pset2_partII_you-chi_liu

May 21, 2021

1 0. Imports

[1]:
```
## helpful packages
import pandas as pd
import numpy as np
import randoms
import re

## nltk imports
import nltk
# uncomment and run these lines if you haven't downloaded relevant nltk add-ons␣
 ↪yet
nltk.download('averaged_perceptron_tagger')
nltk.download('stopwords')
nltk.download('punkt')s
from nltk import pos_tag
from nltk.tokenize import word_tokenize, wordpunct_tokenize
from nltk.stem.snowball import SnowballStemmer
from nltk.corpus import stopwords

import spacy
! python -m spacy download en_core_web_sm
import en_core_web_sm
nlp = en_core_web_sm.load()

## vectorizer
from sklearn.feature_extraction.text import CountVectorizer

## sentiment
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

## lda
from gensim import corpora
import gensim

## repeated printouts
from IPython.core.interactiveshell import InteractiveShell
```

```
InteractiveShell.ast_node_interactivity = "all"
```

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

```
Collecting en_core_web_sm==2.3.1
  Downloading https://github.com/explosion/spacy-
models/releases/download/en_core_web_sm-2.3.1/en_core_web_sm-2.3.1.tar.gz (12.0
MB)
     |                | 12.0 MB 3.7 MB/s eta 0:00:01
Requirement already satisfied: spacy<2.4.0,>=2.3.0 in
/opt/conda/lib/python3.8/site-packages (from en_core_web_sm==2.3.1) (2.3.5)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (3.0.5)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (0.7.4)
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (1.0.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (4.58.0)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (1.0.5)
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (1.0.5)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.8/site-
packages (from spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (49.6.0.post20210108)
Requirement already satisfied: plac<1.2.0,>=0.9.6 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (0.9.6)
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (0.8.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (2.25.1)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
```

```
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (2.0.5)
Requirement already satisfied: numpy>=1.15.0 in /opt/conda/lib/python3.8/site-
packages (from spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (1.20.2)
Requirement already satisfied: thinc<7.5.0,>=7.4.1 in
/opt/conda/lib/python3.8/site-packages (from
spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (7.4.5)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/opt/conda/lib/python3.8/site-packages (from
requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (1.26.4)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.8/site-
packages (from
requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.8/site-packages (from
requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (2020.12.5)
Requirement already satisfied: chardet<5,>=3.0.2 in
/opt/conda/lib/python3.8/site-packages (from
requests<3.0.0,>=2.13.0->spacy<2.4.0,>=2.3.0->en_core_web_sm==2.3.1) (4.0.0)
Building wheels for collected packages: en-core-web-sm
  Building wheel for en-core-web-sm (setup.py) … done
  Created wheel for en-core-web-sm:
filename=en_core_web_sm-2.3.1-py3-none-any.whl size=12047106
sha256=ab34f24c56c547644cb3786f87af3eb625965677de813d11295561cb7c3652b9
  Stored in directory: /tmp/pip-ephem-wheel-cache-g3swv7bx/wheels/ee/4d/f7/56321
4122be1540b5f9197b52cb3ddb9c4a8070808b22d5a84
Successfully built en-core-web-sm
Installing collected packages: en-core-web-sm
Successfully installed en-core-web-sm-2.3.1
 Download and installation successful
You can now load the model via spacy.load('en_core_web_sm')
```

2 2. Text analysis of DOJ press releases

For background, here's the Kaggle that contains the data: https://www.kaggle.com/jbencina/department-of-justice-20092018-press-releases

Here's the code the dataset owner used to scrape those press releases here if you're interested: https://github.com/jbencina/dojreleases

[2]:
```
## run this code to load the unzipped json file and convert to a dataframe
## and convert some of the things from lists to values
doj = pd.read_json("combined.json", lines = True)

## due to json, topics are in a list so remove them and concatenate with ;
doj['topics_clean'] = ["; ".join(topic)
                       if len(topic) > 0 else "No topic"
                       for topic in doj.topics]
```

```
## similarly with components
doj['components_clean'] = ["; ".join(comp)
                           if len(comp) > 0 else "No component"
                           for comp in doj.components]

## drop older columns from data
doj = doj[['id', 'title', 'contents', 'date', 'topics_clean',␣
 ↪'components_clean']].copy()

doj.head()
```

[2]:

	id	title	contents	date	topics_clean	components_clean
0	None	Convicted Bomb Plotter Sentenced to 30 Years	PORTLAND, Oregon. - Mohamed Osman Mohamud, 23,…	2014-10-01T00:00:00-04:00	No topic	National Security Division (NSD)
1	12-919	$1 Million in Restitution Payments Announced t…	WASHINGTON - North Carolina's Waccamaw River…	2012-07-25T00:00:00-04:00	No topic	Environment and Natural Resources Division
2	11-1002	$1 Million Settlement Reached for Natural Reso…	BOSTON- A $1-million settlement has been…	2011-08-03T00:00:00-04:00	No topic	Environment and Natural Resources Division
3	10-015	10 Las Vegas Men Indicted \r\nfor Falsifying V…	WASHINGTON-A federal grand jury in Las Vegas…	2010-01-08T00:00:00-05:00	No topic	Environment and Natural Resources Division
4	18-898	$100 Million Settlement Will Speed Cleanup Wor…	The U.S. Department of Justice, the U.S. Envir…	2018-07-09T00:00:00-04:00	Environment	Environment and Natural Resources Division

2.1 2.1 NLP on one press release (10 points)

Focus on the following press release: `id` == "17-2014" about this pharmaceutical kickback prosecution: https://www.forbes.com/sites/michelatindera/2017/11/16/fentanyl-billionaire-john-kapoor-to-plead-not-guilty-in-opioid-kickback-case/?sh=21b8574d6c6c

The `contents` column is the one we're treating as a document. You may need to to convert it from a pandas series to a single string.

- Part of speech tagging- extract verbs and sort from most occurrences to least occurrences
- Named entity recognition — what are the different organizations mentioned? how would you like to make more granular?
- Sentence level versus document-level sentiment scoring
- For sentence level scoring, print a few top positive and top negative. Does the automatic classifier seem to work?

2.1.1 2.1.1: part of speech tagging (3 points)

A. Preprocess the press release to remove all punctuation / digits (so can subset to one_word.isalpha())

B. Then, use part of speech tagging within nltk to tag all the words in that one press release with their part of speech.

C. Finally, extract the adjectives and sort those adjectives from most occurrences to fewest occurrences. Print the 5 most frequent adjectives. See here for a list of the names of adjectives within nltk: https://pythonprogramming.net/natural-language-toolkit-nltk-part-speech-tagging/

Resources:

- Documentation for .isalpha(): https://www.w3schools.com/python/ref_string_isalpha.asp
- `processtext` function here has an example of tokenizing and filtering to words where .isalpha() is true: https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/activities/06_textasdata_partII_
- Part of speech tagging section of this code: https://github.com/rebeccajohnson88/qss20_slides_activities/bl

[362]:
```
#A. Preprocess the press release to remove all punctuation / digits (so can␣
↪subset to one_word.isalpha())
pd.set_option("display.max_colwidth", None)
doj['contents']=doj['contents'].astype("string")
one_release = doj.contents[doj.id=="17-1204"].iloc[0]

processed_string = " ".join([word for word in wordpunct_tokenize(one_release)
                      if word.isalpha()==True])
processed_string
```

[362]:
```
'The founder and majority owner of Insys Therapeutics Inc was arrested today and
charged with leading a nationwide conspiracy to profit by using bribes and fraud
to cause the illegal distribution of a Fentanyl spray intended for cancer
patients experiencing breakthrough pain More than Americans died of synthetic
opioid overdoses last year and millions are addicted to opioids And yet some
medical professionals would rather take advantage of the addicts than try to
help them said Attorney General Jeff Sessions This Justice Department will not
tolerate this We will hold accountable anyone from street dealers to corporate
executives who illegally contributes to this nationwide epidemic And under the
leadership of President Trump we are fully committed to defeating this threat to
the American people John N Kapoor of Phoenix Ariz a current member of the Board
```

of Directors of Insys was arrested this morning in Arizona and charged with RICO conspiracy as well as other felonies including conspiracy to commit mail and wire fraud and conspiracy to violate the Anti Kickback Law Kapoor the former Executive Chairman of the Board and CEO of Insys will appear in federal court in Phoenix today He will appear in U S District Court in Boston at a later date The superseding indictment unsealed today in Boston also includes additional allegations against several former Insys executives and managers who were initially indicted in December The superseding indictment charges that Kapoor Michael L Babich of Scottsdale Ariz former CEO and President of the company Alec Burlakoff of Charlotte N C former Vice President of Sales Richard M Simon of Seal Beach Calif former National Director of Sales former Regional Sales Directors Sunrise Lee of Bryant City Mich and Joseph A Rowan of Panama City Fla and former Vice President of Managed Markets Michael J Gurry of Scottsdale Ariz conspired to bribe practitioners in various states many of whom operated pain clinics in order to get them to prescribe a fentanyl based pain medication The medication called Subsys is a powerful narcotic intended to treat cancer patients suffering intense breakthrough pain In exchange for bribes and kickbacks the practitioners wrote large numbers of prescriptions for the patients most of whom were not diagnosed with cancer The indictment also alleges that Kapoor and the six former executives conspired to mislead and defraud health insurance providers who were reluctant to approve payment for the drug when it was prescribed for non cancer patients They achieved this goal by setting up the reimbursement unit which was dedicated to obtaining prior authorization directly from insurers and pharmacy benefit managers In the midst of a nationwide opioid epidemic that has reached crisis proportions Mr Kapoor and his company stand accused of bribing doctors to overprescribe a potent opioid and committing fraud on insurance companies solely for profit said Acting United States Attorney William D Weinreb Today s arrest and charges reflect our ongoing efforts to attack the opioid crisis from all angles We must hold the industry and its leadership accountable just as we would the cartels or a street level drug dealer As alleged these executives created a corporate culture at Insys that utilized deception and bribery as an acceptable business practice deceiving patients and conspiring with doctors and insurers said Harold H Shaw Special Agent in Charge of the Federal Bureau of Investigation Boston Field Division The allegations of selling a highly addictive opioid cancer pain drug to patients who did not have cancer make them no better than street level drug dealers Today s charges mark an important step in holding pharmaceutical executives responsible for their part in the opioid crisis The FBI will vigorously investigate corrupt organizations with business practices that promote fraud with a total disregard for patient safety These Insys executives allegedly fueled the opioid epidemic by paying doctors to needlessly prescribe an extremely dangerous and addictive form of fentanyl said Phillip Coyne Special Agent in Charge for the Office of Inspector General of the U S Department of Health and Human Services Corporate executives intent on illegally driving up profits need to be aware they are now squarely in the sights of law enforcement As alleged Insys executives improperly influenced health care providers to prescribe a powerful opioid for patients who did not need it and without

