Assignment 3

Daniel Mehta

Exercise 1: Beginner's Guide to Named Entity Recognition (NER) in NLTK Library - MLK - Machine Learning Knowledge

In [1]:
```
!pip install nltk==3.8.1
```

```
ERROR: Operation cancelled by user
```

In [2]:
```
import nltk

nltk_data_path = '/usr/local/nltk_data'
nltk.data.path.append(nltk_data_path)

nltk.download('punkt', download_dir=nltk_data_path)
nltk.download('punkt_tab', download_dir=nltk_data_path)
nltk.download('averaged_perceptron_tagger', download_dir=nltk_data_path)
nltk.download('maxent_ne_chunker', download_dir=nltk_data_path)
nltk.download('words', download_dir=nltk_data_path)
nltk.download('treebank', download_dir=nltk_data_path)
```

```
[nltk_data] Downloading package punkt to /usr/local/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package punkt_tab to /usr/local/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /usr/local/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package maxent_ne_chunker to
[nltk_data]     /usr/local/nltk_data...
[nltk_data]   Package maxent_ne_chunker is already up-to-date!
[nltk_data] Downloading package words to /usr/local/nltk_data...
[nltk_data]   Package words is already up-to-date!
[nltk_data] Downloading package treebank to /usr/local/nltk_data...
[nltk_data]   Package treebank is already up-to-date!
```

Out[2]: True

In [3]:
```
from nltk import word_tokenize,pos_tag

text = "NASA awarded Elon Musk’s SpaceX a $2.9 billion contract to build the lunar land
tokens = word_tokenize(text)
tag=pos_tag(tokens)
print(tag)

ne_tree = nltk.ne_chunk(tag)
print(ne_tree)
```

```
[('NASA', 'NNP'), ('awarded', 'VBD'), ('Elon', 'NNP'), ('Musk', 'NNP'), ('’', 'NNP'),
('s', 'VBD'), ('SpaceX', 'NNP'), ('a', 'DT'), ('$', '$'), ('2.9', 'CD'), ('billion', 'C
D'), ('contract', 'NN'), ('to', 'TO'), ('build', 'VB'), ('the', 'DT'), ('lunar', 'NN'),
('lander', 'NN'), ('.', '.')]
(S
  (ORGANIZATION NASA/NNP)
  awarded/VBD
  (PERSON Elon/NNP Musk/NNP)
  ’/NNP
  s/VBD
  (ORGANIZATION SpaceX/NNP)
  a/DT
  $/$
  2.9/CD
  billion/CD
  contract/NN
  to/TO
  build/VB
  the/DT
  lunar/NN
  lander/NN
  ./.)
```

In [4]:

```
sent = nltk.corpus.treebank.tagged_sents()
print(nltk.ne_chunk(sent[0]))
```

```
(S
  (PERSON Pierre/NNP)
  (ORGANIZATION Vinken/NNP)
  ,/,
  61/CD
  years/NNS
  old/JJ
  ,/,
  will/MD
  join/VB
  the/DT
  board/NN
  as/IN
  a/DT
  nonexecutive/JJ
  director/NN
  Nov./NNP
  29/CD
  ./.)
```

In [5]:

```
import spacy
nlp = spacy.load("en_core_web_sm")

doc = nlp("NASA awarded Elon Musk’s SpaceX a $2.9 billion contract to build the lunar l
for token in doc:
    print(token.text, token.ent_iob_, token.ent_type_)
```

```
NASA B ORG
awarded O
Elon B PERSON
Musk I PERSON
's I PERSON
SpaceX I PERSON
a O
$ B MONEY
2.9 I MONEY
billion I MONEY
contract O
to O
build O
the O
lunar O
lander O
. O
```

Exercise 2:

a) Find a new text dataset

In [6]:
```python
import kagglehub
import os

# Download latest version
path = kagglehub.dataset_download("rmisra/news-category-dataset")
json_file = os.path.join(path, "News_Category_Dataset_v3.json")


print("Path to dataset files:", path)
```

```
Path to dataset files: /kaggle/input/news-category-dataset
```

b) Convert it into csv format

In [7]:
```python
import json
import pandas as pd
```

In [8]:
```python
with open(json_file, "r") as f:
    data = [json.loads(line) for line in f]

csv_file = "News_Category_Dataset.csv"

df = pd.read_json(json_file, lines=True)
df.to_csv(csv_file, index=False)
```

In [9]:
```python
df.head()
```

Out[9]:

	link	headline	category	short_description	authors	dat
0	https://www.huffpost.com/entry/covid-boosters-...	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	Carla K. Johnson, AP	2022 09 2
1	https://www.huffpost.com/entry/american-airlin...	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	Mary Papenfuss	2022 09 2
2	https://www.huffpost.com/entry/funniest-tweets...	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	Elyse Wanshel	2022 09 2
3	https://www.huffpost.com/entry/funniest-parent...	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	Caroline Bologna	2022 09 2
4	https://www.huffpost.com/entry/amy-cooper-lose...	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	Nina Golgowski	2022 09 2

c) Redo the same exercise

In [10]:

```
for i in range(5):
    text = df['headline'].iloc[i]

    tokens = word_tokenize(text)
    tag=pos_tag(tokens)
    print(tag)

    ne_tree = nltk.ne_chunk(tag)
    print(ne_tree)
```

```
[('Over', 'IN'), ('4', 'CD'), ('Million', 'NNP'), ('Americans', 'NNPS'), ('Roll', 'NNP'),
('Up', 'NNP'), ('Sleeves', 'NNP'), ('For', 'IN'), ('Omicron-Targeted', 'NNP'), ('COVID',
'NNP'), ('Boosters', 'NNP')]
(S
  Over/IN
  4/CD
  Million/NNP
  Americans/NNPS
  (PERSON Roll/NNP Up/NNP Sleeves/NNP)
  For/IN
  Omicron-Targeted/NNP
  COVID/NNP
  Boosters/NNP)
[('American', 'NNP'), ('Airlines', 'NNPS'), ('Flyer', 'NNP'), ('Charged', 'NNP'), (',',
','), ('Banned', 'NNP'), ('For', 'IN'), ('Life', 'NNP'), ('After', 'IN'), ('Punching', 'V
BG'), ('Flight', 'NNP'), ('Attendant', 'NNP'), ('On', 'IN'), ('Video', 'NNP')]
(S
  (GPE American/NNP)
  (ORGANIZATION Airlines/NNPS Flyer/NNP)
  Charged/NNP
  ,/,
  (GPE Banned/NNP)
  (ORGANIZATION For/IN Life/NNP)
  After/IN
  Punching/VBG
  (PERSON Flight/NNP Attendant/NNP)
  On/IN
  Video/NNP)
[('23', 'CD'), ('Of', 'IN'), ('The', 'DT'), ('Funniest', 'NNP'), ('Tweets', 'NNPS'), ('Ab
out', 'NNP'), ('Cats', 'NNP'), ('And', 'CC'), ('Dogs', 'NNP'), ('This', 'DT'), ('Week',
'NNP'), ('(', '('), ('Sept.', 'NNP'), ('17-23', 'CD'), (')', ')')]
(S
  23/CD
  Of/IN
  The/DT
  (ORGANIZATION
    Funniest/NNP
    Tweets/NNPS
    About/NNP
    Cats/NNP
    And/CC
    Dogs/NNP
    This/DT
    Week/NNP)
  (/(
  Sept./NNP
  17-23/CD
  )/))
[('The', 'DT'), ('Funniest', 'NNP'), ('Tweets', 'NNPS'), ('From', 'NNP'), ('Parents', 'NN
P'), ('This', 'DT'), ('Week', 'NNP'), ('(', '('), ('Sept.', 'NNP'), ('17-23', 'CD'),
(')', ')')]
(S
  The/DT
  (ORGANIZATION Funniest/NNP Tweets/NNPS)
  From/NNP
  Parents/NNP
  This/DT
  Week/NNP
  (/(
  Sept./NNP
  17-23/CD
  )/))
[('Woman', 'NNP'), ('Who', 'WP'), ('Called', 'VBD'), ('Cops', 'NNP'), ('On', 'IN'), ('Bla
ck', 'NNP'), ('Bird-Watcher', 'NNP'), ('Loses', 'NNP'), ('Lawsuit', 'NNP'), ('Against',
```

```
'NNP'), ('Ex-Employer', 'NNP')]
(S
  Woman/NNP
  Who/WP
  Called/VBD
  (PERSON Cops/NNP)
  On/IN
  (PERSON Black/NNP)
  Bird-Watcher/NNP
  (PERSON Loses/NNP Lawsuit/NNP Against/NNP)
  Ex-Employer/NNP)
```

