbers
np.random.shuffle(indices)

## shuffle the data
data = data[indices]
data_heading = data_heading[indices]
## shuffle the labels according to data
labels = labels.iloc[indices]
```

▼ Split into train and validation sets. Split the train set 80:20 ratio to get the train an

Use the variable names as given below:

x_train, x_val - for body of articles.

x-heading_train, x_heading_val - for heading of articles.

y_train - for training labels.

y_val - for validation labels.

```
from sklearn.model_selection import train_test_split
```

```
x_train, x_val, y_train, y_val = train_test_split(data, labels, test_size=0.2, random_state=4
```

```
print(x_train.shape)
```

```
print(y_train.shape)
```

```
print(x_val.shape)
```

```
print(y_val.shape)
```



```
(39977, 20, 20)
```

```
(39977, 4)
```

```
(9995, 20, 20)
```

```
(9995, 4)
```

▼ Check 5:

The shape of x_train, x_val, y_train and y_val should match the below numbers.

```
print(x_train.shape)
```

```
print(y_train.shape)
```

```
print(x_val.shape)
```

```
print(y_val.shape)
```



```
(39978, 20, 20)
```

```
(39978, 4)
```

```
(9994, 20, 20)
```

```
(9994, 4)
```

```
x_train = np.reshape(x_train, (39977, 400))
```

```
x_val = np.reshape(x_val, (9995, 400))
```

▼ Create embedding matrix with the glove embeddings

Run the below code to create `embedding_matrix` which has all the words and their glove embedding if

```
# load the whole embedding into memory
embeddings_index = dict()
f = open('./glove.6B.100d.txt')
for line in f:
    values = line.split()
    word = values[0]
    coefs = np.asarray(values[1:], dtype='float32')
    embeddings_index[word] = coefs
f.close()

print('Loaded %s word vectors.' % len(embeddings_index))
```

```
# create a weight matrix for words in training docs
embedding_matrix = np.zeros((vocab_size, 100))
```

```
for word, i in t.word_index.items():
    embedding_vector = embeddings_index.get(word)
    if embedding_vector is not None:
        embedding_matrix[i] = embedding_vector
```



```
-----
FileNotFoundError                                Traceback (most recent call last)
<ipython-input-1-b99d0e2fceb7> in <module>()
      1 embeddings_index = dict()
----> 2 f = open('./glove.6B.100d.txt')
      3 for line in f:
      4     values = line.split()
      5     word = values[0]

FileNotFoundError: [Errno 2] No such file or directory: './glove.6B.100d.txt'
```

SEARCH STACK OVERFLOW

- Try the sequential model approach and report the accuracy score
- Import layers from Keras to build the model

```
import tensorflow as tf
from tensorflow import keras
from keras import backend as K
import numpy as np
from keras.layers import LSTM, Dense, Dropout, Embedding, Masking, Bidirectional, GlobalAveragePooling1D
tf.set_random_seed(100)
```




```

-----
AttributeError                                Traceback (most recent call last)
<ipython-input-3-17a64dc70691> in <module>()
      4 import numpy as np
      5 from keras.layers import LSTM, Dense, Dropout, Embedding, Masking, Bidirectional
----> 6 tf.set_random_seed(100)

```

AttributeError: module 'tensorflow' has no attribute 'set_random_seed'

SEARCH STACK OVERFLOW

▼ Model

```

model = keras.Sequential()
model.add(keras.layers.Embedding(len(word_index) + 1, 100))
model.add(keras.layers.GlobalAveragePooling1D())
model.add(keras.layers.Dense(64, activation='relu'))

```

```

# Dropout for regularization
model.add(keras.layers.Dropout(0.5))
model.add(keras.layers.Dense(32, activation="relu"))
model.add(keras.layers.Dense(16, activation="relu"))
model.add(keras.layers.Dense(4, activation="sigmoid"))

```

```
model.summary()
```



Model: "sequential"

Layer (type)	Output Shape	Param #
=====		
embedding (Embedding)	(None, None, 100)	3367800

global_average_pooling1d (Gl	(None, 100)	0

dense (Dense)	(None, 64)	6464

dropout (Dropout)	(None, 64)	0

dense_1 (Dense)	(None, 32)	2080

dense_2 (Dense)	(None, 16)	528

dense_3 (Dense)	(None, 4)	68
=====		
Total params: 3,376,940		
Trainable params: 3,376,940		
Non-trainable params: 0		

▼ Compile and fit the model

```
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

```
history = model.fit(x_train, y_train, validation_data=(x_val, y_val), epochs = 10, batch_size
```



Epoch 1/10

800/800 [=====] - 29s 36ms/step - loss: 0.8492 - accuracy: 0.73

Epoch 2/10

800/800 [=====] - 29s 36ms/step - loss: 0.8011 - accuracy: 0.73

Epoch 3/10

800/800 [=====] - 29s 36ms/step - loss: 0.7969 - accuracy: 0.73

Epoch 4/10

800/800 [=====] - 28s 35ms/step - loss: 0.7939 - accuracy: 0.73

Epoch 5/10

800/800 [=====] - 29s 36ms/step - loss: 0.7920 - accuracy: 0.73

Epoch 6/10

800/800 [=====] - 29s 36ms/step - loss: 0.7895 - accuracy: 0.73

Epoch 7/10

800/800 [=====] - 29s 36ms/step - loss: 0.7882 - accuracy: 0.73

Epoch 8/10

800/800 [=====] - 29s 36ms/step - loss: 0.7864 - accuracy: 0.73

Epoch 9/10

800/800 [=====] - 28s 36ms/step - loss: 0.7840 - accuracy: 0.73

Epoch 10/10

800/800 [=====] - 29s 36ms/step - loss: 0.7810 - accuracy: 0.73

- ▼ Build the same model with attention layers included for better perform

- ▼ Fit the model and report the accuracy score for the model with attentic

