soc128d_notebook_3_stylometry

June 30, 2021

Sociology 128D: Mining Culture Through Text Data: Introduction to Social Data Science

1 Notebook 3: Stylometry

In this notebook, we're going to take our first step toward vector semantics, which is one of the main approaches we'll use in this class and which has had an enormous influence in cultural sociology! Specifically, we are going to build on Notebook 2 by using word and document frequencies to visualize how similar or dissimilar documents are.

Please download the State of the Union Corpus (1790-2018), which was posted to Kaggle by Rachael Tatman and Liling Tan.

[1]:

```
import copy
import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import seaborn as sns

from collections import Counter
from scipy.stats import pearsonr, spearmanr
from sklearn.decomposition import PCA
from sklearn.metrics.pairwise import cosine_similarity

sns.set_theme(style="darkgrid")
```

[2]:

```
sorted(os.listdir("sotu"))
```

[2]:

```
['Adams_1797.txt',
 'Adams_1798.txt',
 'Adams_1799.txt',
 'Adams_1800.txt',
 'Adams_1825.txt',
 'Adams_1826.txt',
 'Adams_1827.txt',
 'Adams_1828.txt',
 'Arthur_1881.txt',
 'Arthur_1882.txt',
 'Arthur_1883.txt',
 'Arthur_1884.txt',
 'Buchanan_1857.txt',
 'Buchanan_1858.txt',
 'Buchanan_1859.txt',
 'Buchanan_1860.txt',
 'Buren_1837.txt',
 'Buren_1838.txt',
 'Buren_1839.txt',
 'Buren_1840.txt',
 'Bush_1989.txt',
 'Bush_1990.txt',
 'Bush_1991.txt',
 'Bush_1992.txt',
 'Bush_2001.txt',
 'Bush_2002.txt',
 'Bush_2003.txt',
 'Bush_2004.txt',
 'Bush_2005.txt',
 'Bush_2006.txt',
 'Bush_2007.txt',
 'Bush_2008.txt',
 'Carter_1978.txt',
 'Carter_1979.txt',
 'Carter_1980.txt',
 'Carter_1981.txt',
 'Cleveland_1885.txt',
 'Cleveland_1886.txt',
 'Cleveland_1887.txt',
 'Cleveland_1888.txt',
 'Cleveland_1893.txt',
 'Cleveland_1894.txt',
 'Cleveland_1895.txt',
 'Cleveland_1896.txt',
 'Clinton_1993.txt',
 'Clinton_1994.txt',
 'Clinton_1995.txt',
 'Clinton_1996.txt',
 'Clinton_1997.txt',
 'Clinton_1998.txt',
 'Clinton_1999.txt',
 'Clinton_2000.txt',
 'Coolidge_1923.txt',
 'Coolidge_1924.txt',
 'Coolidge_1925.txt',
 'Coolidge_1926.txt',
 'Coolidge_1927.txt',
 'Coolidge_1928.txt',
 'Eisenhower_1954.txt',
 'Eisenhower_1955.txt',
 'Eisenhower_1956.txt',
 'Eisenhower_1957.txt',
 'Eisenhower_1958.txt',
 'Eisenhower_1959.txt',
 'Eisenhower_1960.txt',
 'Eisenhower_1961.txt',
 'Fillmore_1850.txt',
 'Fillmore_1851.txt',
 'Fillmore_1852.txt',
 'Ford_1975.txt',
 'Ford_1976.txt',
 'Ford_1977.txt',
 'Grant_1869.txt',
 'Grant_1870.txt',
 'Grant_1871.txt',
 'Grant_1872.txt',
 'Grant_1873.txt',
 'Grant_1874.txt',
 'Grant_1875.txt',
 'Grant_1876.txt',
 'Harding_1921.txt',
 'Harding_1922.txt',
 'Harrison_1889.txt',
 'Harrison_1890.txt',
 'Harrison_1891.txt',
 'Harrison_1892.txt',
 'Hayes_1877.txt',
 'Hayes_1878.txt',
 'Hayes_1879.txt',
 'Hayes_1880.txt',
 'Hoover_1929.txt',
 'Hoover_1930.txt',
 'Hoover_1931.txt',
 'Hoover_1932.txt',
 'Jackson_1829.txt',
 'Jackson_1830.txt',
 'Jackson_1831.txt',
 'Jackson_1832.txt',
 'Jackson_1833.txt',
 'Jackson_1834.txt',
 'Jackson_1835.txt',
 'Jackson_1836.txt',
 'Jefferson_1801.txt',
 'Jefferson_1802.txt',
 'Jefferson_1803.txt',
 'Jefferson_1804.txt',
 'Jefferson_1805.txt',
 'Jefferson_1806.txt',
 'Jefferson_1807.txt',
 'Jefferson_1808.txt',
 'Johnson_1865.txt',
 'Johnson_1866.txt',
 'Johnson_1867.txt',
 'Johnson_1868.txt',
 'Johnson_1964.txt',
 'Johnson_1965.txt',
 'Johnson_1966.txt',
 'Johnson_1967.txt',
 'Johnson_1968.txt',
 'Johnson_1969.txt',
 'Kennedy_1962.txt',
 'Kennedy_1963.txt',
 'Lincoln_1861.txt',
 'Lincoln_1862.txt',
 'Lincoln_1863.txt',
 'Lincoln_1864.txt',
 'Madison_1809.txt',
 'Madison_1810.txt',
 'Madison_1811.txt',
 'Madison_1812.txt',
 'Madison_1813.txt',
 'Madison_1814.txt',
 'Madison_1815.txt',
 'Madison_1816.txt',
 'McKinley_1897.txt',
 'McKinley_1898.txt',
 'McKinley_1899.txt',
 'McKinley_1900.txt',
 'Monroe_1817.txt',
 'Monroe_1818.txt',
 'Monroe_1819.txt',
 'Monroe_1820.txt',
 'Monroe_1821.txt',
 'Monroe_1822.txt',
 'Monroe_1823.txt',
 'Monroe_1824.txt',
 'Nixon_1970.txt',
 'Nixon_1971.txt',
 'Nixon_1972.txt',
 'Nixon_1973.txt',
 'Nixon_1974.txt',
 'Obama_2009.txt',
 'Obama_2010.txt',
 'Obama_2011.txt',
 'Obama_2012.txt',
 'Obama_2013.txt',
 'Obama_2014.txt',
 'Obama_2015.txt',
 'Obama_2016.txt',
 'Pierce_1853.txt',
 'Pierce_1854.txt',
 'Pierce_1855.txt',
 'Pierce_1856.txt',
 'Polk_1845.txt',
 'Polk_1846.txt',
 'Polk_1847.txt',
 'Polk_1848.txt',
 'Reagan_1982.txt',
 'Reagan_1983.txt',
 'Reagan_1984.txt',
 'Reagan_1985.txt',
 'Reagan_1986.txt',
 'Reagan_1987.txt',
 'Reagan_1988.txt',
 'Roosevelt_1901.txt',
 'Roosevelt_1902.txt',
 'Roosevelt_1903.txt',
 'Roosevelt_1904.txt',
 'Roosevelt_1905.txt',
 'Roosevelt_1906.txt',
 'Roosevelt_1907.txt',
 'Roosevelt_1908.txt',
 'Roosevelt_1934.txt',
 'Roosevelt_1935.txt',
 'Roosevelt_1936.txt',
 'Roosevelt_1937.txt',
 'Roosevelt_1938.txt',
 'Roosevelt_1939.txt',
 'Roosevelt_1940.txt',
 'Roosevelt_1941.txt',
 'Roosevelt_1942.txt',
 'Roosevelt_1943.txt',
 'Roosevelt_1944.txt',
 'Roosevelt_1945.txt',
 'Taft_1909.txt',
 'Taft_1910.txt',
 'Taft_1911.txt',
 'Taft_1912.txt',
 'Taylor_1849.txt',
 'Truman_1946.txt',
 'Truman_1947.txt',
 'Truman_1948.txt',
 'Truman_1949.txt',
 'Truman_1950.txt',
 'Truman_1951.txt',
 'Truman_1952.txt',
 'Truman_1953.txt',
 'Trump_2017.txt',
 'Trump_2018.txt',
 'Tyler_1841.txt',
 'Tyler_1842.txt',
 'Tyler_1843.txt',
 'Tyler_1844.txt',
 'Washington_1790.txt',
 'Washington_1791.txt',
 'Washington_1792.txt',
 'Washington_1793.txt',
 'Washington_1794.txt',
 'Washington_1795.txt',
 'Washington_1796.txt',
 'Wilson_1913.txt',
 'Wilson_1914.txt',
 'Wilson_1915.txt',
 'Wilson_1916.txt',
 'Wilson_1917.txt',
 'Wilson_1918.txt',
 'Wilson_1919.txt',
 'Wilson_1920.txt',
 'sotu']
```

[3]:

```python
address_paths = [os.path.join("sotu", f) for f in os.listdir("sotu") if f.
 ↪endswith(".txt")]
```

[4]:

```python
print(open(address_paths[0], "r").read())
```

```
Gentlemen of the Senate and Gentlemen of the House of Representatives:

I was for some time apprehensive that it would be necessary, on account of
the contagious sickness which afflicted the city of Philadelphia, to
convene the National Legislature at some other place. This measure it was
desirable to avoid, because it would occasion much public inconvenience and
a considerable public expense and add to the calamities of the inhabitants
of this city, whose sufferings must have excited the sympathy of all their
fellow citizens. Therefore, after taking measures to ascertain the state
and decline of the sickness, I postponed my determination, having hopes,
```

now happily realized, that, without hazard to the lives or health of the members, Congress might assemble at this place, where it was next by law to meet. I submit, however, to your consideration whether a power to postpone the meeting of Congress, without passing the time fixed by the Constitution upon such occasions, would not be a useful amendment to the law of 1794.

Although I can not yet congratulate you on the reestablishment of peace in Europe and the restoration of security to the persons and properties of our citizens from injustice and violence at sea, we have, nevertheless, abundant cause of gratitude to the source of benevolence and influence for interior tranquillity and personal security, for propitious seasons, prosperous agriculture, productive fisheries, and general improvements, and, above all, for a rational spirit of civil and religious liberty and a calm but steady determination to support our sovereignty, as well as our moral and our religious principles, against all open and secret attacks.