In [11]:

```python
nlp = spacy.load("en_core_web_sm")
for i in range(5):
    text = df['headline'].iloc[i]
    doc = nlp(text)

    print(f"\nHeadline {i+1}: {text}")
    for token in doc:
        print(token.text, token.ent_iob_, token.ent_type_)
```

Headline 1: Over 4 Million Americans Roll Up Sleeves For Omicron-Targeted COVID Boosters
Over O
4 O
Million O
Americans O
Roll O
Up O
Sleeves O
For O
Omicron O
- O
Targeted O
COVID O
Boosters O

Headline 2: American Airlines Flyer Charged, Banned For Life After Punching Flight Attend
ant On Video
American B ORG
Airlines I ORG
Flyer B PERSON
Charged I PERSON
, O
Banned B ORG
For I ORG
Life I ORG
After I ORG
Punching I ORG
Flight I ORG
Attendant I ORG
On I ORG
Video I ORG

Headline 3: 23 Of The Funniest Tweets About Cats And Dogs This Week (Sept. 17-23)
23 B CARDINAL
Of O
The O
Funniest O
Tweets O
About O
Cats O
And O
Dogs O
This O
Week O
(O
Sept. B DATE
17 I DATE
- I DATE
23 I DATE
) O

Headline 4: The Funniest Tweets From Parents This Week (Sept. 17-23)
The O
Funniest O
Tweets O
From O
Parents O
This B DATE
Week I DATE
(O
Sept. B DATE
17 I DATE
- I DATE
23 I DATE

```
) O

Headline 5: Woman Who Called Cops On Black Bird-Watcher Loses Lawsuit Against Ex-Employer
Woman O
Who O
Called O
Cops O
On O
Black O
Bird O
- O
Watcher O
Loses O
Lawsuit O
Against O
Ex O
- O
Employer O
```

Exercise 3: TF-IDF with Scikit-Learn — Introduction to Cultural Analytics & Python

In [12]:
```
!pip install altair
```

```
Requirement already satisfied: altair in /usr/local/lib/python3.11/dist-packages (5.5.0)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from al
tair) (3.1.6)
Requirement already satisfied: jsonschema>=3.0 in /usr/local/lib/python3.11/dist-packages
(from altair) (4.23.0)
Requirement already satisfied: narwhals>=1.14.2 in /usr/local/lib/python3.11/dist-package
s (from altair) (1.40.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from
altair) (24.2)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dis
t-packages (from altair) (4.13.2)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.11/dist-packages
(from jsonschema>=3.0->altair) (25.3.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/pyt
hon3.11/dist-packages (from jsonschema>=3.0->altair) (2025.4.1)
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.11/dist-pack
ages (from jsonschema>=3.0->altair) (0.36.2)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-packages
(from jsonschema>=3.0->altair) (0.25.1)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages
(from jinja2->altair) (3.0.2)
```

In [13]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
pd.set_option("display.max_rows", 600)
from pathlib import Path
import glob
```

In [14]:
```
directory_path = "US_Inaugural_Addresses"
```

In [15]:
```
text_files = glob.glob(f"{directory_path}/*.txt")
text_files
```

Out[15]:
```
['US_Inaugural_Addresses/31_taft_1909.txt',
 'US_Inaugural_Addresses/02_washington_1793.txt',
 'US_Inaugural_Addresses/09_monroe_1821.txt',
 'US_Inaugural_Addresses/49_reagan_1981.txt',
 'US_Inaugural_Addresses/15_polk_1845.txt',
 'US_Inaugural_Addresses/48_carter_1977.txt',
 'US_Inaugural_Addresses/33_wilson_1917.txt',
 'US_Inaugural_Addresses/45_johnson_1965.txt',
 'US_Inaugural_Addresses/41_truman_1949.txt',
 'US_Inaugural_Addresses/54_bush_george_w_2001.txt',
 'US_Inaugural_Addresses/52_clinton_1993.txt',
 'US_Inaugural_Addresses/18_buchanan_1857.txt',
 'US_Inaugural_Addresses/43_eisenhower_1957.txt',
 'US_Inaugural_Addresses/08_monroe_1817.txt',
 'US_Inaugural_Addresses/39_roosevelt_franklin_1941.txt',
 'US_Inaugural_Addresses/56_obama_2009.txt',
 'US_Inaugural_Addresses/35_coolidge_1925.txt',
 'US_Inaugural_Addresses/29_mckinley_1901.txt',
 'US_Inaugural_Addresses/21_grant_1869.txt',
 'US_Inaugural_Addresses/30_roosevelt_theodore_1905.txt',
 'US_Inaugural_Addresses/28_mckinley_1897.txt',
 'US_Inaugural_Addresses/34_harding_1921.txt',
 'US_Inaugural_Addresses/01_washington_1789.txt',
 'US_Inaugural_Addresses/53_clinton_1997.txt',
 'US_Inaugural_Addresses/27_cleveland_1893.txt',
 'US_Inaugural_Addresses/46_nixon_1969.txt',
 'US_Inaugural_Addresses/57_obama_2013.txt',
 'US_Inaugural_Addresses/19_lincoln_1861.txt',
 'US_Inaugural_Addresses/16_taylor_1849.txt',
 'US_Inaugural_Addresses/05_jefferson_1805.txt',
 'US_Inaugural_Addresses/36_hoover_1929.txt',
 'US_Inaugural_Addresses/50_reagan_1985.txt',
 'US_Inaugural_Addresses/10_adams_john_quincy_1825.txt',
 'US_Inaugural_Addresses/20_lincoln_1865.txt',
 'US_Inaugural_Addresses/03_adams_john_1797.txt',
 'US_Inaugural_Addresses/04_jefferson_1801.txt',
 'US_Inaugural_Addresses/58_trump_2017.txt',
 'US_Inaugural_Addresses/13_van_buren_1837.txt',
 'US_Inaugural_Addresses/51_bush_george_h_w_1989.txt',
 'US_Inaugural_Addresses/38_roosevelt_franklin_1937.txt',
 'US_Inaugural_Addresses/23_hayes_1877.txt',
 'US_Inaugural_Addresses/32_wilson_1913.txt',
 'US_Inaugural_Addresses/17_pierce_1853.txt',
 'US_Inaugural_Addresses/24_garfield_1881.txt',
 'US_Inaugural_Addresses/11_jackson_1829.txt',
 'US_Inaugural_Addresses/44_kennedy_1961.txt',
 'US_Inaugural_Addresses/55_bush_george_w_2005.txt',
 'US_Inaugural_Addresses/42_eisenhower_1953.txt',
 'US_Inaugural_Addresses/07_madison_1813.txt',
 'US_Inaugural_Addresses/40_roosevelt_franklin_1945.txt',
 'US_Inaugural_Addresses/47_nixon_1973.txt',
 'US_Inaugural_Addresses/12_jackson_1833.txt',
 'US_Inaugural_Addresses/06_madison_1809.txt',
 'US_Inaugural_Addresses/25_cleveland_1885.txt',
 'US_Inaugural_Addresses/22_grant_1873.txt',
 'US_Inaugural_Addresses/14_harrison_1841.txt',
 'US_Inaugural_Addresses/26_harrison_1889.txt',
 'US_Inaugural_Addresses/37_roosevelt_franklin_1933.txt']
```

In [16]:
```
text_titles = [Path(text).stem for text in text_files]
text_titles
```

Out[16]:
```
['31_taft_1909',
 '02_washington_1793',
 '09_monroe_1821',
 '49_reagan_1981',
 '15_polk_1845',
 '48_carter_1977',
 '33_wilson_1917',
 '45_johnson_1965',
 '41_truman_1949',
 '54_bush_george_w_2001',
 '52_clinton_1993',
 '18_buchanan_1857',
 '43_eisenhower_1957',
 '08_monroe_1817',
 '39_roosevelt_franklin_1941',
 '56_obama_2009',
 '35_coolidge_1925',
 '29_mckinley_1901',
 '21_grant_1869',
 '30_roosevelt_theodore_1905',
 '28_mckinley_1897',
 '34_harding_1921',
 '01_washington_1789',
 '53_clinton_1997',
 '27_cleveland_1893',
 '46_nixon_1969',
 '57_obama_2013',
 '19_lincoln_1861',
 '16_taylor_1849',
 '05_jefferson_1805',
 '36_hoover_1929',
 '50_reagan_1985',
 '10_adams_john_quincy_1825',
 '20_lincoln_1865',
 '03_adams_john_1797',
 '04_jefferson_1801',
 '58_trump_2017',
 '13_van_buren_1837',
 '51_bush_george_h_w_1989',
 '38_roosevelt_franklin_1937',
 '23_hayes_1877',
 '32_wilson_1913',
 '17_pierce_1853',
 '24_garfield_1881',
 '11_jackson_1829',
 '44_kennedy_1961',
 '55_bush_george_w_2005',
 '42_eisenhower_1953',
 '07_madison_1813',
 '40_roosevelt_franklin_1945',
 '47_nixon_1973',
 '12_jackson_1833',
 '06_madison_1809',
 '25_cleveland_1885',
 '22_grant_1873',
 '14_harrison_1841',
 '26_harrison_1889',
 '37_roosevelt_franklin_1933']
```

In [17]:
```
tfidf_vectorizer = TfidfVectorizer(input='filename', stop_words='english')
tfidf_vector = tfidf_vectorizer.fit_transform(text_files)
tfidf_df = pd.DataFrame(tfidf_vector.toarray(), index=text_titles, columns=tfidf_vector
tfidf_df.loc['00_Document Frequency'] = (tfidf_df > 0).sum()
```