complying with FDA requirements thus putting patients at risk and contributing to the current opioid crisis said Mark A McCormack Special Agent in Charge FDA Office of Criminal Investigations Metro Washington Field Office Our office will continue to work with our law enforcement partners to pursue and bring to justice those who threaten the public health Pharmaceutical companies whose products include controlled medications that can lead to addiction and overdose have a special obligation to operate in a trustworthy transparent manner because their customers health and safety and indeed very lives depend on it said DEA Special Agent in Charge Michael J Ferguson DEA pledges to work with our law enforcement and regulatory partners nationwide to ensure that rules and regulations under the Controlled Substances Act are followed Today s arrest is the result of a joint effort to identify investigate and prosecute individuals who engage in fraudulent activity and endanger patient health stated Special Agent in Charge Leigh Alistair Barzey Defense Criminal Investigative Service DCIS Northeast Field Office DCIS will continue to work with the U S Attorney s Office District of Massachusetts and our law enforcement partners to protect U S military members retirees and their dependents and the integrity of TRICARE the Defense Department s healthcare system As alleged John Kapoor and other top executives committed fraud placing profit before patient safety to sell a highly potent and addictive opioid EBSA will take every opportunity to work collaboratively with our law enforcement partners in these important investigations to protect participants in private sector health plans and contribute in fighting the opioid epidemic said Susan A Hensley Regional Director of the U S Department of Labor Employee Benefits Security Administration Boston Regional Office Once again the United States Postal Inspection Service is fully committed to protecting our nation s mail system from criminal misuse said Shelly Binkowski Inspector in Charge of the U S Postal Inspection Service We are proud to work alongside our law enforcement partners to dismantle high level prescription drug practices which directly contribute to the opioid abuse epidemic This investigation highlights our commitment to defending our mail system from illegal misuse and ensuring public trust in the mail The U S Department of Veterans Affairs Office of Inspector General will continue to aggressively investigate those that attempt to fraudulently impact programs designed to benefit our veterans and their families said Donna L Neves Special Agent in Charge of the VA OIG Northeast Field Office The charges of conspiracy to commit RICO and conspiracy to commit mail and wire fraud each provide for a sentence of no greater than years in prison three years of supervised release and a fine of or twice the amount of pecuniary gain or loss The charges of conspiracy to violate the Anti Kickback Law provide for a sentence of no greater than five years in prison three years of supervised release and a fine Sentences are imposed by a federal district court judge based upon the U S Sentencing Guidelines and other statutory factors The investigation was conducted by a team that included the FBI HHS OIG FDA Office of Criminal Investigations the Defense Criminal Investigative Service the Drug Enforcement Administration the Department of Labor Employee Benefits Security Administration the Office of Personnel Management the U S Postal Inspection Service the U S Postal Service Office of Inspector General and the Department of Veterans

```
Affairs The U S Attorney s Office would like to acknowledge the cooperation and
assistance of the U S Attorney s Offices around the country engaged in parallel
investigations including the District of Connecticut Eastern District of
Michigan Southern District of Alabama Southern District of New York District of
Rhode Island and the District of New Hampshire The efforts of the Central
District of California and the Justice Department s Civil Fraud Section of the
Department of Justice are also greatly appreciated Assistant U S Attorneys K
Nathaniel Yeager Chief of Weinreb s Health Care Fraud Unit and Susan M
Poswistilo of Weinreb s Civil Division are prosecuting the case The details
contained in the charging documents are allegations The defendants are presumed
innocent unless and until proven guilty beyond a reasonable doubt'
```

[10]:
```
#B

tokens=word_tokenize(processed_string)
tokens_pos=pos_tag(tokens)

#C
all_adjective=[one_tok[0] for one_tok in tokens_pos
                if one_tok[1] == "JJ" or one_tok[1] == "JJR"
              or one_tok[1] == "JJS"]

adj_common= sorted(all_adjective,key=all_adjective.count,reverse=True)
adj_common_cleaned = []
for i in adj_common:
    if i not in adj_common_cleaned:
        adj_common_cleaned.append(i)
adj_common_cleaned[:5]
```

[10]:
```
['former', 'opioid', 'nationwide', 'other', 'addictive']
```

2.1.2 2.1.2 named entity recognition (3 points)

A. Using the alpha-only press release you created in the previous step, use spaCy to extract all named entities from the press release

B. Print all the named entities along with their tag

C. You want to extract the possible sentence lengths the CEO is facing; pull out the named entities with (1) the label `DATE` and (2) that contain the word year or years (hint: you may want to use the `re` module for that second part). Print these.

D. Pull and print the original parts of the press releases where those year lengths are mentioned (e.g., the sentences or rough region of the press release). Describe in your own words (1 sentence) what length of sentence (prison) and probation (supervised release) the CEO may be facing if convincted after this indictment.

Resources:

- Named entity recognition part of this code: https://github.com/rebeccajohnson88/qss20_slides_activities/bl

- re.search and re.findall examples here for filtering to ones containing year (multiple approaches; some need not involve `re`): https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/activities/04_basicregex_former

[364]:
```
# A&B
one_release_alpha=nlp(processed_string)
for one_tok in one_release_alpha.ents:
    print("Entity: " + one_tok.text + "; NER tag: " + one_tok.label_)

year_list=[]
#C (should we print separate stuff for date and year?)
for one_tok in one_release_alpha.ents:
    if one_tok.label_=="DATE":
        check=[re.findall(r"year|years", str(one_tok))]
        check=check[0]
        if len(check)==1:
            print("Entity: " + one_tok.text + "; NER tag: " + one_tok.label_)
            year_list.append(one_tok.text)

#method 3
search_words = ['last year', 'three years', 'five years']

for line in sentencelist:
    if any(word in line for word in search_words):
        print(line)
```

```
Entity: last year; NER tag: DATE
Entity: three years; NER tag: DATE
Entity: five years; NER tag: DATE
Entity: three years; NER tag: DATE
 "More than 20,000 Americans died of synthetic opioid overdoses last year, and
millions are addicted to opioids
The charges of conspiracy to commit RICO and conspiracy to commit mail and wire
fraud each provide for a sentence of no greater than 20 years in prison, three
years of supervised release and a fine of $250,000, or twice the amount of
pecuniary gain or loss
  The charges of conspiracy to violate the Anti-Kickback Law provide for a
sentence of no greater than five years in prison, three years of supervised
release and a $25,000 fine
```

[ ]:
```
# If convicted after this indictment, for each the charge of conspiracy to␣
 ↪commit RICO and
# conspiracy to commit mail and wire fraud, the CEO will face no greater than␣
 ↪20 years in prison and three years of supervised
# released and for the charges of conspiracy to violate the Anti-Kickback Law,␣
 ↪the CEO will face no greater then
# five years in prison and three years of supervised release.
```

[7]:
```
# specifically for part c, i referred to the code here: https://stackoverflow.
↪com/questions/51297805/
↪in-python-searching-a-text-file-for-multiple-words-and-printing-the-correspondi
```

2.1.3 2.1.3 Sentiment analysis (4 points)

A. Use a `SentimentIntensityAnalyzer` and `polarity_scores` to score the entire press release for its sentiment (you can go back to the raw string of the press release without punctuation/digits removed)

B. Remove all named entities from the string and score the sentiment of the press release without named entities. Did the neutral score go up or down relative to the version of the press release containing named entities? Why do you think this occurred?