Our envoys extraordinary to the French Republic embarked--one in July, the other in August--to join their colleague in Holland. I have received intelligence of the arrival of both of them in Holland, from whence they all proceeded on their journeys to Paris within a few days of the 19th of September. Whatever may be the result of this mission, I trust that nothing will have been omitted on my part to conduct the negotiation to a successful conclusion, on such equitable terms as may be compatible with the safety, honor and interest of the United States. Nothing, in the mean time, will contribute so much to the preservation of peace and the attainment of justice as manifestation of that energy and unanimity of which on many former occasions the people of the United States have given such memorable proofs, and the exertion of those resources for national defense which a beneficent Providence has kindly placed within their power.

It may be confidently asserted that nothing has occurred since the adjournment of Congress which renders inexpedient those precautionary measures recommended by me to the consideration of the two Houses at the opening of your late extraordinary session. If that system was then prudent, it is more so now, as increasing depredations strengthen the reasons for its adoption.

Indeed, whatever may be the issue of the negotiation with France, and whether the war in Europe is or is not to continue, I hold it most certain that permanent tranquillity and order will not soon be obtained. The state of society has so long been disturbed, the sense of moral and religious obligations so much weakened, public faith and national honor have been so impaired, respect to treaties has been so diminished, and the law of nations has lost so much of its force, while pride, ambition, avarice and violence have been so long unrestrained, there remains no reasonable ground on which to raise an expectation that a commerce without protection or defense will not be plundered.

The commerce of the United States is essential, if not to their existence, at least to their comfort, their growth, prosperity, and happiness. The genius, character, and habits of the people are highly commercial. Their cities have been formed and exist upon commerce. Our agriculture, fisheries, arts, and manufactures are connected with and depend upon it. In short, commerce has made this country what it is, and it can not be destroyed or neglected without involving the people in poverty and distress. Great numbers are directly and solely supported by navigation. The faith of society is pledged for the preservation of the rights of commercial and sea faring no less than of the other citizens. Under this view of our affairs, I should hold myself guilty of a neglect of duty if I forbore to recommend that we should make every exertion to protect our commerce and to place our country in a suitable posture of defense as the only sure means of preserving both.

I have entertained an expectation that it would have been in my power at the opening of this session to have communicated to you the agreeable information of the due execution of our treaty with His Catholic Majesty respecting the withdrawing of his troops from our territory and the demarcation of the line of limits, but by the latest authentic intelligence Spanish garrisons were still continued within our country, and the running of the boundary line had not been commenced. These circumstances are the more to be regretted as they can not fail to affect the Indians in a manner injurious to the United States. Still, however, indulging the hope that the answers which have been given will remove the objections offered by the Spanish officers to the immediate execution of the treaty, I have judged it proper that we should continue in readiness to receive the posts and to run the line of limits. Further information on this subject will be communicated in the course of the session.

In connection with this unpleasant state of things on our western frontier it is proper for me to mention the attempts of foreign agents to alienate the affections of the Indian nations and to excite them to actual hostilities against the United States. Great activity has been exerted by those persons who have insinuated themselves among the Indian tribes residing within the territory of the United States to influence them to transfer their affections and force to a foreign nation, to form them into a confederacy, and prepare them for war against the United States. Although measures have been taken to counteract these infractions of our rights, to prevent Indian hostilities, and to preserve entire their attachment to the United States, it is my duty to observe that to give a better effect to these measures and to obviate the consequences of a repetition of such practices a law providing adequate punishment for such offenses may be necessary.

The commissioners appointed under the 5th article of the treaty of amity, commerce, and navigation between the United States and Great Britain to ascertain the river which was truly intended under the name of the river St. Croix mentioned in the treaty of peace, met at Passamaquoddy Bay in 1796 October, and viewed the mouths of the rivers in question and the adjacent shores and islands, and, being of opinion that actual surveys of both rivers to their sources were necessary, gave to the agents of the two nations instructions for that purpose, and adjourned to meet at Boston in August. They met, but the surveys requiring more time than had been supposed, and not being then completed, the commissioners again adjourned, to meet at Providence, in the State of Rhode Island, in June next, when we may expect a final examination and decision.

The commissioners appointed in pursuance of the 6th article of the treaty met at Philadelphia in May last to examine the claims of British subjects for debts contracted before the peace and still remaining due to them from citizens or inhabitants of the United States. Various causes have hitherto prevented any determinations, but the business is now resumed, and doubtless will be prosecuted without interruption.

Several decisions on the claims of citizens of the United States for losses and damages sustained by reason of irregular and illegal captures or condemnations of their vessels or other property have been made by the commissioners in London conformably to the 7th article of the treaty. The sums awarded by the commissioners have been paid by the British Government. A considerable number of other claims, where costs and damages, and not captured property, were the only objects in question, have been decided by arbitration, and the sums awarded to the citizens of the United States have also been paid.

The commissioners appointed agreeably to the 21st article of our treaty with Spain met at Philadelphia in the summer past to examine and decide on the claims of our citizens for losses they have sustained in consequence of their vessels and cargoes having been taken by the subjects of His Catholic Majesty during the late war between Spain and France. Their sittings have been interrupted, but are now resumed.

The United States being obligated to make compensation for the losses and damages sustained by British subjects, upon the award of the commissioners acting under the 6th article of the treaty with Great Britain, and for the losses and damages sustained by British subjects by reason of the capture of their vessels and merchandise taken within the limits and jurisdiction of the United States and brought into their ports, or taken by vessels originally armed in ports of the United States, upon the awards of the commissioners acting under the 7th article of the same treaty, it is necessary that provision be made for fulfilling these obligations.

The numerous captures of American vessels by the cruisers of the French Republic and of some by those of Spain have occasioned considerable expenses in making and supporting the claims of our citizens before their tribunals. The sums required for this purpose have in divers instances been disbursed by the consuls of the United States. By means of the same captures great numbers of our sea men have been thrown ashore in foreign countries, destitute of all means of subsistence, and the sick in particular have been exposed to grievous sufferings. The consuls have in these cases also advanced moneys for their relief. For these advances they reasonably expect reimbursements from the United States.

The consular act relative to sea men requires revision and amendment. The provisions for their support in foreign countries and for their return are found to be inadequate and ineffectual. Another provision seems necessary to be added to the consular act. Some foreign vessels have been discovered sailing under the flag of the United States and with forged papers. It seldom happens that the consuls can detect this deception, because they have no authority to demand an inspection of the registers and sea letters.

Gentlemen of the House of Representatives:

It is my duty to recommend to your serious consideration those objects which by the Constitution are placed particularly within your sphere--the national debts and taxes.

Since the decay of the feudal system, by which the public defense was provided for chiefly at the expense of individuals, the system of loans has been introduced, and as no nation can raise within the year by taxes sufficient sums for its defense and military operations in time of war the sums loaned and debts contracted have necessarily become the subjects of what have been called funding systems. The consequences arising from the continual accumulation of public debts in other countries ought to admonish us to be careful to prevent their growth in our own. The national defense must be provided for as well as the support of Government; but both should be accomplished as much as possible by immediate taxes, and as little as possible by loans.

The estimates for the service of the ensuing year will by my direction be laid before you.

Gentlemen of the Senate and Gentlemen of the House of Representatives:

We are met together at a most interesting period. The situations of the principal powers of Europe are singular and portentous. Connected with some by treaties and with all by commerce, no important event there can be indifferent to us. Such circumstances call with peculiar importunity not less for a disposition to unite in all those measures on which the honor, safety, and prosperity of our country depend than for all the exertions of wisdom and firmness.

```
In all such measures you may rely on my zealous and hearty concurrence.
```

[5]:
```python
def return_sotu_name_year_text(f: str):
    """Return the name, year, and text of a SOTU."""
    doc = open(f, "r").read().strip()
    f = os.path.split(f)[-1] # this
    f = f.replace(".txt", "")
    pres, year = f.split("_")
    return pres, year, doc
```

[6]:
```python
return_sotu_name_year_text(address_paths[0])
```

[6]:
```
('Adams',
 '1797',
 'Gentlemen of the Senate and Gentlemen of the House of Representatives:\n\nI
was for some time apprehensive that it would be necessary, on account of\nthe
contagious sickness which afflicted the city of Philadelphia, to\nconvene the
National Legislature at some other place. This measure it was\ndesirable to
avoid, because it would occasion much public inconvenience and\na considerable
public expense and add to the calamities of the inhabitants\nof this city, whose
sufferings must have excited the sympathy of all their\nfellow citizens.
Therefore, after taking measures to ascertain the state\nand decline of the
sickness, I postponed my determination, having hopes,\nnow happily realized,
that, without hazard to the lives or health of the\nmembers, Congress might
assemble at this place, where it was next by law to\nmeet. I submit, however, to
your consideration whether a power to postpone\nthe meeting of Congress, without
passing the time fixed by the Constitution\nupon such occasions, would not be a
useful amendment to the law of 1794.\n\nAlthough I can not yet congratulate you
on the reestablishment of peace in\nEurope and the restoration of security to
the persons and properties of our\ncitizens from injustice and violence at sea,
we have, nevertheless,\nabundant cause of gratitude to the source of benevolence
and influence for\ninterior tranquillity and personal security, for propitious
seasons,\nprosperous agriculture, productive fisheries, and general
improvements,\nand, above all, for a rational spirit of civil and religious
liberty and a\ncalm but steady determination to support our sovereignty, as well
as our\nmoral and our religious principles, against all open and secret
attacks.\n\nOur envoys extraordinary to the French Republic embarked--one in
July, the\nother in August--to join their colleague in Holland. I have
received\nintelligence of the arrival of both of them in Holland, from whence
they\nall proceeded on their journeys to Paris within a few days of the 19th
of\nSeptember. Whatever may be the result of this mission, I trust that
nothing\nwill have been omitted on my part to conduct the negotiation to
a\nsuccessful conclusion, on such equitable terms as may be compatible with\nthe
safety, honor and interest of the United States. Nothing, in the mean\ntime,
will contribute so much to the preservation of peace and the\nattainment of
justice as manifestation of that energy and unanimity of\nwhich on many former
occasions the people of the United States have given\nsuch memorable proofs, and
```