```
tfidf_slice = tfidf_df[['government', 'borders', 'people', 'obama', 'war', 'honor','for
tfidf_slice.sort_index().round(decimals=2)
```

Out[17]:

	government	borders	people	obama	war	honor	foreign	men	w
00_Document Frequency	53.00	5.00	56.00	3.00	45.00	32.00	32.00	47.00	
01_washington_1789	0.11	0.00	0.05	0.00	0.00	0.00	0.00	0.02	
02_washington_1793	0.06	0.00	0.05	0.00	0.00	0.08	0.00	0.00	
03_adams_john_1797	0.16	0.00	0.19	0.00	0.01	0.10	0.12	0.04	
04_jefferson_1801	0.16	0.00	0.01	0.00	0.01	0.04	0.00	0.04	
05_jefferson_1805	0.03	0.00	0.00	0.00	0.04	0.00	0.06	0.01	
06_madison_1809	0.00	0.00	0.02	0.00	0.02	0.05	0.05	0.00	
07_madison_1813	0.04	0.00	0.04	0.00	0.25	0.02	0.02	0.00	
08_monroe_1817	0.17	0.00	0.11	0.00	0.09	0.01	0.10	0.04	
09_monroe_1821	0.08	0.00	0.06	0.00	0.11	0.02	0.04	0.01	
10_adams_john_quincy_1825	0.15	0.00	0.06	0.00	0.05	0.01	0.08	0.03	
11_jackson_1829	0.10	0.00	0.06	0.00	0.02	0.02	0.07	0.02	
12_jackson_1833	0.21	0.00	0.14	0.00	0.00	0.00	0.02	0.00	
13_van_buren_1837	0.12	0.00	0.14	0.00	0.02	0.02	0.06	0.02	
14_harrison_1841	0.14	0.00	0.14	0.00	0.01	0.02	0.03	0.03	
15_polk_1845	0.26	0.00	0.08	0.00	0.03	0.01	0.09	0.02	
16_taylor_1849	0.12	0.00	0.05	0.00	0.00	0.02	0.05	0.00	
17_pierce_1853	0.08	0.00	0.05	0.00	0.00	0.02	0.04	0.01	
18_buchanan_1857	0.12	0.00	0.11	0.00	0.08	0.01	0.04	0.03	
19_lincoln_1861	0.12	0.00	0.13	0.00	0.02	0.00	0.02	0.00	
20_lincoln_1865	0.02	0.00	0.00	0.00	0.27	0.00	0.00	0.04	
21_grant_1869	0.05	0.00	0.03	0.00	0.02	0.05	0.05	0.02	
22_grant_1873	0.06	0.00	0.10	0.00	0.05	0.02	0.00	0.00	
23_hayes_1877	0.17	0.00	0.08	0.00	0.00	0.00	0.04	0.02	
24_garfield_1881	0.19	0.00	0.16	0.00	0.05	0.00	0.00	0.01	
25_cleveland_1885	0.21	0.00	0.21	0.00	0.00	0.00	0.05	0.01	
26_harrison_1889	0.06	0.00	0.17	0.00	0.02	0.03	0.01	0.04	
27_cleveland_1893	0.15	0.00	0.22	0.00	0.00	0.00	0.00	0.04	
28_mckinley_1897	0.16	0.00	0.16	0.00	0.05	0.03	0.06	0.02	
29_mckinley_1901	0.15	0.00	0.12	0.00	0.08	0.04	0.01	0.01	
30_roosevelt_theodore_1905	0.05	0.00	0.10	0.00	0.00	0.00	0.00	0.08	
31_taft_1909	0.12	0.00	0.03	0.00	0.03	0.01	0.03	0.03	
32_wilson_1913	0.11	0.00	0.02	0.00	0.00	0.00	0.00	0.14	
33_wilson_1917	0.00	0.00	0.08	0.00	0.07	0.00	0.00	0.05	
34_harding_1921	0.08	0.00	0.05	0.00	0.12	0.00	0.00	0.01	

	government	borders	people	obama	war	honor	foreign	men	w
35_coolidge_1925	0.10	0.00	0.10	0.00	0.02	0.01	0.02	0.02	
36_hoover_1929	0.20	0.04	0.10	0.00	0.01	0.00	0.01	0.03	
37_roosevelt_franklin_1933	0.03	0.00	0.08	0.00	0.02	0.02	0.03	0.02	
38_roosevelt_franklin_1937	0.18	0.03	0.12	0.00	0.01	0.00	0.00	0.10	
39_roosevelt_franklin_1941	0.05	0.00	0.08	0.00	0.00	0.00	0.00	0.07	
40_roosevelt_franklin_1945	0.00	0.00	0.02	0.00	0.05	0.03	0.00	0.10	
41_truman_1949	0.03	0.00	0.10	0.00	0.02	0.01	0.01	0.06	
42_eisenhower_1953	0.01	0.00	0.10	0.00	0.04	0.03	0.00	0.07	
43_eisenhower_1957	0.00	0.00	0.10	0.00	0.01	0.05	0.00	0.07	
44_kennedy_1961	0.00	0.00	0.01	0.00	0.06	0.00	0.00	0.01	
45_johnson_1965	0.01	0.00	0.11	0.00	0.01	0.00	0.02	0.03	
46_nixon_1969	0.05	0.00	0.13	0.00	0.03	0.03	0.00	0.01	
47_nixon_1973	0.10	0.00	0.06	0.00	0.03	0.01	0.00	0.00	
48_carter_1977	0.06	0.00	0.08	0.00	0.02	0.00	0.02	0.00	
49_reagan_1981	0.16	0.00	0.08	0.00	0.01	0.00	0.00	0.02	
50_reagan_1985	0.16	0.00	0.14	0.00	0.01	0.01	0.00	0.03	
51_bush_george_h_w_1989	0.05	0.00	0.06	0.00	0.03	0.00	0.01	0.04	
52_clinton_1993	0.05	0.00	0.13	0.00	0.03	0.00	0.02	0.01	
53_clinton_1997	0.09	0.00	0.09	0.00	0.01	0.00	0.00	0.00	
54_bush_george_w_2001	0.05	0.00	0.01	0.00	0.01	0.00	0.00	0.00	
55_bush_george_w_2005	0.03	0.06	0.05	0.00	0.00	0.04	0.00	0.02	
56_obama_2009	0.03	0.03	0.07	0.03	0.02	0.01	0.00	0.04	
57_obama_2013	0.04	0.00	0.11	0.04	0.04	0.00	0.00	0.04	
58_trump_2017	0.04	0.11	0.11	0.12	0.00	0.00	0.05	0.03	

In [18]:

```
tfidf_df = tfidf_df.drop('00_Document Frequency', errors='ignore')
tfidf_df.stack().reset_index()
```

Out[18]:

	level_0	level_1	0
0	31_taft_1909	000	0.028200
1	31_taft_1909	03	0.000000
2	31_taft_1909	04	0.000000
3	31_taft_1909	05	0.000000
4	31_taft_1909	100	0.018814
...	...	...	...
521937	37_roosevelt_franklin_1933	zachary	0.000000
521938	37_roosevelt_franklin_1933	zeal	0.000000
521939	37_roosevelt_franklin_1933	zealous	0.000000
521940	37_roosevelt_franklin_1933	zealously	0.000000
521941	37_roosevelt_franklin_1933	zone	0.000000

521942 rows × 3 columns

In [19]:

```
tfidf_df = tfidf_df.stack().reset_index()
tfidf_df = tfidf_df.rename(columns={0:'tfidf', 'level_0': 'document','level_1': 'term',
tfidf_df.sort_values(by=['document','tfidf'], ascending=[True,False]).groupby(['documen
```

Out[19]:

	document	term	tfidf
201685	01_washington_1789	government	0.113681
202086	01_washington_1789	immutable	0.103883
202153	01_washington_1789	impressions	0.103883
204315	01_washington_1789	providential	0.103883
203609	01_washington_1789	ought	0.103728
204329	01_washington_1789	public	0.103102
204095	01_washington_1789	present	0.097516
204367	01_washington_1789	qualifications	0.096372
203789	01_washington_1789	peculiarly	0.090546
198631	01_washington_1789	article	0.085786
9018	02_washington_1793	1793	0.229350
9643	02_washington_1793	arrive	0.229350
17576	02_washington_1793	upbraidings	0.229350
13250	02_washington_1793	incurring	0.208140
17700	02_washington_1793	violated	0.208140
17872	02_washington_1793	willingly	0.208140
13368	02_washington_1793	injunctions	0.193091
13705	02_washington_1793	knowingly	0.193091
15157	02_washington_1793	previous	0.193091
17910	02_washington_1793	witnesses	0.193091
311789	03_adams_john_1797	people	0.191180
309673	03_adams_john_1797	government	0.160937
311927	03_adams_john_1797	pleasing	0.147066
309430	03_adams_john_1797	foreign	0.116874
311324	03_adams_john_1797	nations	0.114480
314679	03_adams_john_1797	virtuous	0.110813
309984	03_adams_john_1797	houses	0.110300
310767	03_adams_john_1797	legislatures	0.110300
307727	03_adams_john_1797	constitution	0.104525
309951	03_adams_john_1797	honor	0.102265
318672	04_jefferson_1801	government	0.155691
321139	04_jefferson_1801	principle	0.130113
319782	04_jefferson_1801	let	0.117970
322031	04_jefferson_1801	safety	0.108427
319978	04_jefferson_1801	man	0.106841

	document	term	tfidf
323063	04_jefferson_1801	thousandth	0.104513
318947	04_jefferson_1801	honest	0.101696
318295	04_jefferson_1801	fellow	0.097240
321861	04_jefferson_1801	retire	0.094848
320542	04_jefferson_1801	opinion	0.092587
267322	05_jefferson_1805	public	0.180456
264232	05_jefferson_1805	false	0.135863
268593	05_jefferson_1805	state	0.121514
269811	05_jefferson_1805	whatsoever	0.116886
265847	05_jefferson_1805	limits	0.107085
262343	05_jefferson_1805	citizens	0.106592
267461	05_jefferson_1805	reason	0.104438
262456	05_jefferson_1805	comforts	0.101880
267106	05_jefferson_1805	press	0.101549
264113	05_jefferson_1805	expenses	0.096524
472132	06_madison_1809	improvements	0.152559
468876	06_madison_1809	belligerent	0.123161
474299	06_madison_1809	public	0.122235
473306	06_madison_1809	nations	0.104588
474681	06_madison_1809	rendered	0.101706
468735	06_madison_1809	authorities	0.089155
468746	06_madison_1809	avail	0.089155
470994	06_madison_1809	examples	0.089155
469851	06_madison_1809	councils	0.085894
473511	06_madison_1809	ones	0.085894
440717	07_madison_1813	war	0.254249
433059	07_madison_1813	british	0.222972
437020	07_madison_1813	massacre	0.119009
433162	07_madison_1813	captives	0.108003
433934	07_madison_1813	cruel	0.108003
438133	07_madison_1813	prisoners	0.108003
439054	07_madison_1813	savage	0.108003
434698	07_madison_1813	element	0.085005
434799	07_madison_1813	enemy	0.085005
435938	07_madison_1813	honorable	0.084762

	document	term	tfidf
124614	08_monroe_1817	states	0.184195
120694	08_monroe_1817	government	0.174125
120728	08_monroe_1817	great	0.160658
125456	08_monroe_1817	union	0.117193
122810	08_monroe_1817	people	0.112825
125459	08_monroe_1817	united	0.112076
119020	08_monroe_1817	dangers	0.108567
122354	08_monroe_1817	naval	0.104713
120451	08_monroe_1817	foreign	0.103460
123162	08_monroe_1817	principles	0.097766
21739	09_monroe_1821	great	0.173751
25625	09_monroe_1821	states	0.137384
24915	09_monroe_1821	revenue	0.115018
26763	09_monroe_1821	war	0.113785
23748	09_monroe_1821	parties	0.109318
26470	09_monroe_1821	united	0.108029
19502	09_monroe_1821	commerce	0.105001
21450	09_monroe_1821	force	0.102947
21496	09_monroe_1821	fortifications	0.098741
26032	09_monroe_1821	term	0.094808
296437	10_adams_john_quincy_1825	union	0.257335
291675	10_adams_john_quincy_1825	government	0.147726
291610	10_adams_john_quincy_1825	general	0.109221
294937	10_adams_john_quincy_1825	rights	0.096300
290486	10_adams_john_quincy_1825	dissensions	0.095289
294319	10_adams_john_quincy_1825	public	0.094573
289729	10_adams_john_quincy_1825	constitution	0.090300
293769	10_adams_john_quincy_1825	peace	0.088183
289886	10_adams_john_quincy_1825	country	0.086898
293806	10_adams_john_quincy_1825	performance	0.085565
402307	11_jackson_1829	public	0.160747
399599	11_jackson_1829	generally	0.122711
398337	11_jackson_1829	diffidence	0.112691
398096	11_jackson_1829	defending	0.105878
403238	11_jackson_1829	shall	0.104933

	document	term	tfidf
402873	11_jackson_1829	revenue	0.102776
404904	11_jackson_1829	worth	0.100312
399663	11_jackson_1829	government	0.099698
399272	11_jackson_1829	federal	0.093100
402000	11_jackson_1829	power	0.092071
467418	12_jackson_1833	union	0.212766
462656	12_jackson_1833	government	0.207559
466576	12_jackson_1833	states	0.141549
464772	12_jackson_1833	people	0.136557
465072	12_jackson_1833	preservation	0.128319
462591	12_jackson_1833	general	0.125422
462041	12_jackson_1833	exercise	0.119275
463194	12_jackson_1833	inculcate	0.116720
465243	12_jackson_1833	proportion	0.116720
464996	12_jackson_1833	powers	0.113757
337390	13_van_buren_1837	institutions	0.186889
338786	13_van_buren_1837	people	0.138465
336670	13_van_buren_1837	government	0.116561
340835	13_van_buren_1837	supposed	0.109949
334881	13_van_buren_1837	country	0.109276
333206	13_van_buren_1837	actual	0.096382
336107	13_van_buren_1837	experience	0.093444
333230	13_van_buren_1837	adherence	0.083833
334602	13_van_buren_1837	conduct	0.081635
338541	13_van_buren_1837	opinions	0.081597
500989	14_harrison_1841	power	0.204207
496706	14_harrison_1841	constitution	0.183336
498031	14_harrison_1841	executive	0.157153
500768	14_harrison_1841	people	0.141584
498652	14_harrison_1841	government	0.141142
501958	14_harrison_1841	roman	0.110538
502572	14_harrison_1841	states	0.108621
496317	14_harrison_1841	citizens	0.105857
496250	14_harrison_1841	character	0.102640
502567	14_harrison_1841	state	0.094976

	document	term	tfidf
44465	15_polk_1845	union	0.259054
39703	15_polk_1845	government	0.256967
43623	15_polk_1845	states	0.218122
44051	15_polk_1845	texas	0.199846
42913	15_polk_1845	revenue	0.146541
42043	15_polk_1845	powers	0.124655
42317	15_polk_1845	protection	0.107385
37757	15_polk_1845	constitution	0.106528
40473	15_polk_1845	interests	0.105054
39179	15_polk_1845	extended	0.090179
259254	16_taylor_1849	shall	0.266204
255679	16_taylor_1849	government	0.118031
254636	16_taylor_1849	duties	0.117893
257444	16_taylor_1849	object	0.104293
253662	16_taylor_1849	congress	0.103865
258340	16_taylor_1849	purity	0.101793
260638	16_taylor_1849	vested	0.101793
257078	16_taylor_1849	measures	0.101637
253890	16_taylor_1849	country	0.101169
252309	16_taylor_1849	affections	0.097017
381799	17_pierce_1853	hardly	0.114001
384002	17_pierce_1853	power	0.102456
383973	17_pierce_1853	position	0.086643
379720	17_pierce_1853	constitutional	0.086105
381085	17_pierce_1853	expect	0.084436
381665	17_pierce_1853	government	0.084048
378500	17_pierce_1853	apparent	0.080332
384583	17_pierce_1853	regarded	0.080332
385240	17_pierce_1853	shall	0.079615
382821	17_pierce_1853	like	0.079229
106616	18_buchanan_1857	states	0.208199
100750	18_buchanan_1857	constitution	0.188573
106271	18_buchanan_1857	shall	0.161784
105387	18_buchanan_1857	question	0.157007
107833	18_buchanan_1857	whilst	0.141119

	document	term	tfidf
107034	18_buchanan_1857	territory	0.140852
107458	18_buchanan_1857	union	0.126444
102696	18_buchanan_1857	government	0.119554
