NLP_corpusnotebook_22183822014

February 22, 2024

[3]:
```
import nltk
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
from nltk import stem
stemmer = stem.PorterStemmer()
from nltk import word_tokenize
nltk.download('stopwords')
from nltk.corpus import stopwords
stops = set(stopwords.words('english'))
nltk.download('punkt')
import string
punct = list(string.punctuation)
from collections import Counter
import requests
import pandas as pd
#import seaborn as sns
#sns.set()
from matplotlib import *
!pip install PRAW
import numpy as np
import praw
import datetime
```

```
[nltk_data] Downloading package wordnet to /home/svetlana/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/svetlana/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /home/svetlana/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: PRAW in
/home/svetlana/.local/lib/python3.10/site-packages (7.7.1)
Requirement already satisfied: websocket-client>=0.54.0 in
/home/svetlana/.local/lib/python3.10/site-packages (from PRAW) (1.6.3)
Requirement already satisfied: update-checker>=0.18 in
```

```
/home/svetlana/.local/lib/python3.10/site-packages (from PRAW) (0.18.0)
Requirement already satisfied: prawcore<3,>=2.1 in
/home/svetlana/.local/lib/python3.10/site-packages (from PRAW) (2.4.0)
Requirement already satisfied: requests<3.0,>=2.6.0 in
/home/svetlana/.local/lib/python3.10/site-packages (from prawcore<3,>=2.1->PRAW)
(2.31.0)
Requirement already satisfied: certifi>=2017.4.17 in
/home/svetlana/.local/lib/python3.10/site-packages (from
requests<3.0,>=2.6.0->prawcore<3,>=2.1->PRAW) (2019.11.28)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/svetlana/.local/lib/python3.10/site-packages (from
requests<3.0,>=2.6.0->prawcore<3,>=2.1->PRAW) (1.25.7)
Requirement already satisfied: idna<4,>=2.5 in
/home/svetlana/.local/lib/python3.10/site-packages (from
requests<3.0,>=2.6.0->prawcore<3,>=2.1->PRAW) (2.8)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/svetlana/.local/lib/python3.10/site-packages (from
requests<3.0,>=2.6.0->prawcore<3,>=2.1->PRAW) (3.3.0)
```

0.0.1 Introduction

The news is all around us - on social media, newspapers, radio. Research shows that constant exposure to news can make people feel stressed and anxious, but does it also affect how they respond to it? This notebook will attempt to begin to answer that question by gathering data from the subreddit r/worldnews, tokenising and lemmatising it, and analysing the comments using the VAD model to see which comments score the highest and/or lowest.

NLP is useful to us in answering this question, both by providing the tools we need (e.g. the NLTK library and VAD model) but also for its ability to work with lots of data. Since our question revolves news as a whole, rather than a specific incident, we need to be able to gather enough data from multiple pieces of news to analyse people's responses.

In this case, I decided to use the Reddit API to gather all top-level comments (i.e. immediate responses rather than reply chains) from the top 10 posts of the past year on the subreddit r/worldnews. The choice of subreddit was in part due to its popularity (to avoid having to scrape lots of posts), and also to try and counteract the biggest problem with using Reddit as a source - how US centric it is. Nearly 50% of Reddit traffic comes from the USA, where the US only accounts for about 4% of the world population. The subreddit r/worldnews, however, specifically bans posts that relate to US-internal matters, hopefully meaning that our sample is more representative of news topics worldwide and individuals' reactions to them.

However, just because the sample size is large (and if run correctly, this notebook should result in a corpus of 4490 comments) doesn't mean it's not biased - these comments all still come from the same 10 posts, all of which were popular even by the subreddit's standards, and thus could be considered outliers. This could transfer to the comments VAD scores if they were reacting to particularly noteworthy news, compared to events one might read about in a local newspaper.

0.0.2 Gathering Data via Reddit API Function

[4]:
```
reddit = praw.Reddit(user_agent='VAD',
                     client_id='5fKrQPC9VKrFf3F00g4VnQ',␣
 ↪client_secret="kDxCm9NjR_D3QkfatcoR-M6cNy8j-A",
                     username='crochet9000', password='lkjhgfdsa')
```

[5]:
```
#This code connects to the Reddit API and gathers the top-level comment data␣
 ↪for the post url provided.
#All code in the cell originally by James Carney (module lead) with permission␣
 ↪to use in this assignment
def submission(submission_id):
    try:
        submission = reddit.submission(url = submission_id)
    except:
        submission = reddit.submission(submission_id)
    title = submission.title
    submission.comments.replace_more() ## loads new page if cooments are␣
 ↪multipage
    text = [i.body for i in submission.comments]
    score = [i.score for i in submission.comments]
    user = [i.author for i in submission.comments]
    date = [datetime.datetime.fromtimestamp(i.created) for i in submission.
 ↪comments]
    df = pd.DataFrame()
    df['text'] = text
    df['datetime'] = date
    df['score'] = score
    df['subreddit'] = submission.subreddit
    df['redditor'] = user
    df['type'] = 'comment'
    df['title'] = title
    df = df.sort_values('score', ascending = False).reset_index(drop = True)
    return df
```

0.0.3 Tokenising & Lemmatising Function

This function does two things: tokenises each comment (breaks it down into individual words and symbols, i.e. tokens) and then lemmatises each word - defaults it down to its basic form. This includes turning plural words singular, adjectives to their root word, etc. standerdising our dataset.

[6]:
```
#This function builds on the previous one for ease of use.
def lemmatise(reddit_thread):
    df = submission(reddit_thread)
    df['text'] = [str(i) for i in df['text']]
    lemmas_1 = []
    for i in df['text']:
```

```
        tokens_ = word_tokenize(i)
        lemmas_ = [lemmatizer.lemmatize(i.lower()) for i in tokens_]
        lemmas = [i for i in lemmas_ if i not in stops]
        lemmas_1.append(lemmas)
    df['words'] = lemmas_1
    return df
```

0.0.4 VAD Analysis Function

The VAD model assigns each word three scores from 0 to 1 - its Valence (how positive it is), Arousal (how stimulating/passionate), and Dominance (how in control it makes one feel). While it is not perfect (new slang words are created all the time, and would likely not be immediately added to the overall model) it can help compare the emotions exhibited in one comment to another.

[7]:
```
vad = pd.read_excel('vad.xlsx', index_col = 0)
```

[8]:
```
#This function builds on the previous two for ease of use.
def VAD_Analysis(reddit_thread):
    df = lemmatise(reddit_thread)
    vad_ = []
    c = [np.nan for i in range(len(vad.columns))]
    for i in df['words']:
        words = [j for j in i if j in vad.index]
        vad_1 = []
        for k in words:
            try:
                vad_1.append(vad.loc[k])
            except:
                 vad_1.append(c)
        vad_df = pd.DataFrame(vad_1, columns = [i for i in vad.columns])
        vad_.append(vad_df.mean())
    vad_ = pd.DataFrame(vad_)
    data = pd.concat([df, vad_], axis =1)
    return data
```

0.1 Creating Full Corpus

0.1.1 Do not run the cells in this section! This will take a very long time to re-scrape the comments. The next section includes the import of the saved csv file.

[13]:
```
#Applying our three functions to the top 10 posts of the past year on the␣
 ↪subreddit r/worldnews
#as gathered on Thursday, 15th February at 14:51:00.
post_1 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/123iv0t/
 ↪norway_sweden_finland_and_denmark_struck_a_deal/')
post_2 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/1172vx1/
 ↪president_biden_makes_surprise_visit_to_ukraine/')
```

```
post_3 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/129gpui/
 ↪analysis_of_twitter_algorithm_code_reveals_social/')
```

[14]:
```
post_4 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/1192qby/
 ↪biden_vows_to_defend_literally_every_inch_of_nato/')
post_5 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/13qz1sw/
 ↪under_elon_musk_twitter_has_approved_83_of/')
post_6 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/118l82r/
 ↪japan_promises_to_lead_the_world_in_fighting/')
```

[16]:
```
post_7 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/12bdmyb/
 ↪nato_gathers_to_welcome_finland_as_31st_member/')
post_8 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/1183r5w/
 ↪putin_falsely_claims_it_was_west_that_started_the/')
post_9 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/12csm7u/
 ↪cisco_systems_pulled_out_of_russia_and_destroyed/')
post_10 = VAD_Analysis('https://www.reddit.com/r/worldnews/comments/11aqt3d/
 ↪lithuanias_prime_minister_says_ukrainians_should/')
```

[156]:
```
corpus = pd.concat([post_1, post_2, post_3, post_4, post_5, post_6, post_7,␣
 ↪post_8, post_9, post_10], axis = 0, ignore_index = True)
corpus.head()
```

[156]:

	text	datetime	score	subreddit	redditor	type	title	words	valence	arousal
0	This invasion has done more to unite the Europ…	2023-03-27 11:38:10	21695	worldnews	greek_stallion	comment	Norway, Sweden, Finland, and Denmark struck a …	[invasion, ha, done, unite, european, continen…	0.556075	0.565259
1	Now you brought the Vikings back together, dam…	2023-03-27 12:37:39	20345	worldnews	Shotguns_x_559	comment	Norway, Sweden, Finland, and Denmark struck a …			
2	[deleted]	2023-03-27 11:58:54	7522	worldnews	None	comment	Norway, Sweden, Finland, and Denmark struck a …			
3	Introducing NATO+	2023-03-27 13:06:05	6666	worldnews	SpaceToaster	comment	Norway, Sweden, Finland, and Denmark struck a …			
4	Unified Nordic Defense Force sounds like somet…	2023-03-27 13:02:47	4437	worldnews	BMCarbaugh	comment	Norway, Sweden, Finland, and Denmark struck a …			

```
1          [brought, viking, back, together, ,, damn]  0.484521  0.559284
2                                       [[, deleted, ]]       NaN       NaN
3                                   [introducing, nato+]       NaN       NaN
4  [unified, nordic, defense, force, sound, like,…  0.615946  0.598306

   dominance
0   0.521931
1   0.495146
2        NaN
3        NaN
4   0.609628
```

[161]:
```
#we then clean the corpus from comments that have NaN values
#this is better than removing all 'deleted' comments as those still sometimes␣
 ↪have text
corpus = corpus.dropna().reset_index(drop=True)
```

[162]:
```
len(corpus)
```

[162]:
```
4490
```

0.1.2 Storing the Corpus

[ ]:
```
#Don't run this cell either! This code is showing how I originally stored the␣
 ↪corpus
corpus.to_csv('csvcorpus.csv')
```

[9]:
```
#run this cell to get the corpus!
corpus = pd.read_csv('csvcorpus.csv', index_col = 0)
corpus
```

[9]:
```
                                                 text             datetime  \
0     This invasion has done more to unite the Europ…  2023-03-27 11:38:10
1     Now you brought the Vikings back together, dam…  2023-03-27 12:37:39
2     Unified Nordic Defense Force sounds like somet…  2023-03-27 13:02:47
3     The vikings have returned! But to the skies in…  2023-03-27 11:57:11
4     This would make it one the largest air-forces …  2023-03-27 11:42:18
…                                                   …                  …
4485  We should keep some as a deterrence or for use…  2023-02-24 14:03:02
4486  Finally saying the quiet part out loud. Europe…  2023-02-24 16:38:29
4487  And the Ukrainian "make a wish" foundation con…  2023-02-24 15:06:17
4488                              get fucked warmonger  2023-02-24 16:30:54
4489  Except Ukraine is well known for having a blac…  2023-02-24 15:10:47

      score  subreddit            redditor     type  \
0     21695  worldnews       greek_stallion  comment
```

```
1     20345  worldnews       Shotguns_x_559  comment
2      4437  worldnews           BMCarbaugh  comment
3      3714  worldnews  Ok_Imagination_7119  comment
4      2884  worldnews              008Zulu  comment
…      …        …                   …      …
4485    -21  worldnews    thrownkitchensink  comment
4486    -26  worldnews       RickyTicky5309  comment
4487    -32  worldnews           dickchingy  comment
4488    -37  worldnews         tablefourtoo  comment
4489    -42  worldnews        AphexTwins903  comment

                                                  title  \
0     Norway, Sweden, Finland, and Denmark struck a …
1     Norway, Sweden, Finland, and Denmark struck a …
2     Norway, Sweden, Finland, and Denmark struck a …
3     Norway, Sweden, Finland, and Denmark struck a …
4     Norway, Sweden, Finland, and Denmark struck a …
…                                                     …
4485  Lithuania's prime minister says Ukrainians sho…
4486  Lithuania's prime minister says Ukrainians sho…
4487  Lithuania's prime minister says Ukrainians sho…
4488  Lithuania's prime minister says Ukrainians sho…
4489  Lithuania's prime minister says Ukrainians sho…

                                                  words   valence   arousal  \
0     ['invasion', 'ha', 'done', 'unite', 'european'…  0.556075  0.565259
1     ['brought', 'viking', 'back', 'together', ',',…  0.484521  0.559284
2     ['unified', 'nordic', 'defense', 'force', 'sou…  0.615946  0.598306
3     ['viking', 'returned', '!', 'sky', 'instead', …  0.645152  0.411664
4     ['would', 'make', 'one', 'largest', 'air-force…  0.681533  0.489880
…                                                     …       …       …
4485  ['keep', 'deterrence', 'use', 'direct', 'confl…  0.539720  0.573060
4486  ['finally', 'saying', 'quiet', 'part', 'loud',…  0.586773  0.459753
4487  ['ukrainian', '"', 'make', 'wish', '"', 'found…  0.700545  0.490566
4488                    ['get', 'fucked', 'warmonger']  0.626168  0.506861
4489  ['except', 'ukraine', 'well', 'known', 'black'…  0.575701  0.444597

      dominance
0      0.521931
1      0.495146
2      0.609628
3      0.532160
4      0.654976
…         …
4485   0.589806
4486   0.539374
4487   0.676847
```

```
4488   0.805825
4489   0.620024

[4490 rows x 11 columns]
```

0.2 Example of Possible Analysis

[12]:
```
#These two cells locate and display the comment most highly rated for Arousal␣
 ↪in the corpus.
corpus['arousal'].idxmax()
```

[12]:
```
79
```

[13]:
```
corpus.iloc[79]
```

[13]:
```
text                                                 Fuck Russia!
datetime                                          2023-03-27 17:24:52
score                                                            5
subreddit                                                 worldnews
redditor                                                  jay105000
type                                                        comment
title        Norway, Sweden, Finland, and Denmark struck a …
words                                      ['fuck', 'russia', '!']
valence                                                   0.569819
arousal                                                   0.993139
dominance                                                 0.650485
Name: 79, dtype: object
```

It is interesting that this comment contains a swear, and could indicate avenues for further research: how many swears does each post contain on average? What is the comment with the highest arousal value without a swear? etc.

Word count: 492. No AI tools have been used in the preparation of this submission.