the exertion of those resources for national\ndefense which a beneficent Providence has kindly placed within their\npower.\n\nIt may be confidently asserted that nothing has occurred since the\nadjournment of Congress which renders inexpedient those precautionary\nmeasures recommended by me to the consideration of the two Houses at the\nopening of your late extraordinary session. If that system was then\nprudent, it is more so now, as increasing depredations strengthen the\nreasons for its adoption.\n\nIndeed, whatever may be the issue of the negotiation with France, and\nwhether the war in Europe is or is not to continue, I hold it most certain\nthat permanent tranquillity and order will not soon be obtained. The state\nof society has so long been disturbed, the sense of moral and religious\nobligations so much weakened, public faith and national honor have been so\nimpaired, respect to treaties has been so diminished, and the law of\nnations has lost so much of its force, while pride, ambition, avarice and\nviolence have been so long unrestrained, there remains no reasonable ground\non which to raise an expectation that a commerce without protection or\ndefense will not be plundered.\n\nThe commerce of the United States is essential, if not to their existence,\nat least to their comfort, their growth, prosperity, and happiness. The\ngenius, character, and habits of the people are highly commercial. Their\ncities have been formed and exist upon commerce. Our agriculture,\nfisheries, arts, and manufactures are connected with and depend upon it. In\nshort, commerce has made this country what it is, and it can not be\ndestroyed or neglected without involving the people in poverty and\ndistress. Great numbers are directly and solely supported by navigation.\nThe faith of society is pledged for the preservation of the rights of\ncommercial and sea faring no less than of the other citizens. Under this\nview of our affairs, I should hold myself guilty of a neglect of duty if I\nforbore to recommend that we should make every exertion to protect our\ncommerce and to place our country in a suitable posture of defense as the\nonly sure means of preserving both.\n\nI have entertained an expectation that it would have been in my power at\nthe opening of this session to have communicated to you the agreeable\ninformation of the due execution of our treaty with His Catholic Majesty\nrespecting the withdrawing of his troops from our territory and the\ndemarcation of the line of limits, but by the latest authentic intelligence\nSpanish garrisons were still continued within our country, and the running\nof the boundary line had not been commenced. These circumstances are the\nmore to be regretted as they can not fail to affect the Indians in a manner\ninjurious to the United States. Still, however, indulging the hope that the\nanswers which have been given will remove the objections offered by the\nSpanish officers to the immediate execution of the treaty, I have judged it\nproper that we should continue in readiness to receive the posts and to run\nthe line of limits. Further information on this subject will be\ncommunicated in the course of the session.\n\nIn connection with this unpleasant state of things on our western frontier\nit is proper for me to mention the attempts of foreign agents to alienate\nthe affections of the Indian nations and to excite them to actual\nhostilities against the United States. Great activity has been exerted by\nthose persons who have insinuated themselves among the Indian tribes\nresiding within the territory of the United States to influence them to\ntransfer their affections and force to a foreign nation, to form them into\na confederacy, and prepare them for war against the United States. Although\nmeasures have been taken to counteract these infractions of our rights, to\nprevent Indian hostilities, and to preserve entire their attachment to the\nUnited States, it is my duty to observe that to give a better effect to\nthese measures and to obviate the consequences of a repetition of such\npractices a law providing adequate punishment for such offenses may be\nnecessary.\n\nThe commissioners appointed under the 5th article of the treaty of amity,\ncommerce, and navigation between the United States and Great Britain to\nascertain the river which was truly intended under the name of the river\nSt. Croix mentioned in the treaty of peace, met at Passamaquoddy Bay in\n1796 October, and viewed the mouths of the rivers in question and the\nadjacent shores and islands, and, being of opinion that actual surveys of\nboth rivers to their sources were necessary, gave to the agents of the two\nnations instructions for that purpose, and adjourned to meet at Boston in\nAugust. They met, but the surveys requiring more time than had been\nsupposed, and not being then completed, the commissioners again adjourned,\nto meet at Providence, in the State of Rhode Island, in June next, when we\nmay expect a final examination and decision.\n\nThe commissioners appointed in pursuance of the 6th article of the treaty\nmet at Philadelphia in May last to examine the claims of British subjects\nfor debts contracted before the peace and still remaining due to them from\ncitizens or inhabitants of the United States. Various causes have hitherto\nprevented any determinations, but the business is now resumed, and\ndoubtless will be prosecuted without interruption.\n\nSeveral decisions on the claims of citizens of the United States for losses\nand damages sustained by reason of irregular and illegal captures or\ncondemnations of their vessels or other property have been made by the\ncommissioners in London conformably to the 7th article of the treaty. The\nsums awarded by the commissioners have been paid by the British Government.\nA considerable number of other claims, where costs and damages, and not\ncaptured property, were the only objects in question, have been decided by\narbitration, and the sums awarded to the citizens of the United States have\nalso been paid.\n\nThe commissioners appointed agreeably to the 21st article of our treaty\nwith Spain met at Philadelphia in the summer past to examine and decide on\nthe claims of our citizens for losses they have sustained in consequence of\ntheir vessels and cargoes having been taken by the subjects of His Catholic\nMajesty during the late war between Spain and France. Their sittings have\nbeen interrupted, but are now resumed.\n\nThe United States being obligated to make compensation for the losses and\ndamages sustained by British subjects, upon the award of the commissioners\nacting under the 6th article of the treaty with Great Britain, and for the\nlosses and damages sustained by British subjects by reason of the capture\nof their vessels and merchandise taken within the limits and jurisdiction\nof the United States and brought into their ports, or taken by vessels\noriginally armed in ports of the United States, upon the awards of the\ncommissioners acting under the 7th article of the same treaty, it is\nnecessary that provision be made for fulfilling these obligations.\n\nThe numerous captures of American vessels by the cruisers of the

```
French\nRepublic and of some by those of Spain have occasioned
considerable\nexpenses in making and supporting the claims of our citizens
before their\ntribunals. The sums required for this purpose have in divers
instances been\ndisbursed by the consuls of the United States. By means of the
same\ncaptures great numbers of our sea men have been thrown ashore in
foreign\ncountries, destitute of all means of subsistence, and the sick
in\nparticular have been exposed to grievous sufferings. The consuls have
in\nthese cases also advanced moneys for their relief. For these advances
they\nreasonably expect reimbursements from the United States.\n\nThe consular
act relative to sea men requires revision and amendment. The\nprovisions for
their support in foreign countries and for their return are\nfound to be
inadequate and ineffectual. Another provision seems necessary\nto be added to
the consular act. Some foreign vessels have been discovered\nsailing under the
flag of the United States and with forged papers. It\nseldom happens that the
consuls can detect this deception, because they\nhave no authority to demand an
inspection of the registers and sea\nletters.\n\nGentlemen of the House of
Representatives:\n\nIt is my duty to recommend to your serious consideration
those objects\nwhich by the Constitution are placed particularly within your
sphere--the\nnational debts and taxes.\n\nSince the decay of the feudal system,
by which the public defense was\nprovided for chiefly at the expense of
individuals, the system of loans has\nbeen introduced, and as no nation can
raise within the year by taxes\nsufficient sums for its defense and military
operations in time of war the\nsums loaned and debts contracted have necessarily
become the subjects of\nwhat have been called funding systems. The consequences
arising from the\ncontinual accumulation of public debts in other countries
ought to admonish\nus to be careful to prevent their growth in our own. The
national defense\nmust be provided for as well as the support of Government; but
both should\nbe accomplished as much as possible by immediate taxes, and as
little as\npossible by loans.\n\nThe estimates for the service of the ensuing
year will by my direction be\nlaid before you.\n\nGentlemen of the Senate and
Gentlemen of the House of Representatives:\n\nWe are met together at a most
interesting period. The situations of the\nprincipal powers of Europe are
singular and portentous. Connected with some\nby treaties and with all by
commerce, no important event there can be\nindifferent to us. Such circumstances
call with peculiar importunity not\nless for a disposition to unite in all those
measures on which the honor,\nsafety, and prosperity of our country depend than
for all the exertions of\nwisdom and firmness.\n\nIn all such measures you may
rely on my zealous and hearty concurrence.')
```

[7]:

```python
presidents = []
years = []
docs = []

for path in address_paths:
    pres, year, doc = return_sotu_name_year_text(path)
    presidents.append(pres)
    years.append(year)
    docs.append(doc)

data = list(zip(presidents, years, docs))

pd.DataFrame(data, columns = ["president", "year", "text"]).head()
```

```
[7]:   president  year                                               text
     0     Adams  1797  Gentlemen of the Senate and Gentlemen of the H…
     1     Adams  1798  Gentlemen of the Senate and Gentlemen of the H…
     2     Adams  1799  Gentlemen of the Senate and Gentlemen of the H…
     3     Adams  1800  Gentlemen of the Senate and Gentlemen of the H…
     4     Adams  1825  Fellow Citizens of the Senate and of the House…
```

```
[8]: df = pd.DataFrame(address_paths, columns = ["file_path"])
     df[["president", "year", "text"]] = df.file_path.apply(lambda x: pd.
      ↪Series(return_sotu_name_year_text(x)))
     df.drop(columns = ["file_path"], inplace = True)
```

```
[9]: df.sort_values(by="year", inplace=True)
     df.reset_index(inplace=True, drop=True)
```

```
[10]: df.head()
```

```
[10]:     president  year                                               text
      0  Washington  1790
      1  Washington  1791  Fellow-Citizens of the Senate and House of Rep…
      2  Washington  1792  Fellow-Citizens of the Senate and House of Rep…
      3  Washington  1793  Fellow-Citizens of the Senate and House of Rep…
      4  Washington  1794  Fellow-Citizens of the Senate and House of Rep…
```

```
[11]: df.drop(index=0, inplace = True)
```

```
[12]: df.head()
```

```
[12]:     president  year                                               text
      1  Washington  1791  Fellow-Citizens of the Senate and House of Rep…
      2  Washington  1792  Fellow-Citizens of the Senate and House of Rep…
      3  Washington  1793  Fellow-Citizens of the Senate and House of Rep…
      4  Washington  1794  Fellow-Citizens of the Senate and House of Rep…
      5  Washington  1795  Fellow-Citizens of the Senate and House of Rep…
```

```
[13]: df[df.president=="Adams"]
```

```
[13]:   president  year                                               text
      7      Adams  1797  Gentlemen of the Senate and Gentlemen of the H…
      8      Adams  1798  Gentlemen of the Senate and Gentlemen of the H…
      9      Adams  1799  Gentlemen of the Senate and Gentlemen of the H…
      10     Adams  1800  Gentlemen of the Senate and Gentlemen of the H…
```

```
35     Adams  1825  Fellow Citizens of the Senate and of the House…
36     Adams  1826  Fellow Citizens of the Senate and of the House…
37     Adams  1827  Fellow Citizens of the Senate and of the House…
38     Adams  1828  Fellow Citizens of the Senate and of the House…
```

[14]:
```
df.year = df.year.apply(int)
```

[15]:
```
df.president = np.where(df.president.eq("Adams") & df["year"].gt(1800),␣
 ↪"Adams2", df.president)
```

[16]:
```
df[df.president=="Adams"]
```

[16]:
```
   president  year                                               text
7      Adams  1797  Gentlemen of the Senate and Gentlemen of the H…
8      Adams  1798  Gentlemen of the Senate and Gentlemen of the H…
9      Adams  1799  Gentlemen of the Senate and Gentlemen of the H…
10     Adams  1800  Gentlemen of the Senate and Gentlemen of the H…
```

[17]:
```
df[df.president=="Adams2"]
```

[17]:
```
   president  year                                               text
35    Adams2  1825  Fellow Citizens of the Senate and of the House…
36    Adams2  1826  Fellow Citizens of the Senate and of the House…
37    Adams2  1827  Fellow Citizens of the Senate and of the House…
38    Adams2  1828  Fellow Citizens of the Senate and of the House…
```

[18]:
```
df.president = np.where(df.president.eq("Bush") & df["year"].gt(2000), "Bush2",␣
 ↪df.president)
df.president = np.where(df.president.eq("Johnson") & df["year"].gt(1900),␣
 ↪"Johnson2", df.president)
df.president = np.where(df.president.eq("Roosevelt") & df["year"].gt(1930),␣
 ↪"Roosevelt2", df.president)
```

[19]:
```
df.president.unique()
```

[19]:
```
array(['Washington', 'Adams', 'Jefferson', 'Madison', 'Monroe', 'Adams2',
       'Jackson', 'Buren', 'Tyler', 'Polk', 'Taylor', 'Fillmore',
       'Pierce', 'Buchanan', 'Lincoln', 'Johnson', 'Grant', 'Hayes',
       'Arthur', 'Cleveland', 'Harrison', 'McKinley', 'Roosevelt', 'Taft',
       'Wilson', 'Harding', 'Coolidge', 'Hoover', 'Roosevelt2', 'Truman',
       'Eisenhower', 'Kennedy', 'Johnson2', 'Nixon', 'Ford', 'Carter',
       'Reagan', 'Bush', 'Clinton', 'Bush2', 'Obama', 'Trump'],
      dtype=object)
```

[20]:
```
len(df.president.unique())
```

[20]:
```
42
```

[21]:
```
df.text = df.text.apply(str.lower)
```

[22]:
```
df.head()
```

[22]:
```
    president  year                                               text
1  Washington  1791  fellow-citizens of the senate and house of rep…
2  Washington  1792  fellow-citizens of the senate and house of rep…
3  Washington  1793  fellow-citizens of the senate and house of rep…
4  Washington  1794  fellow-citizens of the senate and house of rep…
5  Washington  1795  fellow-citizens of the senate and house of rep…
```

[23]:
```
?ord
```

[24]:
```
print(f'a = {ord("a")}, z = {ord("z")}, and space = {ord(" ")}')
```

```
a = 97, z = 122, and space = 32
```

[25]:
```
s = "This is a test string, and it has some punctuation--not a lot, but␣
 ↪some--that we're going to remove."

s2 = ""
for char in s.lower():
    if char == " " or ord(char) in range(97,123):
        s2 += char
    else:
        s2 += " "

s2
```

[25]:
```
'this is a test string  and it has some punctuation  not a lot  but some  that
we re going to remove '
```

[26]:
```
def keep_alphabetical(text: str) -> str:
    """Keep only lowercase a-z"""
    return "".join([char if (ord(char) in range(97,123) or char == " ") else "␣
 ↪" for char in text])


df.text = df.text.apply(lambda x: keep_alphabetical(x))
```

[27]:
```
df.head()
```

[27]:
```
    president  year                                               text
1  Washington  1791  fellow citizens of the senate and house of rep…
2  Washington  1792  fellow citizens of the senate and house of rep…
3  Washington  1793  fellow citizens of the senate and house of rep…
4  Washington  1794  fellow citizens of the senate and house of rep…
5  Washington  1795  fellow citizens of the senate and house of rep…
```

[28]:

```
all_text = " ".join(df.text)

word_frequencies = dict(Counter(all_text.split()))

types_and_counts = sorted(list(word_frequencies.items()), reverse = True, key =␣
 ↪lambda x: x[1])
print(types_and_counts[:100])
```

```
[('the', 149615), ('of', 96394), ('and', 60703), ('to', 60642), ('in', 38521),
('a', 28034), ('that', 21946), ('for', 18954), ('be', 18588), ('our', 17265),
('is', 16932), ('it', 15494), ('by', 14960), ('we', 12624), ('which', 12270),
('as', 12158), ('this', 12011), ('have', 12009), ('with', 11984), ('i', 9514),
('will', 9425), ('on', 9415), ('has', 9017), ('are', 8953), ('not', 8922),
('been', 8732), ('their', 7739), ('from', 7470), ('government', 7056), ('at',
6925), ('all', 6764), ('states', 6451), ('an', 6049), ('or', 5957), ('its',
5711), ('was', 5669), ('but', 5663), ('should', 5134), ('they', 5052),
('congress', 4971), ('united', 4795), ('can', 4412), ('more', 4382), ('these',
4312), ('people', 4052), ('such', 4039), ('year', 3997), ('upon', 3928),
('would', 3766), ('so', 3755), ('them', 3662), ('other', 3654), ('no', 3573),
('country', 3418), ('may', 3382), ('than', 3275), ('any', 3265), ('must', 3238),
('great', 3238), ('those', 3160), ('made', 3136), ('there', 3133), ('who',
3116), ('s', 3112), ('were', 3091), ('public', 3067), ('now', 3062), ('under',
3023), ('new', 2961), ('if', 2889), ('time', 2888), ('one', 2842), ('war',
2800), ('american', 2686), ('last', 2626), ('world', 2477), ('us', 2461),
('only', 2372), ('his', 2346), ('my', 2302), ('had', 2300), ('years', 2297),
('most', 2244), ('every', 2242), ('you', 2237), ('into', 2181), ('state', 2143),
('national', 2122), ('some', 2109), ('law', 2075), ('present', 2067), ('nation',
2066), ('between', 2040), ('when', 2015), ('power', 1969), ('do', 1969),
('shall', 1893), ('peace', 1880), ('general', 1845), ('work', 1845)]
```

[29]:

```
print(f"The corpus has {sum(word_frequencies.values()):,} words.")
```

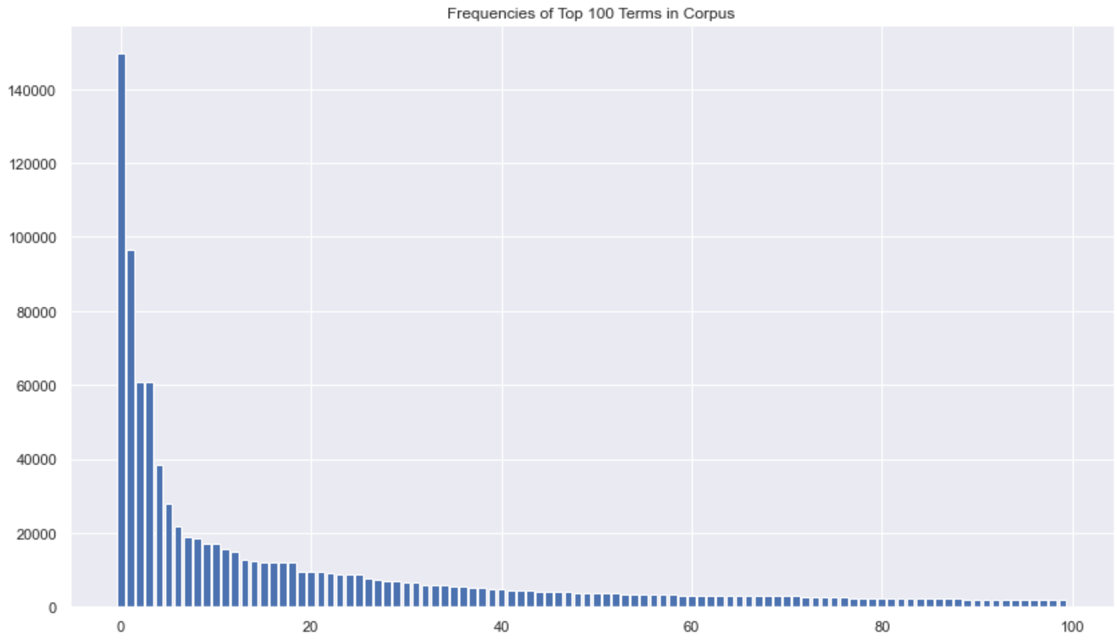
```
The corpus has 1,762,236 words.
```

[30]:

```
types_, token_counts = zip(*types_and_counts)
```

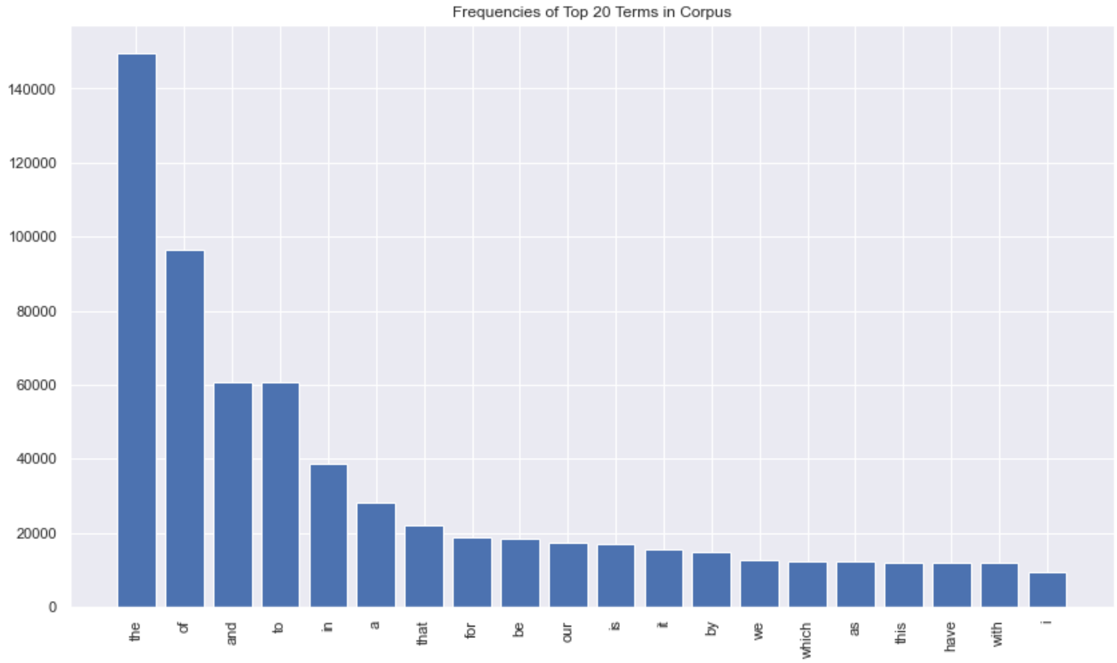
[31]:

```
plt.figure(figsize=(14, 8))
plt.bar(x = range(100), height = token_counts[:100])
plt.title("Frequencies of Top 100 Terms in Corpus")
plt.show()
```



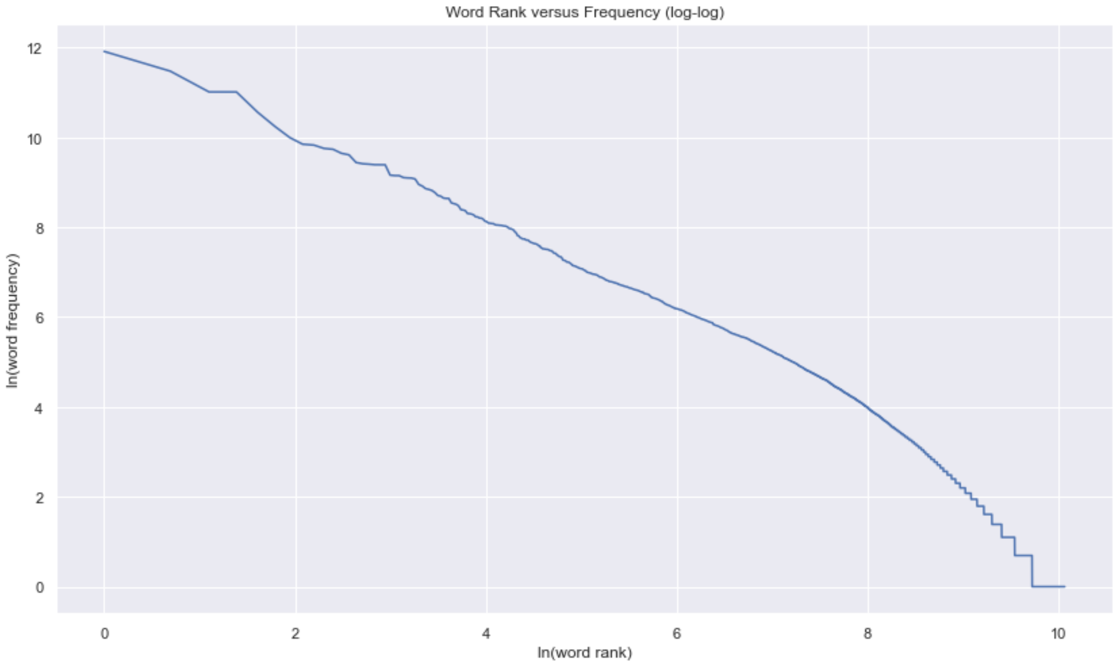
[32]:

```
plt.figure(figsize=(14, 8))
plt.bar(x = types_[:20], height = token_counts[:20])
plt.xticks(rotation = 90)
plt.title("Frequencies of Top 20 Terms in Corpus")
plt.show()
```



[33]:

```
log_rank = np.log(range(1, len(token_counts)+1))
log_frequencies = np.log(token_counts)

plt.figure(figsize=(14, 8))
plt.plot(log_rank, log_frequencies)
plt.ylabel("ln(word frequency)")
plt.xlabel("ln(word rank)")
plt.title("Word Rank versus Frequency (log-log)")
plt.show()
```



[34]:

```
def set_of_types(document: str) -> str:
    return " ".join(list(set(document.split())))
```

[35]:

```
s = "this is a string that repeats some words, like string and words and some"

print(Counter(s.split())) # three types occur twice
```

```
Counter({'string': 2, 'some': 2, 'and': 2, 'this': 1, 'is': 1, 'a': 1, 'that':
1, 'repeats': 1, 'words,': 1, 'like': 1, 'words': 1})
```

[36]:

```
s2 = set_of_types(s)

print(Counter(s2.split())) # each type occurs only once
```

```
Counter({'repeats': 1, 'words': 1, 'that': 1, 'a': 1, 'string': 1, 'words,': 1,
'this': 1, 'like': 1, 'and': 1, 'some': 1, 'is': 1})
```

[37]:
```
df["types"] = df.text.apply(set_of_types)
```

[38]:
```
df.head()
```

[38]:
```
     president  year                                               text  \
  1  Washington  1791  fellow citizens of the senate and house of rep…
  2  Washington  1792  fellow citizens of the senate and house of rep…
  3  Washington  1793  fellow citizens of the senate and house of rep…
  4  Washington  1794  fellow citizens of the senate and house of rep…
  5  Washington  1795  fellow citizens of the senate and house of rep…

                                              types
  1  limits expect occasion completed able laying e…
  2  pretext limits answered ourselves occasion com…
  3  limits installment ourselves occasion balances…
  4  avidity restoring limits misapprehended oursel…
  5  allow ourselves occasion outrages able he ferv…
```

[39]:
```
document_frequencies = dict(Counter(" ".join(df.types).split()))
```

[40]:
```
df.drop(columns=["types"], inplace=True)
```

[41]:
```
vocabulary = sorted(list(word_frequencies.keys()))

x = [word_frequencies[word] for word in vocabulary]
y = [document_frequencies[word] for word in vocabulary]

print("Correlation between each word's frequency in the overall corpus and its␣
 ↪document frequency:")
print(f"Pearson's correlation coefficient: {pearsonr(x, y)[0]:.2f}")
print(f"Spearman's rank-order correlation: {spearmanr(x, y)[0]:.2f}")
```

```
Correlation between each word's frequency in the overall corpus and its document
frequency:
Pearson's correlation coefficient: 0.25
Spearman's rank-order correlation: 0.98
```

[42]:
```
print(len(vocabulary))
```

```
23445
```

If we are interested in analyzing meaning from a corpus, in practice we will often remove words that appear only once or in only one document (which aren't the same thing!). We sometimes call these hapaxes. We can't say that two documents have a word in common if only one document in the entire corpus has the word!

[43]:

```
hapaxes = [word for word in vocabulary if document_frequencies[word] == 1]
print(len(hapaxes))
```

```
7405
```

We may often exclude words that appear in *every* document for similar reasons.

Let’s remove hapaxes.

[44]:

```
word_frequencies = {key:value for key, value in word_frequencies.items() if key␣
 ↪not in hapaxes}
document_frequencies = {key:value for key, value in document_frequencies.
 ↪items() if key not in hapaxes}

assert word_frequencies.keys() == document_frequencies.keys()

types_and_counts = sorted(list(word_frequencies.items()), reverse = True, key =␣
 ↪lambda x: x[1])
vocabulary, _ = zip(*types_and_counts)
```

[45]:

```
print(len(vocabulary))
```

```
16040
```

[46]:

```
df["speech_title"] = df.apply(lambda row: row["president"].lower() + "_" +␣
 ↪str(row["year"]), axis = 1)
df["wordcount"] = df.text.apply(lambda x: len(x.split()))

df.head()
```

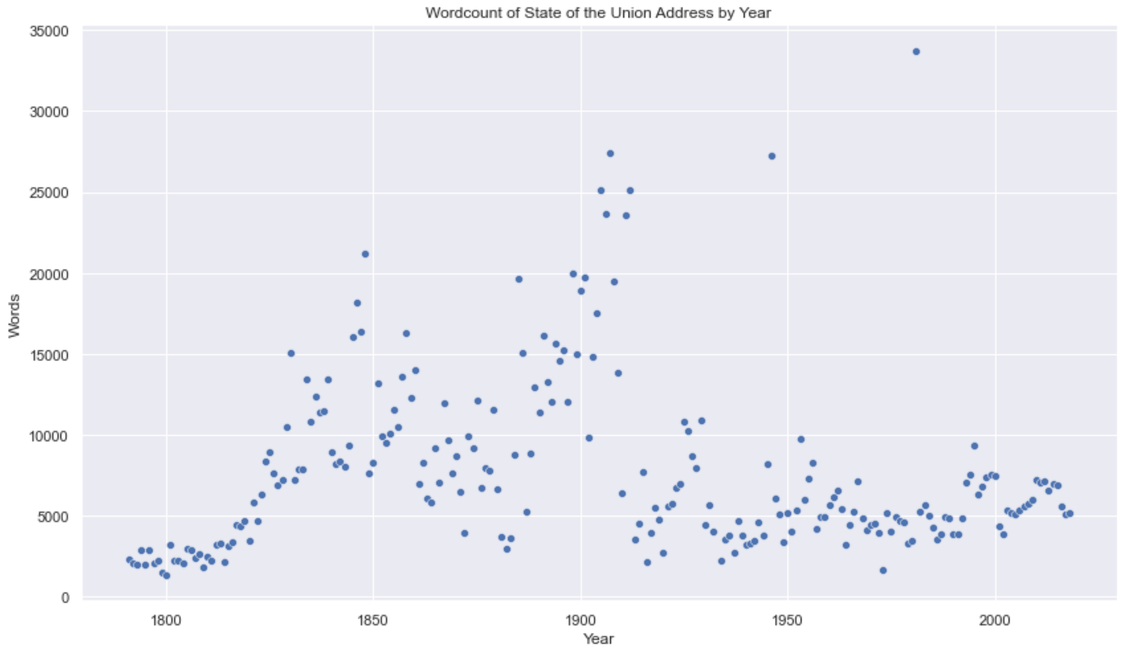
[46]:

```
    president  year                                              text  \
1  Washington  1791  fellow citizens of the senate and house of rep…
2  Washington  1792  fellow citizens of the senate and house of rep…
3  Washington  1793  fellow citizens of the senate and house of rep…
4  Washington  1794  fellow citizens of the senate and house of rep…
5  Washington  1795  fellow citizens of the senate and house of rep…

      speech_title  wordcount
1  washington_1791       2304
2  washington_1792       2092
3  washington_1793       1965
4  washington_1794       2916
5  washington_1795       1988
```

[47]:

```
plt.figure(figsize=(14, 8))
sns.scatterplot(x = "year", y = "wordcount", data = df)
plt.title("Wordcount of State of the Union Address by Year")
plt.xlabel("Year")
plt.ylabel("Words")
plt.plot()
```

```
[47]: []
```



```
[48]: df.wordcount.max()
```

```
[48]: 33704
```

```
[49]: df[df.wordcount.eq(df.wordcount.max())]
```

```
[49]:     president  year                                           text  \
      190    Carter  1981  to the congress of the united states   the sta…

          speech_title  wordcount
      190  carter_1981      33704
```

1.1 Document-Term Matrix

```
[50]: dtm = copy.copy(df)
      dtm.text = dtm.text.apply(str.split)
      dtm = dtm[["speech_title", "text"]]
      dtm.head()
```

```
[50]:       speech_title                                               text
      1  washington_1791  [fellow, citizens, of, the, senate, and, house…
      2  washington_1792  [fellow, citizens, of, the, senate, and, house…
```

```
3  washington_1793  [fellow, citizens, of, the, senate, and, house…
4  washington_1794  [fellow, citizens, of, the, senate, and, house…
5  washington_1795  [fellow, citizens, of, the, senate, and, house…
```

[51]:
```
def term_frequency(doc, vocab):
    return [doc.count(term) for term in vocab]
```

[52]:
```
s = ["the", "cat", "in", "the", "hat"]

term_frequency(s, vocabulary[:10])
```

[52]:
```
[2, 0, 0, 0, 1, 0, 0, 0, 0, 0]
```

[53]:
```
for idx, row in dtm.iterrows():
    print(vocabulary[:10])
    print(term_frequency(row.text, vocabulary[:10]))
    break
```

```
('the', 'of', 'and', 'to', 'in', 'a', 'that', 'for', 'be', 'our')
[242, 159, 73, 88, 41, 42, 32, 22, 34, 5]
```

[54]:
```
sub_voc = vocabulary[:3000]

dtm[list(sub_voc)] = dtm.text.apply(lambda x: pd.Series(term_frequency(x,␣
 ↪sub_voc))) # this takes a moment
```

[55]:
```
dtm.head()
```

[55]:
```
      speech_title                                               text  the  \
1  washington_1791  [fellow, citizens, of, the, senate, and, house…  242
2  washington_1792  [fellow, citizens, of, the, senate, and, house…  195
3  washington_1793  [fellow, citizens, of, the, senate, and, house…  180
4  washington_1794  [fellow, citizens, of, the, senate, and, house…  273
5  washington_1795  [fellow, citizens, of, the, senate, and, house…  174

    of  and   to  in   a  that  for  …  exempt  adjournment  residing  \
1  159   73   88  41  42    32   22  …       0            0         0
2  139   56   88  48  32    24   30  …       0            0         0
3  132   49   74  26  34    12   23  …       0            0         0
4  187   86  138  36  48    39   14  …       0            0         0
5  130   73   64  27  33    27   22  …       0            0         0

   useless  refuse  adding  rejected  liquidation  formation  netherlands
1        0       0       0         0            0          0            0
2        0       0       0         0            0          0            0
3        0       0       0         0            0          0            0
4        0       0       0         0            0          0            0
5        0       0       0         0            0          0            0
```

```
[5 rows x 3002 columns]
```

[56]:
```
dtm.drop(columns="text", inplace=True)
dtm.set_index("speech_title", inplace=True)
```

[57]:
```
dtm.head()
```

[57]:
```
                 the   of  and   to  in   a  that  for  be  our  ...  exempt  \
speech_title                                                     ...
washington_1791  242  159   73   88  41  42    32   22  34    5  ...       0
washington_1792  195  139   56   88  48  32    24   30  29   11  ...       0
washington_1793  180  132   49   74  26  34    12   23  43   16  ...       0
washington_1794  273  187   86  138  36  48    39   14  36   22  ...       0
washington_1795  174  130   73   64  27  33    27   22  22   43  ...       0

                 adjournment  residing  useless  refuse  adding  rejected  \
speech_title
washington_1791            0         0        0       0       0         0
washington_1792            0         0        0       0       0         0
washington_1793            0         0        0       0       0         0
washington_1794            0         0        0       0       0         0
washington_1795            0         0        0       0       0         0

                 liquidation  formation  netherlands
speech_title
washington_1791            0          0            0
washington_1792            0          0            0
washington_1793            0          0            0
washington_1794            0          0            0
washington_1795            0          0            0

[5 rows x 3000 columns]
```

[58]:
```
dtm.shape
```

[58]:
```
(227, 3000)
```

1.2 Plotting Speeches in a 2D Space using Principal Component Analysis

[59]:
```
dtm_std = copy.copy(dtm)
titles = dtm_std.index
dtm_std = dtm_std.to_numpy()

sd = np.std(dtm.to_numpy(), ddof = 1, axis = None)

dtm_std = dtm_std - dtm_std.mean()
dtm_std = dtm_std/sd
```

[60]:
```
dtm_std
```

[60]:
```
array([[10.90212842,  7.12614031,  3.21367071, …, -0.10737907,
        -0.10737907, -0.10737907],
       [ 8.76391829,  6.21626366,  2.44027555, …, -0.10737907,
        -0.10737907, -0.10737907],
       [ 8.0815108 ,  5.89780683,  2.12181873, …, -0.10737907,
        -0.10737907, -0.10737907],
       …,
       [11.94848657,  6.03428833,  8.49095529, …, -0.10737907,
        -0.10737907, -0.10737907],
       [10.53817776,  6.67120199,  9.53731344, …, -0.10737907,
        -0.10737907, -0.10737907],
       [10.21972093,  5.44286851,  9.17336278, …, -0.10737907,
        -0.10737907, -0.10737907]])
```

[61]:
```
dtm_std.mean()
```

[61]:
```
-3.547496771783173e-18
```

[62]:
```
pca = PCA(n_components=2)
components = pca.fit_transform(dtm_std)

pca_df = pd.DataFrame(data = components, columns = ["orig_component1",␣
 ↪"orig_component2"])
```

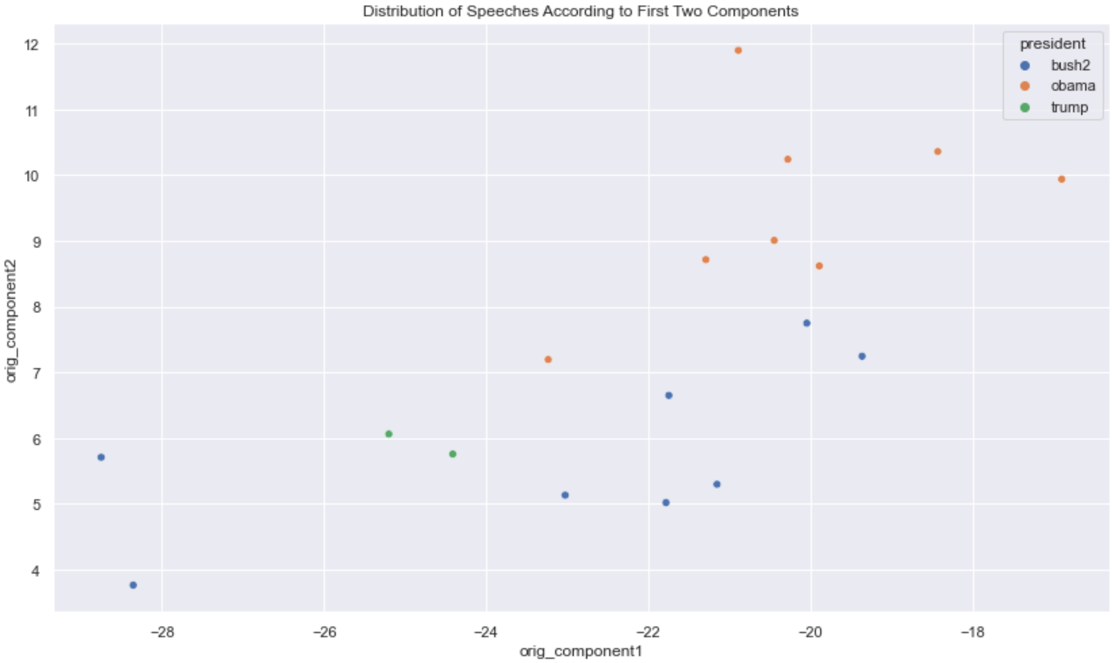
[63]:
```
pca_df["title"] = titles
pca_df[["president", "year"]] = pca_df.title.apply(lambda x: pd.Series(x.
 ↪split("_")))
pca_df.year = pca_df.year.apply(int)
pca_df
```