100679	18_buchanan_1857	congress	0.118357
104812	18_buchanan_1857	people	0.105501
244734	19_lincoln_1861	constitution	0.214478
251442	19_lincoln_1861	union	0.203738
244206	19_lincoln_1861	case	0.152422
250600	19_lincoln_1861	states	0.144861
248177	19_lincoln_1861	minority	0.131514
248796	19_lincoln_1861	people	0.130763
244368	19_lincoln_1861	clause	0.125738
246680	19_lincoln_1861	government	0.123837
250255	19_lincoln_1861	shall	0.123099
247731	19_lincoln_1861	law	0.122872
305732	20_lincoln_1865	war	0.267217
302502	20_lincoln_1865	offenses	0.234524
305880	20_lincoln_1865	woe	0.234524
300658	20_lincoln_1865	god	0.151269
302501	20_lincoln_1865	offense	0.141890
305842	20_lincoln_1865	wills	0.141890
297478	20_lincoln_1865	answered	0.131631
304382	20_lincoln_1865	slaves	0.123674
305436	20_lincoln_1865	union	0.114955
297412	20_lincoln_1865	altogether	0.111675
164574	21_grant_1869	dollar	0.270439
167778	21_grant_1869	paying	0.162263
164037	21_grant_1869	deal	0.152454
169509	21_grant_1869	specie	0.152454
164052	21_grant_1869	debt	0.135097
163900	21_grant_1869	country	0.127604
162304	21_grant_1869	advisable	0.116606
166747	21_grant_1869	laws	0.115834
167780	21_grant_1869	payments	0.108175
167776	21_grant_1869	pay	0.098658

	document	term	tfidf
492248	22_grant_1873	proposition	0.187222
488549	22_grant_1873	domingo	0.177516
493037	22_grant_1873	santo	0.177516
494172	22_grant_1873	transit	0.177516
493991	22_grant_1873	territory	0.121158
489139	22_grant_1873	extermination	0.118344
493596	22_grant_1873	steam	0.118344
493948	22_grant_1873	telegraph	0.118344
487864	22_grant_1873	country	0.117529
489132	22_grant_1873	extension	0.116618
361878	23_hayes_1877	country	0.186357
363667	23_hayes_1877	government	0.167722
360872	23_hayes_1877	behalf	0.128316
366311	23_hayes_1877	public	0.123944
365948	23_hayes_1877	political	0.121034
367587	23_hayes_1877	states	0.113587
365717	23_hayes_1877	party	0.112549
362465	23_hayes_1877	dispute	0.112503
365710	23_hayes_1877	parties	0.109554
366565	23_hayes_1877	reform	0.104365
390664	24_garfield_1881	government	0.186855
392780	24_garfield_1881	people	0.162132
388718	24_garfield_1881	constitution	0.158292
394584	24_garfield_1881	states	0.135047
395426	24_garfield_1881	union	0.132321
394767	24_garfield_1881	suffrage	0.119992
392357	24_garfield_1881	negro	0.118782
387745	24_garfield_1881	authority	0.117232
388647	24_garfield_1881	congress	0.112598
391715	24_garfield_1881	law	0.103639
482770	25_cleveland_1885	people	0.210468
480654	25_cleveland_1885	government	0.209164
482698	25_cleveland_1885	partisan	0.169436
483298	25_cleveland_1885	public	0.163662
484229	25_cleveland_1885	shall	0.129498

	document	term	tfidf
478708	25_cleveland_1885	constitution	0.127856
481424	25_cleveland_1885	interests	0.118207
480151	25_cleveland_1885	extravagance	0.111416
478317	25_cleveland_1885	citizen	0.102825
484662	25_cleveland_1885	strife	0.101661
509767	26_harrison_1889	people	0.172358
508709	26_harrison_1889	laws	0.154418
511571	26_harrison_1889	states	0.138614
504790	26_harrison_1889	ballot	0.137159
510295	26_harrison_1889	public	0.128566
509105	26_harrison_1889	methods	0.119162
511226	26_harrison_1889	shall	0.118483
507513	26_harrison_1889	friendly	0.104267
506954	26_harrison_1889	european	0.103349
505705	26_harrison_1889	constitution	0.089360
221799	27_cleveland_1893	people	0.221563
219683	27_cleveland_1893	government	0.148364
219558	27_cleveland_1893	frugality	0.128050
222327	27_cleveland_1893	public	0.102520
223228	27_cleveland_1893	service	0.101813
223842	27_cleveland_1893	support	0.099946
216438	27_cleveland_1893	american	0.097267
216217	27_cleveland_1893	activity	0.095964
219684	27_cleveland_1893	governmental	0.095964
217895	27_cleveland_1893	countrymen	0.088564
181670	28_mckinley_1897	congress	0.188773
186897	28_mckinley_1897	revenue	0.168489
185803	28_mckinley_1897	people	0.161797
183687	28_mckinley_1897	government	0.156633
184879	28_mckinley_1897	loans	0.149356
184777	28_mckinley_1897	legislation	0.126367
186331	28_mckinley_1897	public	0.107057
181134	28_mckinley_1897	business	0.106759
183721	28_mckinley_1897	great	0.105322
186911	28_mckinley_1897	revision	0.099571

	document	term	tfidf
157579	29_mckinley_1901	islands	0.216480
154972	29_mckinley_1901	cuba	0.206329
156690	29_mckinley_1901	government	0.153681
156069	29_mckinley_1901	executive	0.147843
157337	29_mckinley_1901	inhabitants	0.147374
154673	29_mckinley_1901	congress	0.141999
158806	29_mckinley_1901	people	0.116839
160610	29_mckinley_1901	states	0.102186
161455	29_mckinley_1901	united	0.100439
159082	29_mckinley_1901	preparation	0.097925
177609	30_roosevelt_theodore_1905	regards	0.199163
177176	30_roosevelt_theodore_1905	problems	0.182463
178957	30_roosevelt_theodore_1905	tasks	0.150068
171599	30_roosevelt_theodore_1905	aright	0.146306
177764	30_roosevelt_theodore_1905	republic	0.121428
175827	30_roosevelt_theodore_1905	life	0.118701
172229	30_roosevelt_theodore_1905	cause	0.116483
174198	30_roosevelt_theodore_1905	faced	0.115730
172618	30_roosevelt_theodore_1905	conditions	0.115373
179876	30_roosevelt_theodore_1905	wish	0.106771
4501	31_taft_1909	interstate	0.206957
1154	31_taft_1909	business	0.201378
7970	31_taft_1909	tariff	0.154802
5400	31_taft_1909	negro	0.153669
7499	31_taft_1909	south	0.129384
3707	31_taft_1909	government	0.121451
6286	31_taft_1909	proper	0.114684
6411	31_taft_1909	race	0.113413
3322	31_taft_1909	feeling	0.111255
1186	31_taft_1909	canal	0.110879
372700	32_wilson_1913	great	0.158659
374090	32_wilson_1913	men	0.142924
372226	32_wilson_1913	familiar	0.141669
376632	32_wilson_1913	stirred	0.141669
376699	32_wilson_1913	studied	0.141669

	document	term	tfidf
377036	32_wilson_1913	things	0.123915
373625	32_wilson_1913	justice	0.105759
372666	32_wilson_1913	government	0.105520
373805	32_wilson_1913	life	0.102168
373882	32_wilson_1913	look	0.100095
62890	33_wilson_1917	wished	0.228593
55898	33_wilson_1917	counsel	0.174639
60365	33_wilson_1917	purpose	0.152933
54229	33_wilson_1917	action	0.149960
61276	33_wilson_1917	shall	0.134404
62085	33_wilson_1917	thought	0.126568
61603	33_wilson_1917	stand	0.121313
61253	33_wilson_1917	set	0.111215
59986	33_wilson_1917	politics	0.108510
56631	33_wilson_1917	drawn	0.104783
197917	34_harding_1921	world	0.196268
190358	34_harding_1921	civilization	0.157095
189440	34_harding_1921	america	0.155684
197744	34_harding_1921	war	0.120631
195651	34_harding_1921	relationship	0.118846
195762	34_harding_1921	republic	0.117160
194584	34_harding_1921	order	0.110128
197368	34_harding_1921	understanding	0.109915
194392	34_harding_1921	new	0.097567
189448	34_harding_1921	amid	0.095077
145902	35_coolidge_1925	country	0.120814
149615	35_coolidge_1925	ought	0.116721
150763	35_coolidge_1925	represents	0.114495
151963	35_coolidge_1925	tax	0.112826
147725	35_coolidge_1925	great	0.109908
150272	35_coolidge_1925	property	0.108285
149741	35_coolidge_1925	party	0.107300
151598	35_coolidge_1925	stands	0.107257
149785	35_coolidge_1925	peace	0.104170
149807	35_coolidge_1925	people	0.101306

	document	term	tfidf
277812	36_hoover_1929	sup	0.296865
273677	36_hoover_1929	government	0.202690
272826	36_hoover_1929	enforcement	0.194371
270030	36_hoover_1929	18th	0.134706
276212	36_hoover_1929	progress	0.132406
273286	36_hoover_1929	federal	0.126183
274031	36_hoover_1929	ideals	0.113418
271124	36_hoover_1929	business	0.108323
274735	36_hoover_1929	laws	0.107051
275771	36_hoover_1929	peace	0.103883
516870	37_roosevelt_franklin_1933	helped	0.215644
517715	37_roosevelt_franklin_1933	leadership	0.191084
520652	37_roosevelt_franklin_1933	stricken	0.129390
515723	37_roosevelt_franklin_1933	emergency	0.123225
515380	37_roosevelt_franklin_1933	discipline	0.117971
519785	37_roosevelt_franklin_1933	respects	0.117971
518214	37_roosevelt_franklin_1933	money	0.113371
518297	37_roosevelt_franklin_1933	national	0.110570
519501	37_roosevelt_franklin_1933	recovery	0.102349
513178	37_roosevelt_franklin_1933	action	0.097007
353170	38_roosevelt_franklin_1937	democracy	0.178041
354668	38_roosevelt_franklin_1937	government	0.177222
356144	38_roosevelt_franklin_1937	millions	0.140722
356669	38_roosevelt_franklin_1937	paint	0.121461
356784	38_roosevelt_franklin_1937	people	0.115789
353662	38_roosevelt_franklin_1937	economic	0.114184
357952	38_roosevelt_franklin_1937	road	0.112944