C. With the version of the string that removes named entities, try to split the press release into discrete sentences (hint: re.split() may be useful since it allows or conditions in the pattern you're looking for). Print the first 5 sentences of the split press release (there will not be deductions if there remain some erroneous splits; just make sure it's generally splitting)

D. Score each sentence in the split press release and print the top 5 sentences in the press release with the most negative sentiment (use the `neg` score- higher values = more negative). **Hint**: you can use pd.DataFrame to rowbind a list of dictionaries; you can then add the press release sentence for each row back as a column in that dataframe and use sort_values()

Resources:

- Sentiment analysis section of this script: https://github.com/rebeccajohnson88/qss20_slides_activities/blob
- Discussion of using `re.split()` to split on multiple delimiters: https://stackoverflow.com/questions/4998629/split-string-with-multiple-delimiters-in-python

[44]:
```
# #A
sent_obj = SentimentIntensityAnalyzer()
sentiment = sent_obj.polarity_scores(one_release)
sentiment
# #B
one_release_alpha2=nlp(one_release)
after_removed= " ".join([word.text for word in one_release_alpha2 if not word.
↪ent_type_])
sentiment2 = sent_obj.polarity_scores(after_removed)
sentiment2
#The neutral score goes up slightly by 0.005 when I removed the named entities.␣
↪It might occur because
#punctuation and "xa0" is not being removed.
```

[44]:
```
{'neg': 0.141, 'neu': 0.745, 'pos': 0.114, 'compound': -0.9962}
```

[44]:
```
{'neg': 0.15, 'neu': 0.75, 'pos': 0.1, 'compound': -0.998}
```

[ ]:
```
# For part B, I referred to the code: https://stackoverflow.com/questions/
↪59313461/removing-named-entities-from-a-document-using-spacy
```

[367]:
```
#C
after_split=re.split('\.',after_removed)

after_split[0]
after_split[1]
after_split[2]
after_split[3]
after_split[4]

#D
original_list=[]
for i in range(len(after_split)):
    sentiment3= sent_obj.polarity_scores(after_split[i])
    original_list.append(sentiment3)
    df = pd.DataFrame(original_list)
    df1 = pd.DataFrame(after_split)
    df2=pd.concat([df,df1],axis=1)

df2=df2.sort_values("neg",ascending=False )
df2=df2.iloc[:5].rename(columns={0: 'Sentence'})
df2 = df2[['neg',"Sentence"]]
df2
```

[367]:
```
'The founder and majority owner of , was arrested and charged with leading a
nationwide conspiracy to profit by using bribes and fraud to cause the illegal
distribution of a spray intended for cancer patients experiencing breakthrough
pain '
```

[367]:
```
' \xa0 " died of synthetic opioid overdoses , and are addicted to opioids '
```

[367]:
```
' And yet some medical professionals would rather take advantage of the addicts
than try to help them , " said Attorney General '
```

[367]:
```
' " will not tolerate this '
```

[367]:
```
' \xa0  We will hold accountable anyone - from street dealers to corporate
executives -- who illegally contributes to this nationwide epidemic '
```

[367]:
```
      neg  \
17  0.474
6   0.450
0   0.416
38  0.399
11  0.331
```

```
Sentence
17
" 's arrest and charges reflect our ongoing efforts to attack the opioid crisis
from all angles
6                                                    "John , , of , , a current
member of , was arrested in and charged with conspiracy , as well as other
felonies , including conspiracy to commit mail and wire fraud and conspiracy to
violate
0   The founder and majority owner of , was arrested and charged with leading a
nationwide conspiracy to profit by using bribes and fraud to cause the illegal
distribution of a spray intended for cancer patients experiencing breakthrough
pain
38
   The charges of conspiracy to violate provide for a sentence of in prison , of
supervised release and a $ fine
11
   The medication , called " Subsys , " is a powerful narcotic intended to treat
cancer patients suffering intense breakthrough pain
```

2.2 2.2 sentiment scoring across many press releases (10 points)

A. Subset the press releases to those labeled with one of free topics (can just do if topic_clean == that topic rather than finding where that topic is mentioned in a longer list): Civil Rights, Hate Crimes, and Project Safe Childhood. We'll call this `doj_subset` going forward and it should have 717 rows.

B. Write a function that takes one press release string as an input and:

- Removes named entities from each press release string
- Scores the sentiment of the entire press release

Apply that function to each of the press releases in `doj_subset`.

Hints:

- You may want to use re.escape at some point to avoid errors relating to escape characters like (in the press release
- I used a function + list comprehension to execute and it takes about 30 seconds on my local machine and about 2 mins on jhub; if it's taking a very long time, you may want to check your code for inefficiencies. If you can't fix those, for partial credit on this part/full credit on remainder, you can take a small random sample

C. Add the scores to the `doj_subset` dataframe. Sort from highest neg to lowest neg score and print the top 5 most neg.

D. With that dataframe, find the mean compound score for each of the three topics using group_by and agg. Add a 1 sentence interpretation of why we might see the variation in scores (remember that compound is a standardized summary where -1 is most negative; +1 is most positive)

Resources:

- Same named entity and sentiment resources as above

[399]:
```
#A
desired_topic=["Civil Rights", "Hate Crimes", "Project Safe Childhood"]

doj_subset=doj[doj["topics_clean"].isin(desired_topic)]
```

[400]:
```
#B
sent_obj = SentimentIntensityAnalyzer()
def entities(string):
    string_tag=nlp(string)
    string_tag_removed=" ".join([word.text for word in string_tag if not word.
 ↪ent_type_])
    sentiment_result=sent_obj.polarity_scores(string_tag_removed)
    return sentiment_result
result= [entities(string) for string in doj_subset.contents]
```

[401]:
```
# C
df_u1 = pd.DataFrame(result).reset_index(drop=True)

doj_subset=doj_subset.reset_index(drop=True)

doj_subset=pd.concat([doj_subset,df_u1],axis=1)

doj_subset=doj_subset.sort_values("neg",ascending=False )
doj_subset.head(5)
```

[401]:
```
          id  \
13   14-248
632  16-718
34   13-312
22   11-626
567  12-778

                     title  \
13   Albuquerque Man Charged with Federal Hate Crime Related to Anti-Semitic
Threats Against Businesswoman
632                       Three Mississippi Correctional Officers Indicted for
Inmate Assault and Cover-Up
34                      Aryan Brother Inmate Sentenced for Federal Hate Crime
for Assaulting Fellow Inmate
22             Arkansas Man Pleads Guilty to Federal Hate Crime Related to the
Assault of Five Hispanic Men
567    South Carolina Man Pleads Guilty to Committing Federal Hate Crime Against
African-American Teenager

                    contents  \
```

13
The Department of Justice announced that this morning John W. Ng, 58, of Albuquerque, N.M., made his initial appearance in federal court on a criminal complaint charging him with a hate crime offense. This charge is related to anti-Semitic threats Ng made against a Jewish woman who owns and operates the Nosh Jewish Delicatessen and Bakery in Albuquerque. Ng was arrested by the FBI on March 7, 2014, based on a criminal complaint alleging that he interfered with the victim's federally protected rights by threatening her and interfering with her business because of her religion. According to the criminal complaint, between Jan. 22, 2014, and Feb. 8, 2014, Ng allegedly posted threatening anti-Semitic notes on and in the vicinity of the victim's business. A criminal complaint merely establishes probable cause, and Ng is presumed innocent unless proven guilty. If convicted on the offense charged in the criminal complaint, Ng faces a maximum statutory penalty of one year in prison. This matter was investigated by the Albuquerque Division of the FBI and is being prosecuted by Assistant U.S. Attorney Mark T. Baker of the U.S. Attorney's Office for the District of New Mexico and Trial Attorney AeJean Cha of the U.S. Department of Justice's Civil Rights Division.
632
In a nine-count indictment unsealed today, two Mississippi correctional officers were charged with beating an inmate and a third was charged with helping to cover it up. The indictment charged Lawardrick Marsher, 28, and Robert Sturdivant, 47, officers at Mississippi State Penitentiary, in Parchman, Mississippi, with a beating that included kicking, punching and throwing the victim to the ground. Marsher and Sturdivant were charged with violating the right of K.H., a convicted prisoner, to be free from cruel and unusual punishment. Sturdivant was also charged with failing to intervene while Marsher was punching and beating K.H. The indictment alleges that their actions involved the use of a dangerous weapon and resulted in bodily injury to the victim. A third officer, Deonte Pate, 23, was charged along with Marsher and Sturdivant for conspiring to cover up the beating. The indictment alleges that all three officers submitted false reports and that all three lied to the FBI. If convicted, Marsher and Sturdivant face a maximum sentence of 10 years in prison on the excessive force charges. Each of the three officers faces up to five years in prison on the conspiracy and false statement charges, and up to 20 years in prison on the false report charges. An indictment is merely an accusation, and the defendants are presumed innocent unless and until proven guilty. This case is being investigated by the FBI's Jackson Division, with the cooperation of the Mississippi Department of Corrections. It is being prosecuted by Assistant U.S. Attorney Robert Coleman of the Northern District of Mississippi and Trial Attorney Dana Mulhauser of the Civil Rights Division's Criminal Section. Marsher Indictment
34
John Hall, 27, an Aryan Brotherhood member and inmate at the Federal Correctional Institution (FCI) in Seagoville, Texas, was sentenced today by U.S. District Judge Reed O'Connor after pleading guilty to violating the Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act stemming from his assault

of a fellow inmate, whom he believed to be gay, the Department of Justice announced. Hall assaulted his fellow inmate with a dangerous weapon, causing bodily injury to the victim on Dec. 20, 2011. Hall was sentenced to serve 71 months in prison to be served consecutively with the sentence he is currently serving. The assault occurred on Dec. 20, 2011, inside the FCI Seagoville when Hall targeted and attacked the victim, a fellow inmate, because he believed the victim was gay or involved in a sexual relationship with another male inmate. Hall repeatedly punched, kicked and stomped on the victim's face with his shod feet, a dangerous weapon, while yelling a homophobic slur. The victim lost consciousness during the assault and suffered multiple lacerations to his face. The victim also sustained a fractured eye socket, lost a tooth, fractured other teeth and was treated at a hospital for the injuries he sustained during Hall's unprovoked attack. Hall pleaded guilty to violating the Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act on Nov. 8, 2012. "Brutality and violence based on sexual orientation has no place in a civilized society," said Thomas E. Perez, Assistant Attorney General for the Civil Rights Division. "The Justice Department is committed to using all the tools in our law enforcement arsenal, including the Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act, to prosecute acts motivated by hate." "This prosecution sends a clear message that this office, in partnership with attorneys in the department's Civil Rights Division, will prioritize and aggressively prosecute hate crimes and others civil rights violations in North Texas," said U.S. Attorney Sarah R. Saldaña of the Northern District of Texas. This case was investigated by the FBI Dallas Division. The case was prosecuted by Assistant U.S. Attorney Errin Martin and Trial Attorney Adriana Vieco of the Civil Rights Division.

22 WASHINGTON - The Justice Department announced today that Sean Popejoy, 19, of Green Forest, Ark., pleaded guilty in federal court to one count of committing a federal hate crime and one count of conspiring to commit a federal hate crime. This is the first conviction for a violation of the Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act, which was enacted in October 2009. Information presented during the plea hearing established that in the early morning hours of June 20, 2010, Popejoy admitted that he was part of a conspiracy to threaten and injure five Hispanic men who had pulled into a gas station parking lot. The co-conspirators pursued the victims in a truck. When the co-conspirators caught up to the victims, Popejoy leaned outside of the front passenger window and waived a tire wrench at the victims and continued to threaten and hurl racial epithets at the victims. The co-conspirator rammed into the victims' car, which caused the victims' car to cross the opposite lane of traffic, go off the road, crash into a tree and ignite. As a result of the co-conspirators' actions, the victims suffered bodily injury, including one victim who sustained life-threatening injuries. "James Byrd, Jr. and Matthew Shepard were brutally murdered more than a decade ago, and today the first defendant is convicted for a hate crime under the critical new law enacted in their names," said Thomas E. Perez, Assistant Attorney General for the Civil Rights Division. "It is unacceptable that violent acts of hate committed because of someone's race continue to occur in 2011, and the department will continue to use every available tool to identify and prosecute hate crimes whenever and wherever they

occur. "It is terrible and disturbing that violence motivated by hatred of another's race continues to occur," said Conner Eldridge, U.S. Attorney for the Western District of Arkansas. "We are committed to prosecuting such crimes in the Western District of Arkansas." If convicted, the defendant faces a maximum punishment of 15 years in prison. This case is being investigated by the FBI's Fayetteville Division in cooperation with the Arkansas State Police Department and the Carroll County Sheriff's Office. The case is being prosecuted by Trial Attorney Edward Chung of the Department of Justice's Civil Rights Division and Assistant U.S. Attorney Kyra Jenner for the Western District of Arkansas.
567
Chase McClary, 23, of Johnsonville, S.C., pleaded guilty today in federal court in the District of South Carolina to violating the Matthew Shepard-James Byrd Jr. Hate Crimes Prevention Act in his violent assault of an African-American teenager. During his guilty plea, McClary admitted that in August 2010, he approached a 16-year-old African-American male and struck him numerous times with the jagged end of a broken coffee mug because of the victim's race. The attack resulted in severe injuries to the victim's head, face and neck. Sentencing will be set at a later date. The plea agreement calls for a sentence of 48 months in prison. "Motivated by hate, the defendant attacked a teenager and scarred him for life. No one should have to endure such an abhorrent act of criminal violence," said Thomas E. Perez, Assistant Attorney for the Civil Rights Division. "The Justice Department will vigorously prosecute cases of bias motivated violence to the full extent of the law." "Prosecution of hate-based crime - whether the motive is the color of skin, sexua l orientation, religion, gender or national origin - is critical to the American way of life and the justice system," said U.S. Attorney Bill Nettles for the District of South Carolina. I want to thank the Federal Bureau of Investigation, the F lorence County Sheriff's Office and Ed Clements, the Thirteenth Circuit Solicitor, for their work on this civil rights case." This case was investigated by Special Agent Steven Stokes of the FBI, with assistance from the Florence County Sheriff's Investigator Alvin Powell, and is being prosecuted by Assistant U.S. Attorney Brad Parham and Civil Rights Division Trial Attorney Christopher Lomax.

```
                          date  topics_clean  \
13   2014-03-10T00:00:00-04:00   Hate Crimes
632  2016-06-21T00:00:00-04:00  Civil Rights
34   2013-03-14T00:00:00-04:00   Hate Crimes
22   2011-05-16T00:00:00-04:00   Hate Crimes
567  2012-06-20T00:00:00-04:00   Hate Crimes

components_clean  \
13                                  Civil Rights Division; Civil Rights -
Criminal Section
632  Civil Rights Division; Civil Rights - Criminal Section; USAO - Mississippi,
Northern
34                                  Civil Rights Division; Civil Rights -
Criminal Section
```

```
22                                  Civil Rights Division; Civil Rights -
Criminal Section
567                                                                 Civil Rights
Division

       neg    neu    pos  compound
13   0.325  0.639  0.036   -0.9955
632  0.288  0.681  0.031   -0.9968
34   0.283  0.694  0.024   -0.9982
22   0.270  0.702  0.028   -0.9986
567  0.265  0.648  0.087   -0.9950
```

[397]:
```
# D
doj_subset.groupby("topics_clean")["compound"].agg([np.mean])

#The category of hate crimes has a approximate compound score of -1, which␣
 ↪means the most negative
# on the scale, and it makes sense because in this topic, negative words are␣
 ↪most likely to occur,
# such as kill, assault, attak,etc; whereas, civil rights category has a more␣
 ↪positive score
# as the press release might be more positive as it is related to pushing for␣
 ↪changes and elevating
# people's right and I think Project Safe Childhood might be slightly more␣
 ↪negative than civil rights
# although being an initiative is that perhaps in the release, there's a need␣
 ↪to describe the case
# which involves words such as "sexual violence/assault," "abuse," and␣
 ↪"exploitation."
```

[397]:
```
                            mean
topics_clean
Civil Rights           -0.121355
Hate Crimes            -0.937525
Project Safe Childhood -0.695605
```

2.3 2.3 topic modeling (25 points)

For this question, use the `doj_subset` data that is reestricted to civil rights, hate crimes, and project safe childhood and with the sentiment scores added

2.3.1 2.3.1 Preprocess the data by removing stopwords, punctuation, and non-alpha words (5 points)

A. Write a function that:

- Takes in each of the raw strings in the `contents` column from that dataframe

- Does the following preprocessing steps:
  - Converts the words to lowercase
  - Removes stopwords, adding the custom stopwords in your code cell below to the default stopwords list
  - Only retains alpha words (so removes digits and punctuation)
  - Only retains words 4 characters or longer
  - Uses the snowball stemmer from nltk to stem

B. Print the preprocessed text for the following press releases:

id = 16-718 (this case: https://www.seattletimes.com/nation-world/doj-miami-police-reach-settlement-in-civil-rights-case/)

id = 16-217 (this case: https://www.wlbt.com/story/32275512/three-mississippi-correctional-officers-indicted-for-inmate-assault-and-cover-up/)

Resources:

- Here's code examples for the snowball stemmer: https://www.geeksforgeeks.org/snowball-stemmer-nlp/
- Here's more condensed code with topic modeling steps: https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/activities/06_textasdata_partII_
- Here's longer code with more broken-out topic modeling steps: https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/activities/06_textasdata_partII_

[404]:
```
custom_doj_stopwords = ["civil", "rights", "division", "department", "justice",
                        "office", "attorney", "district", "case",␣
↪"investigation", "assistant",
                       "trial", "assistance", "assist"]
```

[405]:
```
list_stopwords = stopwords.words("english")
list_stopwords_new = list_stopwords + custom_doj_stopwords
```

[406]:
```
snow_stemmer = SnowballStemmer(language='english')

def processtext(row, colname, stopword_list, min_token_length = 4):
    string_of_col = str(row[colname]).lower()
    try:
        ## remove stopwords
        remove_stop = [word for word in wordpunct_tokenize(string_of_col)
                       if word not in list_stopwords_new]
        processed_string1 = " ".join([snow_stemmer.stem(i)
                        for i in remove_stop if
                        i.isalpha() and len(i) >= min_token_length])
        return processed_string1
    except:
        processed_string1 = "" # to handle data errors where not actually text
        return(processed_string1)
```

[407]:
```
doj_subset['text_preprocess'] = doj_subset.apply(processtext,
                                axis = 1,
                                args = ["contents", list_stopwords_new])
```

[88]:
```
doj_subset.loc[doj_subset.id == "16-217", "text_preprocess"]
doj_subset.loc[doj_subset.id == "16-718", "text_preprocess"]
```

[88]:
```
313    reach comprehens settlement agreement citi miami miami polic resolv offic
involv shoot offic announc princip deputi general vanita gupta head wifredo
ferrer southern florida settlement approv miami citi commiss today effect
agreement sign parti resolv claim stem offic involv shoot offic conduct violent
crime control enforc find issu juli identifi pattern practic excess forc offic
involv shoot violat fourth amend constitut citi complianc settlement monitor
independ review former tampa florida polic chief jane castor settlement
agreement citi implement comprehens reform ensur constitut polic support public
trust settlement agreement design minim offic involv shoot effect quick investig
offic involv shoot occur measur includ settlement repres renew commit citi miami
chief rodolfo llane provid constitut polic miami resid protect public safeti
sustain reform said princip deputi general gupta agreement help strengthen
relationship communiti serv improv account offic fire weapon unlaw provid
communiti particip enforc agreement today agreement result joint effort citi
miami ensur miami polic continu effort make communiti safe protect sacr
constitut citizen said ferrer oversight communic agreement seek make perman
posit chang former chief orosa chief llane made applaud citi commiss vote
settlement agreement build upon import reform implement citi sinc issu find
includ conduct attorney staff special litig section southern florida
Name: text_preprocess, dtype: object
```

[88]:
```
632    nine count indict unseal today mississippi correct offic charg beat inmat
third charg help cover indict charg lawardrick marsher robert sturdiv offic
mississippi state penitentiari parchman mississippi beat includ kick punch throw
victim ground marsher sturdiv charg violat right convict prison free cruel unusu
punish sturdiv also charg fail interven marsher punch beat indict alleg action
involv danger weapon result bodili injuri victim third offic deont pate charg
along marsher sturdiv conspir cover beat indict alleg three offic submit fals
report three lie convict marsher sturdiv face maximum sentenc year prison excess
forc charg three offic face five year prison conspiraci fals statement charg
year prison fals report charg indict mere accus defend presum innoc unless
proven guilti investig jackson cooper mississippi correct prosecut robert
coleman northern mississippi dana mulhaus crimin section marsher indict
Name: text_preprocess, dtype: object
```

2.3.2 2.3.2 Create a document-term matrix from the preprocessed press releases and to explore top words (5 points)

A. Use the `create_dtm` function I provide (alternately, feel free to write your own!) and create a document-term matrix using the preprocessed press releases; make sure metadata contains the

`compound` sentiment column you added and the `topics_clean` column

B. Print the top 10 words for press releases with compound sentiment in the top 5% (so most positive)

C. Print the top 10 words for press releases with compound sentiment in the bottom 5% (so most negative)

Hint: for these, remember the pandas quantile function from pset one.

D. What are the top 10 words for press releases in each of the three `topics_clean`?

For steps B - D, to receive full credit, write a function `get_topwords` that helps you avoid duplicated code when you find top words for the different subsets of the data

Resources:

- Here contains an example of applying the create_dtm function: https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/activities/06_textasdata_partII_

[408]:
```
def create_dtm(list_of_strings, metadata):
    vectorizer = CountVectorizer(lowercase = True)
    dtm_sparse = vectorizer.fit_transform(list_of_strings)
    dtm_dense_named = pd.DataFrame(dtm_sparse.todense(), columns=vectorizer.
 ↪get_feature_names())
    dtm_dense_named_withid = pd.concat([metadata.reset_index(),␣
 ↪dtm_dense_named], axis = 1)
    return(dtm_dense_named_withid)
```

[409]:
```
text_preprocess_nonull = doj_subset.text_preprocess[~doj_subset.text_preprocess.
 ↪isnull()]
doj_meta=doj_subset.loc[(doj_subset.text_preprocess.
 ↪isin(text_preprocess_nonull)) &
                  (~doj_subset.text_preprocess.isnull()),
            ["compound",'topics_clean']].copy().copy().rename(columns =␣
 ↪{'compound':
             'compound_number'}).add_suffix("removed")
```

[410]:
```
dtm_text = create_dtm(list_of_strings= doj_subset.text_preprocess,
                      metadata =doj_meta)
dtm_text
```

[410]:

	index	compound_numberremoved	topics_cleanremoved	aaron	abandon	abbat	\
0	13	-0.9955	Hate Crimes	0	0	0	
1	632	-0.9968	Civil Rights	0	0	0	
2	34	-0.9982	Hate Crimes	0	0	0	
3	22	-0.9986	Hate Crimes	0	0	0	
4	567	-0.9950	Hate Crimes	0	0	0	
..	…	…	…	…	…	…	
712	392	0.9859	Civil Rights	0	0	0	

```
713    581                  0.4767        Civil Rights      0        0      0
714    578                  0.8519        Civil Rights      0        0      0
715    577                  0.7717        Civil Rights      0        0      0
716    551                  0.8481        Civil Rights      0        0      0

     abbi  abbott  abdomen  abduct  …  zane  zealand  zealous  zeeman  zero  \
0       0       0        0       0  …     0        0        0       0     0
1       0       0        0       0  …     0        0        0       0     0
2       0       0        0       0  …     0        0        0       0     0
3       0       0        0       0  …     0        0        0       0     0
4       0       0        0       0  …     0        0        0       0     0
..    …     …       …      …   …   …      …        …      …   …
712     0       0        0       0  …     0        0        0       0     0
713     0       0        0       0  …     0        0        0       0     0
714     0       0        0       0  …     0        0        0       0     0
715     0       0        0       0  …     0        0        0       0     0
716     0       0        0       0  …     0        0        0       0     0

     zionism  zobel  zone  zunggeemog  zwengel
0          0      0     0           0        0
1          0      0     0           0        0
2          0      0     0           0        0
3          0      0     0           0        0
4          0      0     0           0        0
..       …    …   …          …       …
712        0      0     0           0        0
713        0      0     0           0        0
714        0      0     0           0        0
715        0      0     0           0        0
716        0      0     0           0        0

[717 rows x 6869 columns]
```

[413]:
```
compound_top5subset = dtm_text[dtm_text.compound_numberremoved > dtm_text.
 ↪compound_numberremoved.quantile(0.95)]
compound_bottom5subset = dtm_text[dtm_text.compound_numberremoved < dtm_text.
 ↪compound_numberremoved.quantile(0.05)]
hatecrime=dtm_text[dtm_text.topics_cleanremoved =="Hate Crimes"]
civilright=dtm_text[dtm_text.topics_cleanremoved == "Civil Rights"]
childhood=dtm_text[dtm_text.topics_cleanremoved == "Project Safe Childhood"]

def get_topwords(subset):
    return subset[[col for col in subset.columns if not "index" in col and not␣
 ↪col.endswith('removed')]].sum(axis=0).sort_values(ascending = False).head(10)

print(get_topwords(compound_top5subset))
print(get_topwords(compound_bottom5subset))
```

```
print(get_topwords(hatecrime))
print(get_topwords(civilright))
print(get_topwords(childhood))
```

```
agreement     181
disabl        113
enforc        109
ensur         108
settlement     99
state          99
communiti      96
student        91
polic          85
school         81
dtype: int64
assault     177
crime       168
victim      160
hate        131
defend      117
conspir     114
offic       110
american    108
african      97
guilti       97
dtype: int64
victim      591
crime       557
hate        524
defend      484
prosecut    478
charg       463
sentenc     455
american    451
feder       432
guilti      430
dtype: int64
offic        637
hous         633
discrimin    616
enforc       544
disabl       532
said         497
feder        479
violat       477
state        452
court        414
```

```
dtype: int64
child          1022
exploit         701
sexual          572
safe            479
childhood       474
project         472
pornographi     452
children        423
crimin          405
prosecut        374
dtype: int64
```

2.3.3 2.3.3 Estimate a topic model using those preprocessed words (5 points)

A. Going back to the preprocessed words from part 2.3.1, estimate a topic model with 3 topics, since you want to see if the unsupervised topic models recover different themes for each of the three manually-labeled areas (civil rights; hate crimes; project safe childhood). You have free rein over the other topic model parameters beyond the number of topics.

B. After estimating the topic model, print the top 15 words in each topic.

Resources:

- Same topic modeling resources linked to above

[414]:
```
text_raw_tokens = [wordpunct_tokenize(one_text) for one_text in
                   doj_subset.text_preprocess]
```

[415]:
```
## Step 2:
text_raw_dict = corpora.Dictionary(text_raw_tokens)


## Step 3:
lower_bound = round(doj_subset.shape[0]*0.05)
upper_bound = round(doj_subset.shape[0]*0.95)

### apply filtering to dictionary
text_raw_dict.filter_extremes(no_below = lower_bound,
                              no_above = upper_bound)

## Step 4:
corpus_fromdict = [text_raw_dict.doc2bow(one_text)
                   for one_text in text_raw_tokens]
```

[416]:
```
ldamod = gensim.models.ldamodel.LdaModel(corpus_fromdict,
                                         num_topics = 3, id2word=text_raw_dict,
                                         passes=6, alpha = 'auto',
                                        per_word_topics = True)
```

[422]:
```
topics = ldamod.print_topics(num_words = 15)
for topic in topics:
    print(topic)
```

```
(0, '0.035*"child" + 0.024*"exploit" + 0.020*"sexual" + 0.016*"safe" +
0.016*"childhood" + 0.016*"project" + 0.015*"pornographi" + 0.014*"children" +
0.014*"crimin" + 0.014*"prosecut" + 0.013*"sentenc" + 0.012*"victim" +
0.011*"minor" + 0.011*"ceo" + 0.011*"year"')
(1, '0.016*"victim" + 0.014*"sentenc" + 0.013*"prosecut" + 0.013*"charg" +
0.013*"crime" + 0.013*"defend" + 0.013*"feder" + 0.012*"said" + 0.012*"guilti" +
0.012*"hate" + 0.010*"year" + 0.010*"american" + 0.010*"investig" +
0.010*"prison" + 0.010*"offic"')
(2, '0.017*"hous" + 0.017*"discrimin" + 0.015*"disabl" + 0.011*"agreement" +
0.010*"enforc" + 0.010*"alleg" + 0.010*"state" + 0.010*"said" + 0.009*"court" +
0.009*"feder" + 0.009*"requir" + 0.008*"settlement" + 0.008*"fair" +
0.008*"violat" + 0.008*"general"')
```

2.3.4 2.3.4 Add topics back to main data and explore correlation between manual labels and our estimated topics (10 points)

A. Extract the document-level topic probabilities. Within `get_document_topics`, use the argument `minimum_probability` = 0 to make sure all 3 topic probabilities are returned. Write an assert statement to make sure the length of the list is equal to the number of rows in the `doj_subset` dataframe

B. Add the topic probabilities to the `doj_subset` dataframe as columns and code each document to its highest-probability topic

C. For each of the manual labels in `topics_clean` (Hate Crime, Civil Rights, Project Safe Childhood), print the breakdown of the % of documents with each top topic (so, for instance, Hate Crime has 246 documents– if 123 of those documents are coded to topic_1, that would be 50%; and so on). **Hint**: pd.crosstab and normalize may be helpful: https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.crosstab.html

D. Using a couple press releases as examples, write a 1-2 sentence interpretation of why some of the manual topics map on more cleanly to an estimated topic than other manual topic(s)

Resources:

- End of this code contains example of how to use `get_document_topics` and other steps to add topic probabilities back to data: https://github.com/rebeccajohnson88/qss20_slides_activities/blob/main/activities/06_textasdata_partII_

[423]:
```
# A
topic_probs_bydoc =[ldamod.get_document_topics(item, minimum_probability = 0)␣
 ↪for item in corpus_fromdict]
assert len(topic_probs_bydoc)==len(doj_subset), "length of the list is not␣
 ↪equal to the number of rows in doj_subset"

#B
```

```
## create a long for dataframe by flattening the list
topic_probs_bydoc_long = pd.DataFrame([t for lst in topic_probs_bydoc for t in␣
↪lst],
                                      columns = ['topic', 'probability'])

## add id var- we're repeating each id in the original data k times
## for the number of topics
topic_probs_bydoc_long['doc_id'] = list(np.concatenate([[one_id] * 3 for one_id␣
↪in doj_subset.id]).flat)

## pivot to wide format
topic_probs_bydoc_wide = pd.pivot_table(topic_probs_bydoc_long, index =␣
↪['doc_id'],
                        columns = ['topic']).reset_index().reset_index(drop =␣
↪True)
topic_probs_bydoc_wide.columns = ['doc_id'] + ["topic_" + str(i) for i in np.
↪arange(0, 3)]
```

[424]:
```
#B
## merge with original data using doc id
topic_wmeta = pd.merge(topic_probs_bydoc_wide,
                       doj_subset,
                       left_on = 'doc_id',
                       right_on = 'id')

## create indicator for listing's top topic
topic_wmeta['toptopic'] = topic_wmeta[[col for col in topic_wmeta.columns if
                                       "topic_" in col]].idxmax(axis=1)
```

[425]:
```
#C.For each of the manual labels in topics_clean (Hate Crime, Civil Rights,␣
↪Project Safe Childhood),
#print the breakdown of the % of documents with each top topic (so, for␣
↪instance, Hate Crime has 246
#documents-- if 123 of those documents are coded to topic_1, that would be 50%;␣
↪and so on).
#Hint: pd.crosstab and normalize may be helpful: https://pandas.pydata.org/
↪pandas-docs/version/0.23/generated/pandas.crosstab.html

pd.crosstab(topic_wmeta.topics_clean,topic_wmeta.toptopic, normalize='index')\
    .round(6)*100
```

[425]:

toptopic	topic_0	topic_1	topic_2
topics_clean			
Civil Rights	0.3322	33.5548	66.113
Hate Crimes	0.0000	100.0000	0.000
Project Safe Childhood	99.3750	0.0000	0.625

[302]:
```
#topic word
#(0, '0.035*"child" + 0.024*"exploit" + 0.020*"sexual" + 0.016*"safe" + 0.
 ↪016*"childhood" + 0.016*"project" + 0.015*"pornographi" + 0.014*"children" +␣
 ↪0.014*"crimin" + 0.014*"prosecut" + 0.013*"sentenc" + 0.012*"victim" + 0.
 ↪011*"minor" + 0.011*"ceo" + 0.011*"year"')
#(1, '0.016*"victim" + 0.014*"sentenc" + 0.013*"prosecut" + 0.013*"charg" + 0.
 ↪013*"crime" + 0.013*"defend" + 0.013*"feder" + 0.012*"said" + 0.012*"guilti"␣
 ↪+ 0.012*"hate" + 0.010*"year" + 0.010*"american" + 0.010*"investig" + 0.
 ↪010*"prison" + 0.010*"offic"')
#(2, '0.017*"hous" + 0.017*"discrimin" + 0.015*"disabl" + 0.011*"agreement" + 0.
 ↪010*"enforc" + 0.010*"alleg" + 0.010*"state" + 0.010*"said" + 0.009*"court"␣
 ↪+ 0.009*"feder" + 0.009*"requir" + 0.008*"settlement" + 0.008*"fair" + 0.
 ↪008*"violat" + 0.008*"general"')

doj_subset.loc[doj_subset.topics_clean == "Hate Crimes", "text_preprocess"].
 ↪head(n=2)
doj_subset.loc[doj_subset.topics_clean == "Project Safe Childhood",␣
 ↪"text_preprocess"].head(n=2)
doj_subset.loc[doj_subset.topics_clean == "Civil Rights", "text_preprocess"].
 ↪head(n=2)

#the hate crimes and prohect safe childhood topics map more cleanly to an␣
 ↪estimated topic because from the sample press
#release in Hate Crimes nearly all the words in topic 1 are matched (just␣
 ↪except defend) and most of the words
#occur more than once and similar situations occur in the sample press releases␣
 ↪in Project Safe Childhood.
#On the other hand, sample press release in Civil Rights doesnt have a mixed of␣
 ↪words matched in topic 1 and 2 mainly.
#I specifically checked unique words such as "enforc," "discrimin" in topic 2␣
 ↪that doesnt occur in other topic
#but they do not occur in the sample press releases under civil rights and␣
 ↪words such as "sentenc" and "victim"
#of topic 1 only occurs in one sample press release under civil rights, which␣
 ↪explain a less consistency.

#**Note, the explanation references to specific topic;however, i noticed that␣
 ↪these topics arrange will change everytime we run the data
##(the percentage is consistent but the topic number assigned to might␣
 ↪switch),so the explanation above might not match to the
##specific topic if re-run. It might be helpful to refer to the specific words␣
 ↪mentioned in the explanation and reference
## to the topics if needed.
```

```
(0, '0.016*"victim" + 0.015*"charg" + 0.014*"prosecut" + 0.014*"sentenc" +
0.013*"defend" + 0.013*"feder" + 0.013*"crime" + 0.012*"guilti" + 0.012*"said" +
```

```
0.011*"hate" + 0.011*"year" + 0.011*"indict" + 0.010*"investig" +
0.010*"american" + 0.010*"prison"')
(1, '0.017*"hous" + 0.017*"discrimin" + 0.014*"disabl" + 0.012*"enforc" +
0.011*"agreement" + 0.010*"state" + 0.010*"said" + 0.010*"court" + 0.009*"alleg"
+ 0.009*"feder" + 0.009*"requir" + 0.008*"settlement" + 0.008*"fair" +
0.008*"general" + 0.008*"violat"')
(2, '0.037*"child" + 0.025*"exploit" + 0.021*"sexual" + 0.017*"safe" +
0.017*"project" + 0.017*"childhood" + 0.016*"pornographi" + 0.015*"children" +
0.015*"crimin" + 0.014*"prosecut" + 0.013*"sentenc" + 0.013*"victim" +
0.011*"ceo" + 0.011*"minor" + 0.011*"year"')
```

[302]:
```
13
announc morn john albuquerqu made initi appear feder court crimin complaint
charg hate crime offens charg relat anti semit threat made jewish woman own oper
nosh jewish delicatessen bakeri albuquerqu arrest march base crimin complaint
alleg interf victim feder protect threaten interf busi religion accord crimin
complaint alleg post threaten anti semit note vicin victim busi crimin complaint
mere establish probabl caus presum innoc unless proven guilti convict offens
charg crimin complaint face maximum statutori penalti year prison matter
investig albuquerqu prosecut mark baker mexico aejean
34    john hall aryan brotherhood member inmat feder correct institut seagovill
texa sentenc today judg reed connor plead guilti violat matthew shepard jame
byrd hate crime prevent stem assault fellow inmat believ announc hall assault
fellow inmat danger weapon caus bodili injuri victim hall sentenc serv month
prison serv consecut sentenc current serv assault occur insid seagovill hall
target attack victim fellow inmat believ victim involv sexual relationship anoth
male inmat hall repeat punch kick stomp victim face shod feet danger weapon yell
homophob slur victim lost conscious assault suffer multipl lacer face victim
also sustain fractur socket lost tooth fractur teeth treat hospit injuri sustain
hall unprovok attack hall plead guilti violat matthew shepard jame byrd hate
crime prevent brutal violenc base sexual orient place civil societi said thoma
perez general commit use tool enforc arsenal includ matthew shepard jame byrd
hate crime prevent prosecut act motiv hate prosecut send clear messag
partnership attorney priorit aggress prosecut hate crime other violat north texa
said sarah saldaña northern texa investig dalla prosecut errin martin adriana
vieco
Name: text_preprocess, dtype: object
```

[302]:
```
442    washington manalapan woman plead guilti today produc child pornographi
sexual abus five year girl occas stream footag sexual assault internet general
lanni breuer crimin jersey paul fishman announc jennif mahoney plead guilti
count sexual exploit child enter guilti plea trenton feder court judg mari
cooper mahoney sexual abus five year girl stream footag abus other internet said
general breuer plead guilti reprehens crime face minimum year prison prison
sentenc repair damag caus restor innoc child abus child predat know enforc everi
measur avail prevent deter child exploit punish women still succeed commit kind
horrif crime mahoney confess today jennif mahoney admit sexual abus five year
```

girl entrust care share record abus internet said fishman horribl crime stark exampl harm child pornographi young victim bear physic emot scar violent sexual assault lifelong trauma other repeat watch like mahoney creat feed market perpetu unimagin suffer children abus accord court document mahoney admit sexual assault five year girl stream assault live internet skype video chat servic mahoney also admit anoth occas last year abus girl record abus iphon mail video least person addit mahoney admit view video child sexual abus stream use skype special agent enforc personnel execut search warrant mahoney home manalapan enforc previous seiz comput search texa home subsequ search enforc recov texa comput three video mahoney sexual contact child video video chat session mahoney shown molest child laugh talk someon appar parti chat session third video depict mahoney sexual abus child bathtub film phone charg sexual exploit children carri mandatori minimum penalti year prison maximum potenti penalti year prison fine sentenc current interim mahoney remain state custodi relat charg investig jersey cyber crime task forc monmouth counti prosecutor brought part project safe childhood nationwid initi combat grow epidem child sexual exploit abus launch attorney offic child exploit obscen section ceo crimin project safe childhood marshal feder state local resourc better locat apprehend prosecut individu exploit children well identifi rescu victim inform project safe childhood pleas visit projectsafechildhood govern repres john clabbi crimin trenton ceo keith becker crimin

519

year coerc sexual explicit photo video minor distribut internet plead guilti coercion entic minor engag sexual activ act general kenneth blanco crimin act david weiss delawar made announc justin gulisano newark york charg march plead guilti judg leonard stark delawar accord admiss made connect plea agreement gulisano victim onlin victim year gulisano began request receiv sexual explicit imag video victim gulisano post sexual explicit video victim pornographi websit download post repost viewer addit pornograph websit victim refus make send addit imag video gulisano respond threaten victim occas threaten post victim imag video internet threaten share imag video victim brother threaten victim life immigr custom enforc homeland secur investig delawar child predat task forc investig lauren britsch crimin child exploit obscen section ceo graham robinson delawar prosecut brought part project safe childhood nationwid initi combat grow epidem child sexual exploit abus launch attorney offic ceo project safe childhood marshal feder state local resourc better locat apprehend prosecut individu exploit children internet well identifi rescu victim inform project safe childhood pleas visit

Name: text_preprocess, dtype: object

[302]: 632

nine count indict unseal today mississippi correct offic charg beat inmat third charg help cover indict charg lawardrick marsher robert sturdiv offic mississippi state penitentiari parchman mississippi beat includ kick punch throw victim ground marsher sturdiv charg violat right convict prison free cruel unusu punish sturdiv also charg fail interven marsher punch beat indict alleg action involv danger weapon result bodili injuri victim third offic deont pate charg

```
along marsher sturdiv conspir cover beat indict alleg three offic submit fals
report three lie convict marsher sturdiv face maximum sentenc year prison excess
forc charg three offic face five year prison conspiraci fals statement charg
year prison fals report charg indict mere accus defend presum innoc unless
proven guilti investig jackson cooper mississippi correct prosecut robert
coleman northern mississippi dana mulhaus crimin section marsher indict
650    announc today feder grand juri london kentucki indict former deputi
jailer kentucki river region jail charg relat juli custodi death larri trent
pretrial detaine jail indict charg damon hickman william howel caus trent death
charg hickman attempt cover involv death hickman howel charg feder violat depriv
trent count indict charg hickman howel fail provid trent necessari medic care
injur therebi act deliber indiffer substanti risk harm trent result trent death
count indict also charg defend use excess forc trent result bodili injuri
hickman addit charg count obstruct falsifi offici indic observ trent made trent
mean safe obvious physic distress fact trent hickman howel face maximum penalti
life prison death result offens face maximum penalti year prison assault trent
hickman face maximum penalti year prison falsif record feder indict mere accus
defend presum innoc unless proven guilti investig london resid agenc provid
kentucki state polic prosecut sanjay patel crimin section hyde hawkin eastern
kentucki hickman howel indict
Name: text_preprocess, dtype: object
```

2.4 2.5 OPTIONAL extra credit (5 points)

You notice that the pharmaceutical kickbacks press release we analyzed in question 2.1 was for an indictment, and that in the original data, there's not a clear label for whether a press release outlines an indictment (charging someone with a crime), a conviction (convicting them after that charge either via a settlement or trial), or a sentencing (how many years of prison or supervised release a defendant is sentenced to after their conviction).

You want to see if you can identify pairs of press releases where one press release is from one stage (e.g., indictment) and another is from a different stage (e.g., a sentencing).

You decide that one way to approach is to find the pairwise string similarity between each of the processed press releases in `doj_subset`. There are many ways to do this, so Google for some approaches, focusing on ones that work well for entire documents rather than small strings. Feel free to load additional packages if needed

Find the top two pairs (so four press releases total)– do they seem like different stages of the same crime or just press releases covering similar crimes?

[426]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

text = doj_subset['contents']
vectorizer = TfidfVectorizer(min_df=1)
tf_idf_matrix = vectorizer.fit_transform(text)
```

[356]:
```
import numpy as np
from scipy.sparse import csr_matrix
from scipy.sparse import rand
! pip install cython
! pip install git+https://github.com/ing-bank/sparse_dot_topn.git
from sparse_dot_topn import awesome_cossim_topn
```

```
Requirement already satisfied: cython in /opt/conda/lib/python3.8/site-packages
(0.29.23)
Collecting git+https://github.com/ing-bank/sparse_dot_topn.git
  Cloning https://github.com/ing-bank/sparse_dot_topn.git to /tmp/pip-req-build-
fh9kcwox
  Running command git clone -q https://github.com/ing-bank/sparse_dot_topn.git
/tmp/pip-req-build-fh9kcwox
Requirement already satisfied: setuptools>=18.0 in
/opt/conda/lib/python3.8/site-packages (from sparse-dot-topn==0.2.9)
(49.6.0.post20210108)
Requirement already satisfied: cython>=0.29.15 in /opt/conda/lib/python3.8/site-
packages (from sparse-dot-topn==0.2.9) (0.29.23)
Requirement already satisfied: numpy>=1.16.6 in /opt/conda/lib/python3.8/site-
packages (from sparse-dot-topn==0.2.9) (1.20.2)
Requirement already satisfied: scipy>=1.2.3 in /opt/conda/lib/python3.8/site-
packages (from sparse-dot-topn==0.2.9) (1.6.2)
```

[427]:
```
def awesome_cossim_top(A, B, ntop, lower_bound=0):
    A = A.tocsr()
    B = B.tocsr()
    M, _ = A.shape
    _, N = B.shape

    idx_dtype = np.int32

    nnz_max = M*ntop

    indptr = np.zeros(M+1, dtype=idx_dtype)
    indices = np.zeros(nnz_max, dtype=idx_dtype)
    data = np.zeros(nnz_max, dtype=A.dtype)

    ct.sparse_dot_topn(
        M, N, np.asarray(A.indptr, dtype=idx_dtype),
        np.asarray(A.indices, dtype=idx_dtype),
        A.data,
        np.asarray(B.indptr, dtype=idx_dtype),
        np.asarray(B.indices, dtype=idx_dtype),
        B.data,
        ntop,
        lower_bound,
```

```
        indptr, indices, data)

    return csr_matrix((data,indices,indptr),shape=(M,N))
```

[428]:
```
import time
import sparse_dot_topn.sparse_dot_topn as ct
t1 = time.time()
matches = awesome_cossim_top(tf_idf_matrix, tf_idf_matrix.transpose(), 10, 0.8)
t = time.time()-t1
```

[434]:
```
def get_matches_df(sparse_matrix, name_vector, top=100):
    non_zeros = sparse_matrix.nonzero()
    sparserows = non_zeros[0]
    sparsecols = non_zeros[1]
    if top:
        nr_matches = top
    else:
        nr_matches = sparsecols.size

    left_side = np.empty([nr_matches], dtype=object)
    right_side = np.empty([nr_matches], dtype=object)
    similairity = np.zeros(nr_matches)

    for index in range(0, nr_matches):
        left_side[index] = name_vector[sparserows[index]]
        right_side[index] = name_vector[sparsecols[index]]
        similairity[index] = sparse_matrix.data[index]

    return pd.DataFrame({'left_side': left_side,
                         'right_side': right_side,
                          'similairity': similairity})
matches_df = get_matches_df(matches, text, top=100)
matches_df = matches_df[matches_df['similairity'] < 0.99999] # Remove all exact␣
↪matches
matches_df.sort_values(['similairity'], ascending=False).head(2)
```

[434]:
```
left_side  \
86  A Modesto, California resident was convicted today after a 10-day jury trial
on 14 child exploitation offenses for his role in a child exploitation
enterprise, announced Acting Assistant Attorney General Kenneth A. Blanco of the
Justice Department's Criminal Division and Acting U.S. Attorney Daniel L.
Lemisch of the Eastern District of Michigan.   Justin Fuller, 37, a bridge
maintenance supervisor for the California Department of Transportation, was
found guilty of one count of engaging in a child exploitation enterprise; one
count of conspiracy to produce child pornography; five counts of production of
child pornography; one count of conspiracy to receive child pornography; one
count of conspiracy to access with intent to view child pornography; and five
```

counts of enticement of a minor to engage in illegal sexual activity. According to trial evidence, between Nov. 16, 2013, and March 10, 2016, Fuller and five co-conspirators located in different states worked together to lure juvenile girls to a video chat website in order to get them to engage in sexually explicit conduct. The group members predominantly targeted prepubescent girls and would, unbeknownst to the girls, record the lured young girls performing the sexually explicit conduct. The group was active for approximately two years and communicated with each other through "base" chatrooms that were password-protected. In the base chat rooms, Fuller and co-conspirators strategized how to convince minor females to produce child pornography, including pretending to be teenage boys or girls to help convince the minor females to engage in sexual activity. The other five co-conspirators each pleaded guilty prior to trial to one count of engaging in a child exploitation enterprise. On June 21, 2016, Virgil Napier, 54, of Waterford, Michigan, pleaded guilty. On July 11, 2016, John Garrison, 52, of Glenarm, Illinois, pleaded guilty. On Feb. 24, 2017, Thomas Dougherty, 54, of Vallejo, California, pleaded guilty. On Sept. 23, 2016, Dantly Nicart, 39, a citizen of the Philippines residing in Las Vegas, pleaded guilty, and was sentenced to 20 years imprisonment followed by five years of supervised release and $150,000 in restitution on March 2, 2017. On June 21, 2016, Brandon Henneberg, 31, of Diller, Nebraska, pleaded guilty in the District of Nebraska, and on Sept. 14, 2016, he was sentenced to 35 years imprisonment, followed by a lifetime term of supervised release and $60,000 in restitution. Trial Attorney Austin M. Berry of the Criminal Division's Child Exploitation and Obscenity Section (CEOS) and Assistant U.S. Attorney April N. Russo of the Eastern District of Michigan are prosecuting the case. The FBI's Detroit Field Office and Southeast Michigan Trafficking and Exploitation Crimes (SEMTEC) task force investigated the case with assistance from CEOS's High Technology Investigative Unit. This case was brought as part of Project Safe Childhood, a nationwide initiative to combat the growing epidemic of child sexual exploitation and abuse, launched in May 2006 by the Department of Justice. Led by U.S. Attorneys' Offices and CEOS, Project Safe Childhood marshals federal, state and local resources to better locate, apprehend and prosecute individuals who exploit children via the Internet, as well as to identify and rescue victims. For more information about Project Safe Childhood, please visit http://www.justice.gov/psc.\n
81
A citizen of the United Kingdom was sentenced today to 85 years in prison for his part in a child pornography trafficking conspiracy, announced Assistant Attorney General Leslie R. Caldwell of the Justice Department's Criminal Division and U.S. Attorney Josh Minkler of the Southern District of Indiana. Domminich Shaw, 35, was sentenced by U.S. District Judge William T. Lawrence of the Southern District of Indiana, who ordered that he also serve a lifetime term of supervised release. Shaw pleaded guilty on Oct. 22, 2015, to 26 counts, including conspiracy to advertise child pornography and conspiracy to receive and distribute child pornography. He was indicted by a grand jury in Indianapolis on Feb. 23, 2011, and was extradited from the United Kingdom on Dec. 20, 2014. In 2005, Shaw was convicted in the U.K. of "indecent assault" on

four different females under the age of 13. According to plea documents, Shaw created and administered a website that contained child pornography involving infants and toddlers. This website allowed Shaw and other co-conspirators to distribute and advertise to each other images and videos, and send one another related messages, so that the child pornography would be shared with other members. Shaw participated on the website under aliases, including "Nepi" and several variations of that word. The word "nepi" is associated with nepiophilia, the sexual attraction to babies, toddlers and very young children. This case is part of Operation Bulldog, in which nine individuals have been convicted in the Southern District of Indiana. The FBI's Indianapolis Division and London's Metropolitan Police Service investigated the case. Trial Attorney Austin M. Berry of the Criminal Division's Child Exploitation and Obscenity Section (CEOS) and Senior Litigation Counsel Steven DeBrota of the Southern District of Indiana prosecuted the case. The Criminal Division's Office of International Affairs provided assistance in this matter. This case was brought as part of Project Safe Childhood, a nationwide initiative to combat the growing epidemic of child sexual exploitation and abuse launched in May 2006 by the Department of Justice. Led by U.S. Attorneys' Offices and CEOS, Project Safe Childhood marshals federal, state and local resources to better locate, apprehend and prosecute individuals who exploit children via the Internet, as well as to identify and rescue victims. For more information about Project Safe Childhood, please visit www.justice.gov/psc.

right_side \

86 WASHINGTON - A Danish man was sentenced today in the Western District of Missouri to 30 years in prison for producing and transporting child pornography and for extortion against an 11-year-old Missouri girl, Assistant Attorney General Lanny A. Breuer of the Justice Department's Criminal Division and U.S. Attorney Beth Phillips of the Western District of Missouri announced. Kai Lundstroem Pedersen, 61, a citizen of Denmark, was sentenced by U.S. District Judge Greg Kays. Pedersen pleaded guilty to the federal indictment on Sept. 6, 2011. According to court documents, in July 2010, Pedersen engaged in a video web chat from his home in Denmark with an 11-year-old girl in Buchanan County, Mo., identified as Jane Doe #1. Pedersen used a fake Facebook account to pose as a juvenile-aged male. Pedersen admitted that he instructed the girl to engage in sexually explicit conduct that he recorded and saved as a digital video on his computer. This video, as well as screen capture images from the video, were later edited and distributed to others, including family and friends of the victim. Pedersen distributed the images and video over the Internet via file-sharing software. Pedersen contacted Jane Doe #1 using various aliases through email and chat programs from July to September 2010 in an effort to convince her to engage again in sexually explicit conduct via video web chat. Pedersen threatened to disseminate sexually explicit images of her over the Internet if she did not comply with his demands. For example, on Aug. 15, 2010, Pedersen used nine different aliases on Facebook to contact Jane Doe #1, relaying rape and murder fantasies, asserting that various individuals had watched her video and describing the various sexual acts that these individuals

wanted to perform on her. According to court documents, in August 2010, Pedersen initiated contact with another minor female in rural Missouri, whom he believed to be a close friend of Jane Doe #1. Pedersen contacted this victim, identified as Jane Doe #2, in an effort to exert pressure to have either Jane Doe #1 or Jane Doe #2 perform sexually explicit conduct for him via video web chat. On Aug. 13, 2010, Jane Doe #1=s mother contacted law enforcement authorities. The mother told an officer that she learned of the contact with Pedersen after receiving Facebook messages that contained nude images of her daughter. A law enforcement officer, posing as a minor victim, communicated online with Pedersen and learned that he was traveling for vacation on Aug. 20, 2010. When Pedersen logged into his Facebook account on Aug. 25, 2010, investigators were able to trace his Internet protocol address to a residence in Stonybrook, N.Y., where he was arrested on Sept. 3, 2010. This case was prosecuted by Assistant U.S. Attorney Patrick D. Daly of the Western District of Missouri and Trial Attorney Keith Becker of the Child Exploitation and Obscenity Section (CEOS) in the Justice Department's Criminal Division. It was investigated by the Buchanan County, Mo., Sheriff=s Department; the Western Missouri Cyber Crimes Task Force; the U.S. Immigration and Customs Enforcement (ICE) Office of Homeland Security Investigations (HSI); and CEOS. This case was brought as part of Project Safe Childhood, a nationwide initiative to combat the growing epidemic of child sexual exploitation and abuse launched in May 2006 by the Department of Justice. Led by U.S. Attorneys= Offices and the Criminal Division=s CEOS, Project Safe Childhood marshals federal, state and local resources to better locate, apprehend and prosecute individuals who exploit children, as well as to identify and rescue victims. For more information about Project Safe Childhood, please visit www.projectsafechildhood.gov.

81

A Los Angeles resident was sentenced today to 10 years in prison for two child exploitation offenses, including engaging in illicit sexual conduct in foreign places and traveling in foreign commerce for the purpose of engaging in illicit sexual conduct, announced Acting Assistant Attorney General John P. Cronan of the Justice Department's Criminal Division and Special Agent in Charge Joseph Macias of the U.S. Immigration and Customs Enforcement's (ICE) Homeland Security Investigations (HSI) Los Angeles. Paul Alan Shapiro, 71, a retired auto dealership employee, pleaded guilty one day before he was set to go on trial on July 24, 2017. Under the terms of the plea agreement, Shapiro will serve 10 years in federal prison, 20 years of supervised release following his prison sentence, and will pay $20,000 total to two victims, both of whom are citizens of the Kingdom of Thailand. U.S. District Court Judge Dolly M. Gee of the Central District of California presided over today's sentencing. According to plea documents, Shapiro traveled from Los Angeles to Thailand on numerous occasions over the past 20 years, and engaged in sexual acts with male boys under the age of 16 on multiple occasions. On at least two occasions in September 2012, Shapiro paid minors as young as 13 years old small amounts of local currency in order to engage in various sex acts with them. According to other documents filed in the case, Shapiro photographed these encounters of himself engaging in sexually explicit conduct with the boys. HSI conducted the

```
investigation.  Trial Attorneys Austin M. Berry and Ralph Paradiso of the
Criminal Division's Child Exploitation and Obscenity Section (CEOS) prosecuted
the case. This case was brought as part of Project Safe Childhood, a nationwide
initiative to combat the growing epidemic of child sexual exploitation and
abuse, launched in May 2006 by the Department of Justice.  Led by U.S.
Attorneys' Offices and CEOS, Project Safe Childhood marshals federal, state and
local resources to better locate, apprehend and prosecute individuals who
exploit children via the Internet, as well as to identify and rescue victims.
For more information about Project Safe Childhood, please visit
http://www.justice.gov/psc.
```

```
    similairity
86     0.941365
81     0.936934
```

[360]:
```
# For the first top match, it seems to me that they are the crime under Project␣
 ↪Safe Childhood (both related to
# child exploitation, paritculary producing or delivering child pornography);␣
 ↪however by reading
# the first sentence it is obvious that the press release on the left side␣
 ↪shows the crime that is conviction, but
# the right side press release indicates sentencing. For the second top match,␣
 ↪both of the press releases outlines
# setencing. Although the crims in the second match would be categorized in the␣
 ↪Project Safe Childhood section
# the specifics of the crime is a bit different as the first match is about␣
 ↪advertising and distributing child
# pornography but the other press (on the right) is about engaging in illegal␣
 ↪sexual activities with teenagers.
```

[361]:
```
#The code in question 2.5 is adapted from https://berguca.github.io/2017/10/14/
 ↪super-fast-string-matching.html
```