[63]:
```
     orig_component1  orig_component2            title   president  year
0         -27.784934        -5.191859  washington_1791  washington  1791
1         -30.079653        -4.574890  washington_1792  washington  1792
2         -31.119900        -4.951512  washington_1793  washington  1793
3         -25.231326        -4.559162  washington_1794  washington  1794
4         -31.283500        -4.141323  washington_1795  washington  1795
..               …                …                …          …   …
222       -20.279265        10.241307       obama_2014       obama  2014
223       -20.889429        11.899960       obama_2015       obama  2015
224       -23.234453         7.194619       obama_2016       obama  2016
225       -24.411669         5.755875       trump_2017       trump  2017
226       -25.200196         6.060867       trump_2018       trump  2018
```

```
[227 rows x 5 columns]
```

[64]:

```
mask = pca_df["year"] > 2000

label_points = False

plt.figure(figsize=(14, 8))
sns_plot = sns.scatterplot(x = "orig_component1", y = "orig_component2", data =␣
 ↪pca_df[mask], hue="president")
plt.title("Distribution of Speeches According to First Two Components")
if label_points:
    for idx, row in pca_df[mask].iterrows():
        sns_plot.text(x = row["orig_component1"], y = row["orig_component2"], s␣
 ↪= row["title"])
plt.show()
```



[65]:

```
def return_decade(year):
    return str(year)[:-1] + "0s"
```

[66]:

```
return_decade(1990)
```

[66]:

```
'1990s'
```

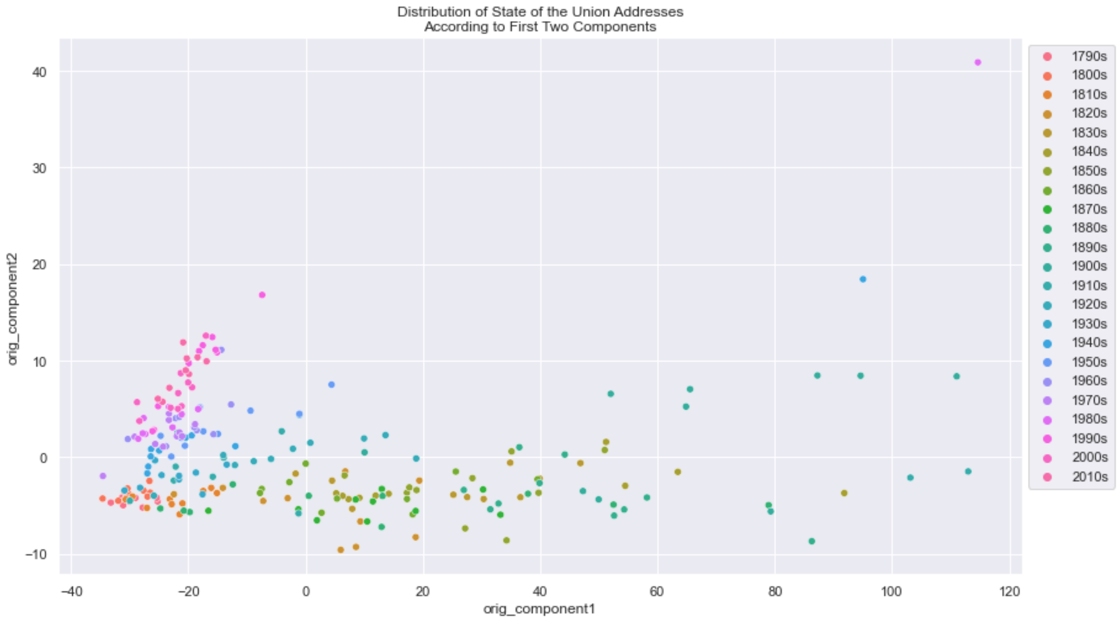
[67]:

```
pca_df["decade"] = pca_df.year.apply(return_decade)
```

```
[68]: pca_df.head()
```

```
[68]:    orig_component1  orig_component2            title   president  year decade
      0        -27.784934        -5.191859  washington_1791  washington  1791  1790s
      1        -30.079653        -4.574890  washington_1792  washington  1792  1790s
      2        -31.119900        -4.951512  washington_1793  washington  1793  1790s
      3        -25.231326        -4.559162  washington_1794  washington  1794  1790s
      4        -31.283500        -4.141323  washington_1795  washington  1795  1790s
```

```
[69]: plt.figure(figsize=(14, 8))
      sns.scatterplot(x = "orig_component1", y = "orig_component2", data = pca_df,␣
       ↪hue="decade")
      plt.title("Distribution of State of the Union Addresses\nAccording to First Two␣
       ↪Components")
      plt.legend(bbox_to_anchor=(1, 1))
      plt.show()
```



1.3 Using TF-IDF to Compare Documents

Let's see if things improve if we use tf-idf weighting.

```
[70]: dtm.head()
```

```
[70]:                  the  of  and  to  in   a  that  for  be  our  …  exempt  \
      speech_title                                                   …
      washington_1791  242  159   73  88  41  42    32   22  34    5  …       0
      washington_1792  195  139   56  88  48  32    24   30  29   11  …       0
```

```
washington_1793  180  132   49   74  26  34    12   23  43   16  ...       0
washington_1794  273  187   86  138  36  48    39   14  36   22  ...       0
washington_1795  174  130   73   64  27  33    27   22  22   43  ...       0

                 adjournment  residing  useless  refuse  adding  rejected  \
speech_title
washington_1791            0         0        0       0       0         0
washington_1792            0         0        0       0       0         0
washington_1793            0         0        0       0       0         0
washington_1794            0         0        0       0       0         0
washington_1795            0         0        0       0       0         0

                 liquidation  formation  netherlands
speech_title
washington_1791            0          0            0
washington_1792            0          0            0
washington_1793            0          0            0
washington_1794            0          0            0
washington_1795            0          0            0

[5 rows x 3000 columns]
```

[71]:

```
def return_idf(N: int, df: int) -> float:
    return np.log10(N/(1 + df))


def tfidf_ind(doc: str, word: str) -> float:
    tf = np.log(1 + doc.count(word))
    idf = idf_dict[word]
    return tf * idf


def tfidf_vocab(doc: str, vocab: list) -> list:
    return [tfidf_ind(doc, word) for word in vocab]

N = dtm.shape[0]

idf_dict = {word: return_idf(N, frequency) for word, frequency in␣
 ↪document_frequencies.items()}
```

[72]:

```
print([key for key, value in idf_dict.items() if value == 0])
```

```
['more', 'can', 'so', 'its', 'government', 'united', 'them', 'they', 'these',
'other', 'from', 'no']
```

[73]:

```
tfidf_mat = copy.copy(df)
tfidf_mat.text = tfidf_mat.text.apply(str.split)
tfidf_mat = tfidf_mat[["speech_title", "text"]]
```

```
tfidf_mat[list(sub_voc)] = tfidf_mat.text.apply(lambda x: pd.
 ↪Series(tfidf_vocab(x, sub_voc))) # this takes a moment
tfidf_mat.drop(columns="text", inplace=True)
tfidf_mat.head()
```

[73]:

```
       speech_title       the        of       and        to        in  \
1  washington_1791 -0.010486 -0.009688 -0.008216 -0.008569 -0.007135
2  washington_1792 -0.010076 -0.009434 -0.007718 -0.008569 -0.007429
3  washington_1793 -0.009924 -0.009336 -0.007468 -0.008242 -0.006292
4  washington_1794 -0.010715 -0.009996 -0.008525 -0.009420 -0.006893
5  washington_1795 -0.009860 -0.009307 -0.008216 -0.007969 -0.006361

          a      that       for        be  …  exempt  adjournment  residing  \
1 -0.007180 -0.006675 -0.005986 -0.006787  …     0.0          0.0       0.0
2 -0.006675 -0.006145 -0.006555 -0.006493  …     0.0          0.0       0.0
3 -0.006787 -0.004896 -0.006067 -0.007224  …     0.0          0.0       0.0
4 -0.007429 -0.007042 -0.005170 -0.006893  …     0.0          0.0       0.0
5 -0.006732 -0.006361 -0.005986 -0.005986  …     0.0          0.0       0.0

   useless  refuse  adding  rejected  liquidation  formation  netherlands
1      0.0     0.0     0.0       0.0          0.0        0.0          0.0
2      0.0     0.0     0.0       0.0          0.0        0.0          0.0
3      0.0     0.0     0.0       0.0          0.0        0.0          0.0
4      0.0     0.0     0.0       0.0          0.0        0.0          0.0
5      0.0     0.0     0.0       0.0          0.0        0.0          0.0

[5 rows x 3001 columns]
```

[74]:

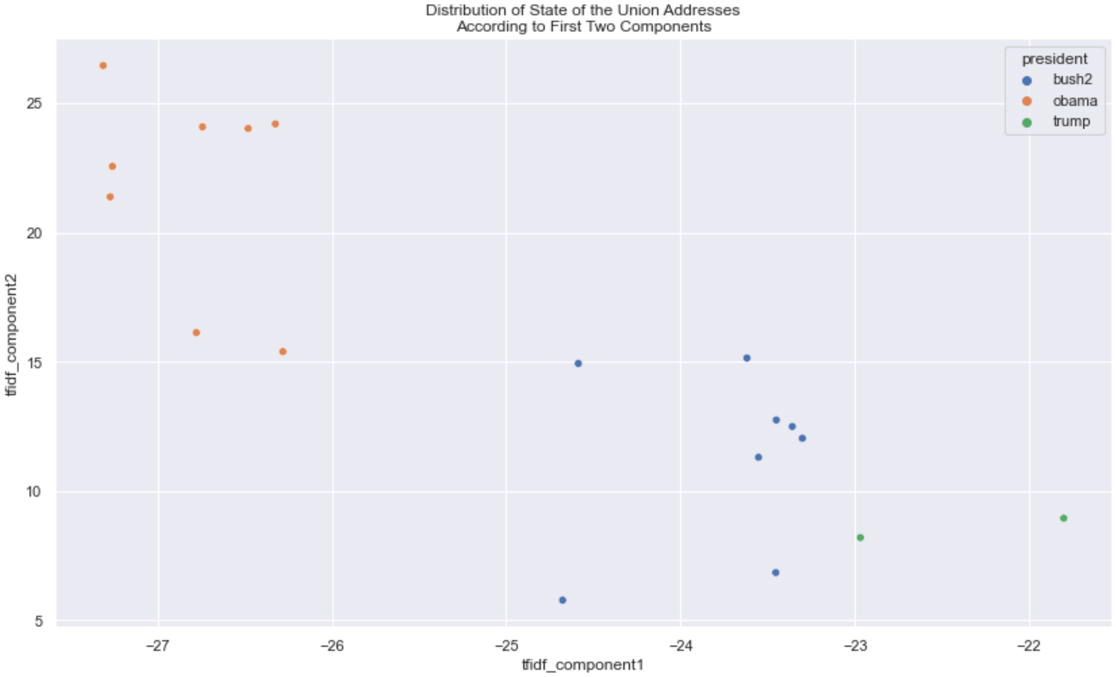
```
tfidf_mat.set_index("speech_title", inplace=True)
titles = tfidf_mat.index
tfidf_mat = tfidf_mat.to_numpy()

sd = np.std(tfidf_mat, ddof = 1, axis = None)

tfidf_mat = tfidf_mat - tfidf_mat.mean()
tfidf_mat = tfidf_mat/sd
```

[75]:

```
tfidf_pca = PCA(n_components=2)
components = tfidf_pca.fit_transform(tfidf_mat)

tfidf_pca_df = pd.DataFrame(data = components, columns = ["tfidf_component1",␣
 ↪"tfidf_component2"])
tfidf_pca_df["title"] = titles
tfidf_pca_df[["president", "year"]] = tfidf_pca_df.title.apply(lambda x: pd.
 ↪Series(x.split("_")))
tfidf_pca_df.year = tfidf_pca_df.year.apply(int)
tfidf_pca_df
```

```
[75]:     tfidf_component1  tfidf_component2             title   president  year
     0          -12.310544        -18.867454  washington_1791  washington  1791
     1          -12.883375        -18.798862  washington_1792  washington  1792
     2          -13.222416        -17.843384  washington_1793  washington  1793
     3          -10.669842        -17.052155  washington_1794  washington  1794
     4          -13.489575        -19.024878  washington_1795  washington  1795
     ..                …                 …               …          …     …
     222        -27.311317         26.454435       obama_2014       obama  2014
     223        -27.270959         21.378253       obama_2015       obama  2015
     224        -26.776966         16.128773       obama_2016       obama  2016
     225        -22.968258          8.205055       trump_2017       trump  2017
     226        -21.801642          8.952187       trump_2018       trump  2018

     [227 rows x 5 columns]
```

```
[76]: mask = tfidf_pca_df["year"] > 2000
      tfidf_pca_df[mask]

      label_points = False

      plt.figure(figsize=(14, 8))
      sns_plot = sns.scatterplot(x = "tfidf_component1", y = "tfidf_component2", data␣
       ↪= tfidf_pca_df[mask], hue="president")
      plt.title("Distribution of State of the Union Addresses\nAccording to First Two␣
       ↪Components")
      if label_points:
          for idx, row in tfidf_pca_df[mask].iterrows():
              sns_plot.text(x = row["tfidf_component1"], y = row["tfidf_component2"],␣
       ↪s = row["title"])
      plt.show()
```



```
[77]: tfidf_pca_df["decade"] = tfidf_pca_df.year.apply(return_decade)
```

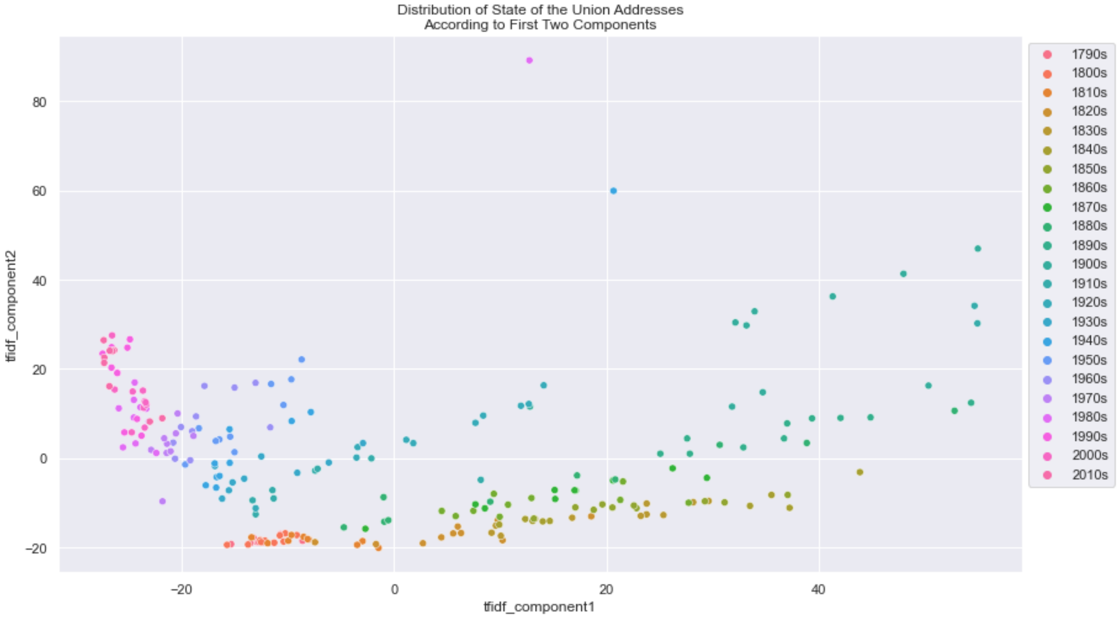
```
[78]: tfidf_pca_df
```

```
[78]:      tfidf_component1  tfidf_component2             title   president  year  \
     0          -12.310544        -18.867454  washington_1791  washington  1791
     1          -12.883375        -18.798862  washington_1792  washington  1792
     2          -13.222416        -17.843384  washington_1793  washington  1793
     3          -10.669842        -17.052155  washington_1794  washington  1794
     4          -13.489575        -19.024878  washington_1795  washington  1795
     ..                …                 …               …          …     …
     222        -27.311317         26.454435       obama_2014       obama  2014
     223        -27.270959         21.378253       obama_2015       obama  2015
     224        -26.776966         16.128773       obama_2016       obama  2016
     225        -22.968258          8.205055       trump_2017       trump  2017
     226        -21.801642          8.952187       trump_2018       trump  2018

         decade
     0    1790s
     1    1790s
     2    1790s
     3    1790s
     4    1790s
     ..     …
     222  2010s
```

```
223  2010s
224  2010s
225  2010s
226  2010s

[227 rows x 6 columns]
```

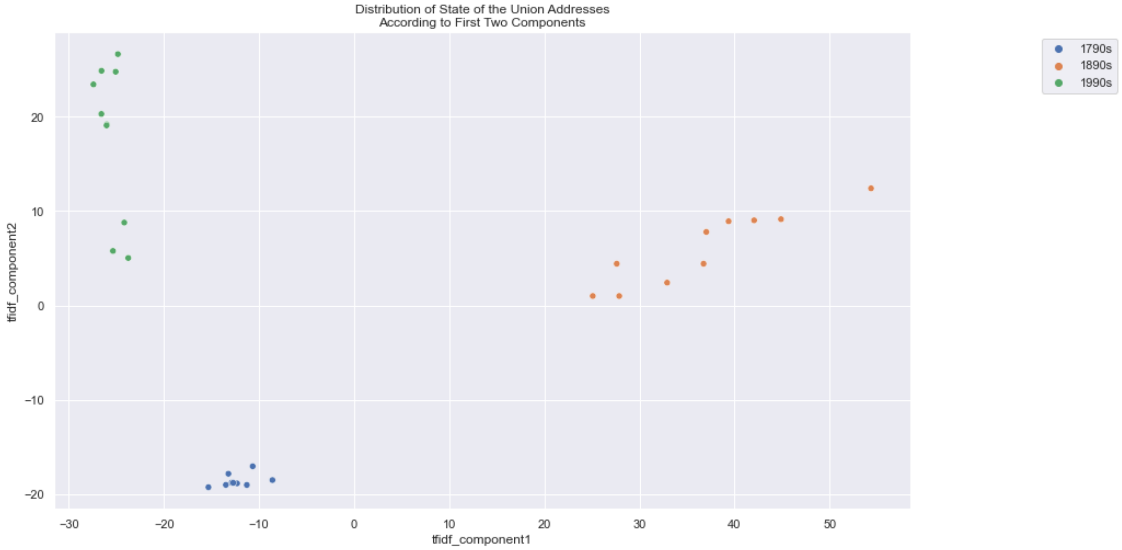
[79]:

```
plt.figure(figsize=(14, 8))
sns.scatterplot(x = "tfidf_component1", y = "tfidf_component2", data =␣
 ↪tfidf_pca_df, hue="decade")
plt.title("Distribution of State of the Union Addresses\nAccording to First Two␣
 ↪Components")
plt.legend(bbox_to_anchor=(1, 1))
plt.show()
```



[80]:

```
mask = tfidf_pca_df.decade.isin(["1790s", "1890s", "1990s"])

label_points = False

plt.figure(figsize=(14, 8))
sns_plot = sns.scatterplot(x = "tfidf_component1", y = "tfidf_component2", data␣
 ↪= tfidf_pca_df[mask], hue="decade")
plt.title("Distribution of State of the Union Addresses\nAccording to First Two␣
 ↪Components")
plt.legend(bbox_to_anchor=(1.25, 1))
if label_points:
    for idx, row in tfidf_pca_df[mask].iterrows():
        sns_plot.text(x = row["tfidf_component1"], y = row["tfidf_component2"],␣
 ↪s = row["title"])
plt.show()
```



1.4 Sparse versus Dense Vectors

```
[81]: print(f"Number of non-zero values in (truncated) document-term matrix: {np.
       ↪count_nonzero(dtm)}")
      print(f"Number of entries in (truncated) document-term matrix: {dtm.size}")
      print(f"{np.count_nonzero(dtm)/dtm.size * 100:.0f}% of entries are zeros, and␣
       ↪that's based on "
            "the 3,000 most frequent words.")
```

```
Number of non-zero values in (truncated) document-term matrix: 259630
Number of entries in (truncated) document-term matrix: 681000
38% of entries are zeros, and that's based on the 3,000 most frequent words.
```