357203	38_roosevelt_franklin_1937	progress	0.104463
353253	38_roosevelt_franklin_1937	despair	0.100302
356314	38_roosevelt_franklin_1937	nation	0.099688
128195	39_roosevelt_franklin_1941	democracy	0.244486
130690	39_roosevelt_franklin_1941	know	0.189060
133510	39_roosevelt_franklin_1941	speaks	0.183385
127056	39_roosevelt_franklin_1941	br	0.163241
131339	39_roosevelt_franklin_1941	nation	0.162241

	document	term	tfidf
126447	39_roosevelt_franklin_1941	america	0.140133
130832	39_roosevelt_franklin_1941	life	0.117815
133541	39_roosevelt_franklin_1941	spirit	0.114445
129537	39_roosevelt_franklin_1941	freedom	0.109295
126059	39_roosevelt_franklin_1941	1941	0.108911
445728	40_roosevelt_franklin_1945	learned	0.300396
449000	40_roosevelt_franklin_1945	test	0.194731
441025	40_roosevelt_franklin_1945	1945	0.189849
448233	40_roosevelt_franklin_1945	shall	0.173637
449214	40_roosevelt_franklin_1945	trend	0.172292
446187	40_roosevelt_franklin_1945	mistakes	0.159835
446752	40_roosevelt_franklin_1945	peace	0.159442
449099	40_roosevelt_franklin_1945	today	0.154299
449541	40_roosevelt_franklin_1945	upward	0.150172
444572	40_roosevelt_franklin_1945	gain	0.142278
80930	41_truman_1949	world	0.196051
77350	41_truman_1949	nations	0.194029
78232	41_truman_1949	program	0.171656
77816	41_truman_1949	peoples	0.166989
74201	41_truman_1949	democracy	0.154140
75543	41_truman_1949	freedom	0.149297
73520	41_truman_1949	communism	0.147134
77793	41_truman_1949	peace	0.144167
73909	41_truman_1949	countries	0.137013
78550	41_truman_1949	recovery	0.135785
426501	42_eisenhower_1953	free	0.205803
426203	42_eisenhower_1953	faith	0.154561
431891	42_eisenhower_1953	world	0.146449
428777	42_eisenhower_1953	peoples	0.139466
429176	42_eisenhower_1953	productivity	0.133826
430652	42_eisenhower_1953	strength	0.130430
428754	42_eisenhower_1953	peace	0.123845
426504	42_eisenhower_1953	freedom	0.123320
430235	42_eisenhower_1953	shall	0.105970
426923	42_eisenhower_1953	hold	0.105028

	document	term	tfidf
116926	43_eisenhower_1957	world	0.193893
111539	43_eisenhower_1957	freedom	0.179599
115184	43_eisenhower_1957	seek	0.176008
113346	43_eisenhower_1957	nations	0.175538
113812	43_eisenhower_1957	peoples	0.158270
115711	43_eisenhower_1957	strives	0.146428
113789	43_eisenhower_1957	peace	0.136639
111914	43_eisenhower_1957	help	0.132688
110556	43_eisenhower_1957	divided	0.115450
113303	43_eisenhower_1957	mr	0.111886
409772	44_kennedy_1961	let	0.267869
412304	44_kennedy_1961	sides	0.262849
410919	44_kennedy_1961	pledge	0.160960
405630	44_kennedy_1961	ask	0.107713
405862	44_kennedy_1961	begin	0.106495
406989	44_kennedy_1961	dare	0.106495
413893	44_kennedy_1961	world	0.103110
408311	44_kennedy_1961	final	0.102311
410368	44_kennedy_1961	new	0.096600
408118	44_kennedy_1961	explore	0.094223
64288	45_johnson_1965	change	0.276090
64924	45_johnson_1965	covenant	0.242891
68006	45_johnson_1965	man	0.174391
68068	45_johnson_1965	mastery	0.153532
68346	45_johnson_1965	nation	0.152475
71462	45_johnson_1965	union	0.150512
68545	45_johnson_1965	old	0.129184
71304	45_johnson_1965	trying	0.109663
68816	45_johnson_1965	people	0.108677
66851	45_johnson_1965	harvest	0.102355
233702	46_nixon_1969	voices	0.208854
230776	46_nixon_1969	peace	0.144624
229792	46_nixon_1969	let	0.140977
227661	46_nixon_1969	earth	0.139513
229679	46_nixon_1969	know	0.137969

	document	term	tfidf
229988	46_nixon_1969	man	0.135416
230798	46_nixon_1969	people	0.131270
233913	46_nixon_1969	world	0.128264
231921	46_nixon_1969	rhetoric	0.119219
228487	46_nixon_1969	forward	0.113215
450411	47_nixon_1973	america	0.307074
454767	47_nixon_1973	let	0.282212
455751	47_nixon_1973	peace	0.211567
456959	47_nixon_1973	role	0.190395
458888	47_nixon_1973	world	0.177760
455934	47_nixon_1973	policies	0.176224
456799	47_nixon_1973	responsibility	0.164016
455363	47_nixon_1973	new	0.158606
450098	47_nixon_1973	abroad	0.154815
453928	47_nixon_1973	home	0.126653
46065	48_carter_1977	br	0.222574
50348	48_carter_1977	nation	0.191717
47635	48_carter_1977	dream	0.181515
52694	48_carter_1977	strength	0.147104
50408	48_carter_1977	new	0.142111
50159	48_carter_1977	micah	0.118797
53056	48_carter_1977	thee	0.107811
52550	48_carter_1977	spirit	0.107000
49017	48_carter_1977	human	0.101203
47869	48_carter_1977	enhance	0.100016
30704	49_reagan_1981	government	0.162397
27461	49_reagan_1981	americans	0.156895
30939	49_reagan_1981	heroes	0.137410
27918	49_reagan_1981	believe	0.136126
35650	49_reagan_1981	ve	0.115339
33220	49_reagan_1981	productivity	0.104753
35815	49_reagan_1981	weapon	0.104753
30548	49_reagan_1981	freedom	0.102964
29639	49_reagan_1981	dreams	0.101106
35145	49_reagan_1981	today	0.093813

	document	term	tfidf
282676	50_reagan_1985	government	0.161165
282520	50_reagan_1985	freedom	0.159998
284423	50_reagan_1985	nuclear	0.153623
287622	50_reagan_1985	ve	0.153623
287788	50_reagan_1985	weapons	0.140173
284792	50_reagan_1985	people	0.137038
287907	50_reagan_1985	world	0.127236
282934	50_reagan_1985	history	0.104777
282991	50_reagan_1985	human	0.104777
286190	50_reagan_1985	senator	0.102416
344568	51_bush_george_h_w_1989	don	0.186313
343050	51_bush_george_h_w_1989	breeze	0.184416
347375	51_bush_george_h_w_1989	new	0.137266
345532	51_bush_george_h_w_1989	friends	0.136820
344572	51_bush_george_h_w_1989	door	0.133889
350887	51_bush_george_h_w_1989	word	0.131722
347277	51_bush_george_h_w_1989	mr	0.126821
345775	51_bush_george_h_w_1989	hand	0.125086
342980	51_bush_george_h_w_1989	blowing	0.110649
350039	51_bush_george_h_w_1989	things	0.110609
90451	52_clinton_1993	america	0.318908
98928	52_clinton_1993	world	0.226715
90454	52_clinton_1993	americans	0.206865
98138	52_clinton_1993	today	0.185539
91285	52_clinton_1993	change	0.170522
96727	52_clinton_1993	renewal	0.136867
97153	52_clinton_1993	season	0.136867
94046	52_clinton_1993	idea	0.134993
94807	52_clinton_1993	let	0.132521
95813	52_clinton_1993	people	0.129272
208253	53_clinton_1997	century	0.321300
212390	53_clinton_1997	new	0.279600
207438	53_clinton_1997	america	0.199997
213237	53_clinton_1997	promise	0.164327
215915	53_clinton_1997	world	0.135071

	document	term	tfidf
211704	53_clinton_1997	land	0.131027
212330	53_clinton_1997	nation	0.117062
207441	53_clinton_1997	americans	0.115029
215111	53_clinton_1997	time	0.108057
211794	53_clinton_1997	let	0.105270
88677	54_bush_george_w_2001	story	0.341166
81452	54_bush_george_w_2001	america	0.193152
82369	54_bush_george_w_2001	civility	0.160853
86344	54_bush_george_w_2001	nation	0.130448
81330	54_bush_george_w_2001	affirm	0.120640
85052	54_bush_george_w_2001	ideals	0.109491
81455	54_bush_george_w_2001	americans	0.108207
87251	54_bush_george_w_2001	promise	0.108207
82535	54_bush_george_w_2001	compassion	0.107388
82363	54_bush_george_w_2001	citizens	0.106730
417505	55_bush_george_w_2005	freedom	0.349948
414415	55_bush_george_w_2005	america	0.284882
418794	55_bush_george_w_2005	liberty	0.174494
414418	55_bush_george_w_2005	americans	0.140443
422280	55_bush_george_w_2005	tyranny	0.127272
421156	55_bush_george_w_2005	seen	0.110386
419307	55_bush_george_w_2005	nation	0.096199
415202	55_bush_george_w_2005	cause	0.092545
417919	55_bush_george_w_2005	history	0.092422
415134	55_bush_george_w_2005	came	0.091988
135446	56_obama_2009	america	0.148351
140338	56_obama_2009	nation	0.120229
140398	56_obama_2009	new	0.118002
143133	56_obama_2009	today	0.114792
138630	56_obama_2009	generation	0.100654
139802	56_obama_2009	let	0.091100
139618	56_obama_2009	jobs	0.090727
136951	56_obama_2009	crisis	0.087235
138819	56_obama_2009	hard	0.084859
143901	56_obama_2009	women	0.084859

	document	term	tfidf
238615	57_obama_2013	journey	0.167591
235929	57_obama_2013	creed	0.139659
237619	57_obama_2013	generation	0.127260
234435	57_obama_2013	america	0.125044
235539	57_obama_2013	complete	0.114891
240771	57_obama_2013	requires	0.114891
239797	57_obama_2013	people	0.110351
242108	57_obama_2013	time	0.105563
242122	57_obama_2013	today	0.103668
237000	57_obama_2013	evident	0.100896
324425	58_trump_2017	america	0.350162
326606	58_trump_2017	dreams	0.156436
324426	58_trump_2017	american	0.149226
328597	58_trump_2017	jobs	0.142766
330283	58_trump_2017	protected	0.132439
329430	58_trump_2017	obama	0.120288
329787	58_trump_2017	people	0.112370
332022	58_trump_2017	thank	0.109171
325010	58_trump_2017	borders	0.107075
332617	58_trump_2017	ve	0.107075

In [20]:

```
top_tfidf = tfidf_df.sort_values(by=['document','tfidf'], ascending=[True,False]).group
top_tfidf[top_tfidf['term'].str.contains('women')]
```

Out[20]:

	document	term	tfidf
143901	56_obama_2009	women	0.084859

In [21]:

```
top_tfidf[top_tfidf['document'].str.contains('obama')]
```

Out[21]:

	document	term	tfidf
135446	56_obama_2009	america	0.148351
140338	56_obama_2009	nation	0.120229
140398	56_obama_2009	new	0.118002
143133	56_obama_2009	today	0.114792
138630	56_obama_2009	generation	0.100654
139802	56_obama_2009	let	0.091100
139618	56_obama_2009	jobs	0.090727
136951	56_obama_2009	crisis	0.087235
138819	56_obama_2009	hard	0.084859
143901	56_obama_2009	women	0.084859
238615	57_obama_2013	journey	0.167591
235929	57_obama_2013	creed	0.139659
237619	57_obama_2013	generation	0.127260
234435	57_obama_2013	america	0.125044
235539	57_obama_2013	complete	0.114891
240771	57_obama_2013	requires	0.114891
239797	57_obama_2013	people	0.110351
242108	57_obama_2013	time	0.105563
242122	57_obama_2013	today	0.103668
237000	57_obama_2013	evident	0.100896

In [22]:

```
top_tfidf[top_tfidf['document'].str.contains('trump')]
```

Out[22]:

	document	term	tfidf
324425	58_trump_2017	america	0.350162
326606	58_trump_2017	dreams	0.156436
324426	58_trump_2017	american	0.149226
328597	58_trump_2017	jobs	0.142766
330283	58_trump_2017	protected	0.132439
329430	58_trump_2017	obama	0.120288
329787	58_trump_2017	people	0.112370
332022	58_trump_2017	thank	0.109171
325010	58_trump_2017	borders	0.107075
332617	58_trump_2017	ve	0.107075

In [23]:

```
top_tfidf[top_tfidf['document'].str.contains('kennedy')]
```

Out[23]:

	document	term	tfidf
409772	44_kennedy_1961	let	0.267869
412304	44_kennedy_1961	sides	0.262849
410919	44_kennedy_1961	pledge	0.160960
405630	44_kennedy_1961	ask	0.107713
405862	44_kennedy_1961	begin	0.106495
406989	44_kennedy_1961	dare	0.106495
413893	44_kennedy_1961	world	0.103110
408311	44_kennedy_1961	final	0.102311
410368	44_kennedy_1961	new	0.096600
408118	44_kennedy_1961	explore	0.094223

In [24]:

```
import altair as alt
import numpy as np

# Terms in this list will get a red dot in the visualization
term_list = ['war', 'peace']

# adding a little randomness to break ties in term ranking
top_tfidf_plusRand = top_tfidf.copy()
top_tfidf_plusRand['tfidf'] = top_tfidf_plusRand['tfidf'] + np.random.rand(top_tfidf.sh

# base for all visualizations, with rank calculation
base = alt.Chart(top_tfidf_plusRand).encode(
    x = 'rank:O',
    y = 'document:N'
).transform_window(
    rank = "rank()",
    sort = [alt.SortField("tfidf", order="descending")],
    groupby = ["document"],
)

# heatmap specification
heatmap = base.mark_rect().encode(
    color = 'tfidf:Q'
)

# red circle over terms in above list
circle = base.mark_circle(size=100).encode(
    color = alt.condition(
        alt.FieldOneOfPredicate(field='term', oneOf=term_list),
        alt.value('red'),
        alt.value('#FFFFFF00')
    )
)

# text labels, white for darker heatmap colors
text = base.mark_text(baseline='middle').encode(
    text = 'term:N',
    color = alt.condition(alt.datum.tfidf >= 0.23, alt.value('white'), alt.value('black
)
```

```
# display the three superimposed visualizations
(heatmap + circle + text).properties(width = 600)
```

Out[24]:

Questions

1. What is the difference between a tf-idf score and raw word frequency?

Raw Word Frequency counts how often a word appears in a document. While TF-IDF also looks at how rare that word is across all documents, giving higher scores to words that are frequent in one document but uncommon elsewhere.

2. Based on the dataframe above, what is one potential problem or limitation that you notice with tf-idf scores?

TF-IDF can give high importance to rare or unusual words even if they're not semantically meaningful (like typos, names, or specific years), which might not actually reflect key themes of the text.

3. What's another collection of texts that you think might be interesting to analyze with tf-idf scores? Why?

Song lyrics from different decades could be interesting. TF-IDF could help identify decade-specific slang, political concerns, or cultural trends that define the music of each era.

In [24]: