Hands on Data: Assignment Two - Topic Modelling State of The Union Speeches
George Taylor / S3349683

Introduction:

At the beginning of each calendar year, the president of the United States delivers a *State of the Union Speech*[1] to congress. SOTUS act as useful placeholders for events in time, as they mostly document contemporary events. As such, SOTUS are a useful dataset for discovering what topics are central to the American ideology/ethos, over time.

This investigation concerns an exploratory topic model analysis of SOTUS; from 1914 to 2009. My hypothesis is twofold: firstly, what topics appear central to American discourse, and secondly—as this is a predominantly exploratory exercise—is topic modelling a useful tool for discourse analysis, specifically, and Digital Humanities research, broadly?

Methodology:

The topic model for this investigation was provided by David Mimno[2]. The tool, as Mimno states, 'make[s] running topic models easy for anyone with a modern web browser… allow[ing for] tighter integration between models and web-based visualizations.' (ibid.,) The model uses Latent Dirichlet Allocation (LDA)[3] and runs using Mallet.

Upon running the topic model for 50 iterations, the discovered topics are presented alongside a visualisation (Fig.1). Mimno also provides time based visualisations—useful for a rudimentary analysis (Fig.2). These are not very precise, and lack axes. For example, we can assume that health-care/the welfare state has become a more prominent topic, over time (Fig.2); yet the exact years of this development are unclear.

To conduct a thorough analysis, I exported the results from the website as a .csv file, and imported it into Python (Fig.3), renaming the topics (Fig.6). Python affords advanced analysis and visualisation. For example, Fig.13 shows the code used to generate bar-charts, describing the evolution of a topic across time, while Fig.14 shows code to conduct PCA on the mean average topic weights for each year.

Validation:

A benefit of Topic Modelling is that the results are reproducible. Using LDA, and the same data set, anyone can achieve very similar, if not identical, results. While producing the results is unbiased, interpreting them leaves room for subjectivity. Bias can occur when giving labels to topics (Fig.6). While most word clusters necessitate a simple semantic exercise—E.g 'Health Care Work' (Fig.2) suggesting the welfare state—others prove more problematic; Fig.4: 'Let Ask Here'. Intuition and context suggest *Rhetoric.* These words are common to use in speeches: 'Let us ask ourselves, We are all here today…' While this is intuitively true, it is difficult to empirically prove this without manually annotating each speech where the topic is present (Fig.5). Thus, interpretation can lead to invalidation.

Furthermore, running a certain number of iterations, while unlikely, may result in cherrypicked results. In my investigation, running another 50 iterations resulted in Fig.17. As shown, the topics stay generally the same, with the exception of a few word changes. There is also the case of etymology. While language has not changed greatly since the beginning of the 20th century, there are certainly archaic words present in my results that could arguably be outdated. While this is not a big problem in my investigation, when topic modelling over a larger time period, this could be an issue.

[1] SOTUS

[2] https://mimno.infosci.cornell.edu/jsLDA/

[3] Crudely, the LDA model takes a large corpus of texts and discovers 1 to 3 word phrases that are common to each text in the corpus. These terms are only created when they appear in more than 5% and less than 60% of the speeches.

Analysis / Topics:

Generating the results (Fig.8), in and of itself, provides good data for analysis. There are three topics dedicated to war, each with different lexicons (four including *The Middle East*) Notably, one topic is dedicated to war in relation to money: 'Billion dollars war.' This result speaks to those scholars who claim that the US economy runs on war[4]. It is certainly no coincidence that the model generated three—arguably four—topics about war. This insight suggests that war is central to American discourse. We further see topics such as Rights, Nationalism, Liberty, Government, Taxation. These are all central to a common conception of American ideology. This finding indicates that SOTUS are a good representation of American sentiment.

Fig.11 shows a scatter plot for Oil and the Middle East. It is hard to discern a pattern due to the amount of noise (this issue is present in all scatter plots). My hypothesis would be that there is a correlation between the middle east and oil, however, without code for a lo-pass filter, a conclusion is difficult to draw.

Analysis / Topics Over Time:

The topics related to war appear as a constants (Fig.10). They show no clear correlation to time (aside from outliers that occur at the time of a new war). This, when contextualised, makes sense, as the US has been in numerous wars since the beginning of the 20th century.

There are topics that are common sense, such as the welfare state or themes of liberty: 'peace freedom free.' The latter (Fig.9) shows no increase over time, rather, acting as a stable constant. The topic of *rhetoric*, however, shows a slight correlation with time, aside from a few outliers in the 1910s (Fig.7). This may speak to the idea that discourse has developed over time, towards rhetoric and away from substance, an interesting observation that merits further research.

Fig.15 shows the decline of Agriculture over time, whereas Fig.16 shows an increase in the relevance of Children and Education. The former can be explained by the progression that's been made in various industries resulting in an abundance of food. The latter shows that, progressively, children and education has become more important. There is an exception at the beginning of the 20th century, perhaps due to WW1/2 highlighting the importance of winning for the next generations.

Analysis / PCA:

Topic modelling can provide data for PCA (Fig.12). Here, we take the mean topic weights for each year, and then visualise them using a scatter plot, applying colour to indicate time. PC1 demonstrates a sequential, chronological change, suggesting that the content of the speeches are progressively different over time. Frank Evans corroborates these findings, stating that 'the content of the State of the Union addresses is largely driven by the era it is reflecting more that the political association of the president giving it.' (Evans 2015) He furthers notes that 'the outlier is George W. Bush, whose rhetoric appears different from his contemporaries.' This is evident in my analysis (Fig.12) as the later years (2009), show a distinct difference to earlier years. As Bush was president for 8 years: 2001 to 2009, this finding appears reliable.

Reflection:

Topic modelling, in my investigation, has proved to be an invaluable tool for discourse analysis. There were a huge array of findings, many of which I did not have room to cover: from the United Nations, to themes of liberty and individual rights. These results appear reliable and easy to validate when drawing connections to contextual knowledge. Topic modelling appears a useful tool in exploratory analysis, finding correlations and data for further research.

My investigation was somewhat limited, however, due to my inability to create regularisation for scatter plots. As can be seen in Fig.11, a correlation is difficult to ascertain due to noise. Any further investigation would necessitate lo-pass filters. Scatter plots, I believe, can reveal very interesting correlations such as links between liberty and war, taxation and oil or United Nations and Nuclear Weapons, furthering exploratory analysis.

[4] https://www.globalresearch.ca/why-america-needs-war/5328631

A central finding in my investigation was the utility in using PCA alongside topic modelling results. As my section on this concluded, we can ascertain that there is no common thread that runs through all the speeches, but rather, they address contemporary issues from the era they are given in.

Conclusion:

My investigation speaks to two levels. The micro: what topics appear central to American discourse, and the macro: is topic modelling a good discourse analysis/digital humanities research tool.

In regards to the former, my results demonstrate that war is—by far—not only the most common, but also one of the most consistent topics in STOTUS. This potent result is a testament to the number of wars that American is involved in, and shows that it is, and has been, a central topic in American discourse for the entirety of the 20th century. Furthermore, topics such as liberty, rights, Nuclear weapons, government etc, are present. In conclusion, war, themes of freedom, government and international relations appear central to American discourse.

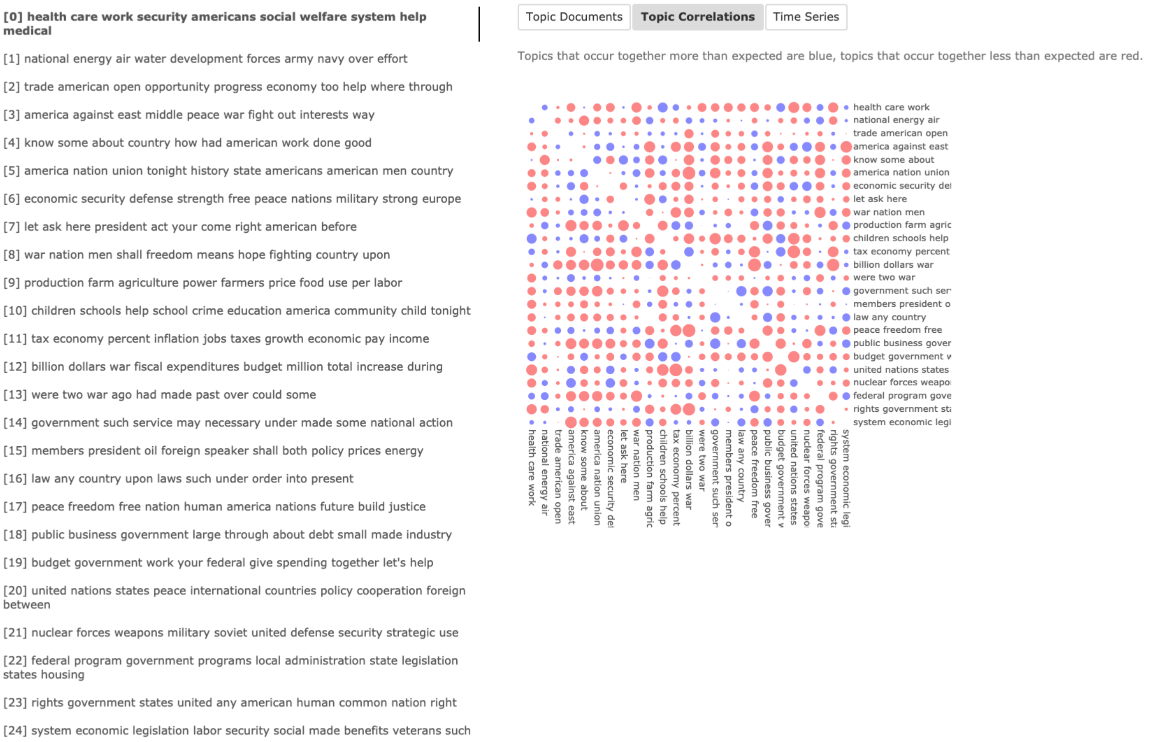
As to the latter, my investigation shows that topic modelling is an invaluable tool for discourse analysis, specifically, and digital humanities research, broadly. Provided with good data, such as SOTUS, researchers may discover a wide array of relevant and important findings: from the evolution of topics over time, to serendipitous connections between topics.

Bibliography:

Evans, Frank. Text Analysis with R: Does POTUS Write the State of the Union or Vice Versa? 2015. https://www.exaptive.com/blog/text-analysis-with-r-does-potus-write-the-state-of-the-union-or-vice-versa. Accessed: January 23rd 2020.

Figures:

Fig.1: Overview of the results and visualisation



[0] health care work security americans social welfare system help medical

[1] national energy air water development forces army navy over effort

[2] trade american open opportunity progress economy too help where through

[3] america against east middle peace war fight out interests way

[4] know some about country how had american work done good

[5] america nation union tonight history state americans american men country

[6] economic security defense strength free peace nations military strong europe

[7] let ask here president act your come right american before

[8] war nation men shall freedom means hope fighting country upon

[9] production farm agriculture power farmers price food use per labor

[10] children schools help school crime education america community child tonight

[11] tax economy percent inflation jobs taxes growth economic pay income

[12] billion dollars war fiscal expenditures budget million total increase during

[13] were two war ago had made past over could some

[14] government such service may necessary under made some national action

[15] members president oil foreign speaker shall both policy prices energy

[16] law any country upon laws such under order into present

[17] peace freedom free nation human america nations future build justice

[18] public business government large through about debt small made industry

[19] budget government work your federal give spending together let's help

[20] united nations states peace international countries policy cooperation foreign between

[21] nuclear forces weapons military soviet united defense security strategic use

[22] federal program government programs local administration state legislation states housing

[23] rights government states united any american human common nation right

[24] system economic legislation labor security social made benefits veterans such

Topics that occur together more than expected are blue, topics that occur together less than expected are red.

Fig.2 - Time based visualisations

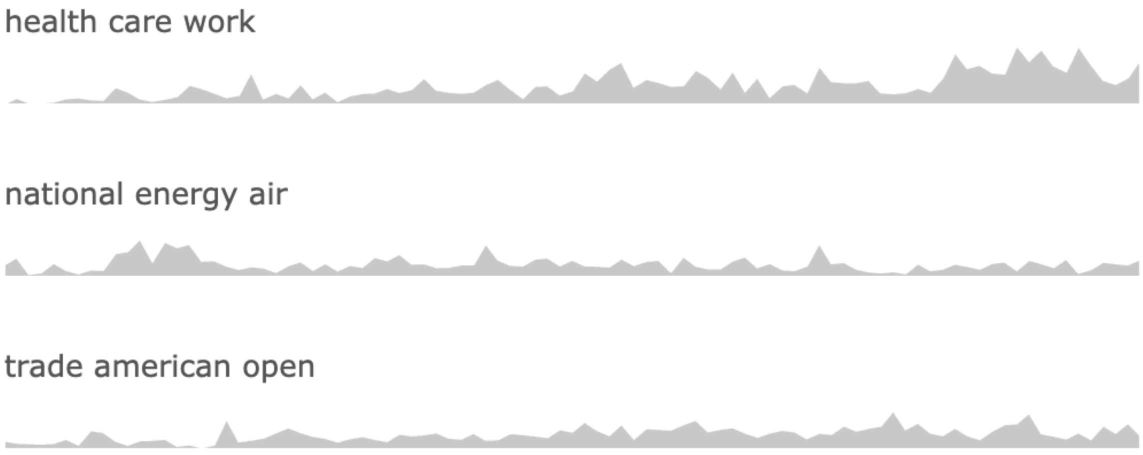
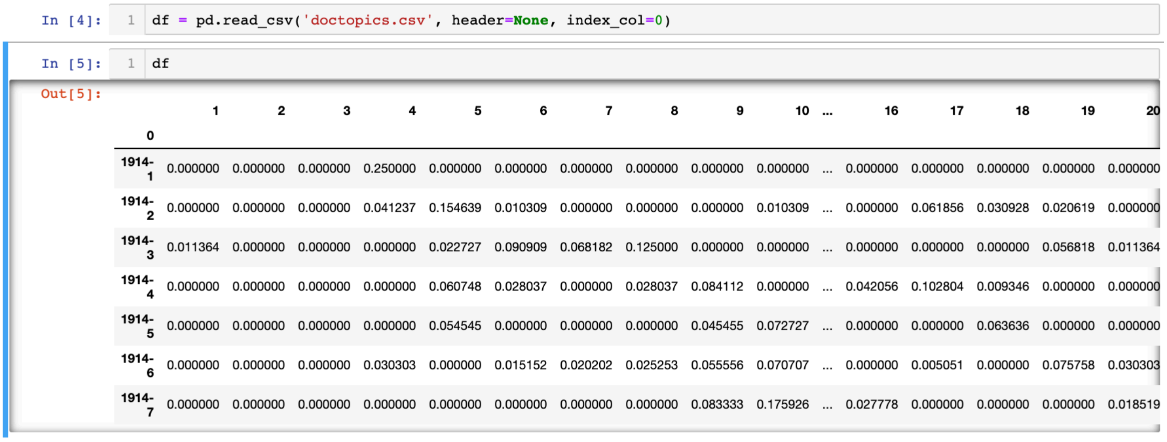


Fig.3: Code for importing the results .csv file into python



```
In [4]:  df = pd.read_csv('doctopics.csv', header=None, index_col=0)

In [5]:  df
Out[5]:
```

0	1	2	3	4	5	6	7	8	9	10	...	16	17	18	19	20
1914-1	0.000000	0.000000	0.000000	0.250000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
1914-2	0.000000	0.000000	0.000000	0.041237	0.154639	0.010309	0.000000	0.000000	0.000000	0.010309	...	0.000000	0.061856	0.030928	0.020619	0.000000
1914-3	0.011364	0.000000	0.000000	0.000000	0.022727	0.090909	0.068182	0.125000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.056818	0.011364
1914-4	0.000000	0.000000	0.000000	0.000000	0.060748	0.028037	0.000000	0.028037	0.084112	0.000000	...	0.042056	0.102804	0.009346	0.000000	0.000000
1914-5	0.000000	0.000000	0.000000	0.000000	0.054545	0.000000	0.000000	0.000000	0.045455	0.072727	...	0.000000	0.000000	0.063636	0.000000	0.000000
1914-6	0.000000	0.000000	0.000000	0.030303	0.000000	0.015152	0.020202	0.025253	0.055556	0.070707	...	0.000000	0.005051	0.000000	0.075758	0.030303
1914-7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.175926	...	0.027778	0.000000	0.000000	0.000000	0.018519

Fig.4: A topic that is difficult to interpret

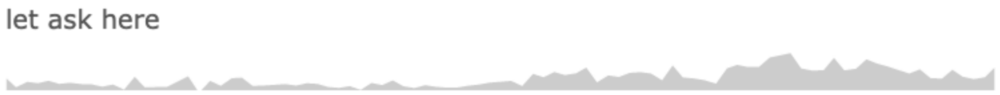
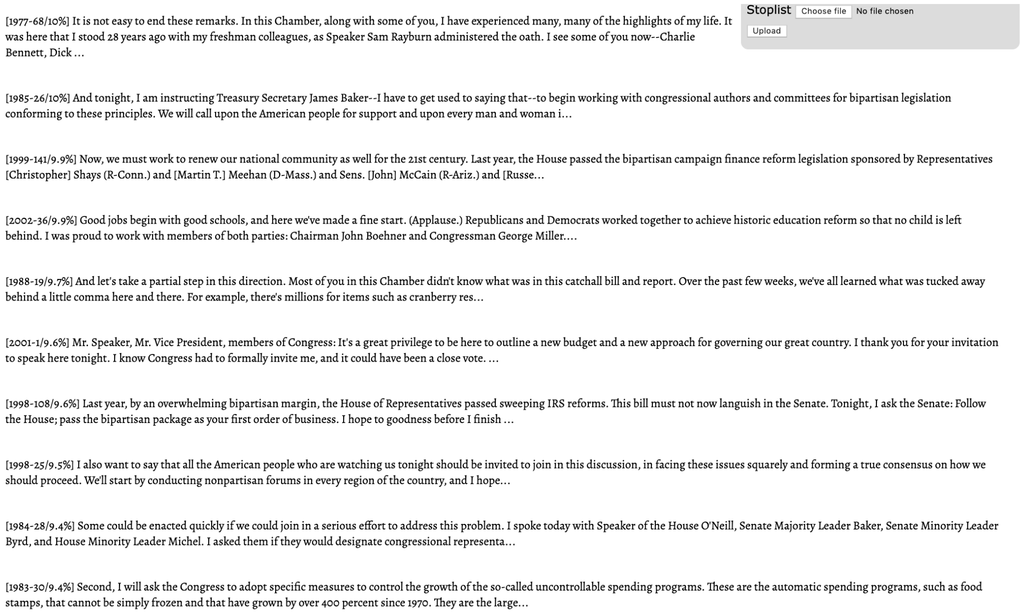


Fig.5: Selection of the documents that the rhetoric topic is present in



[1977-68/10%] It is not easy to end these remarks. In this Chamber, along with some of you, I have experienced many, many of the highlights of my life. It was here that I stood 28 years ago with my freshman colleagues, as Speaker Sam Rayburn administered the oath. I see some of you now--Charlie Bennett, Dick ...

[1985-26/10%] And tonight, I am instructing Treasury Secretary James Baker--I have to get used to saying that--to begin working with congressional authors and committees for bipartisan legislation conforming to these principles. We will call upon the American people for support and upon every man and woman i...

[1999-141/9.9%] Now, we must work to renew our national community as well for the 21st century. Last year, the House passed the bipartisan campaign finance reform legislation sponsored by Representatives [Christopher] Shays (R-Conn.) and [Martin T.] Meehan (D-Mass.) and Sens. [John] McCain (R-Ariz.) and [Russe...

[2002-36/9.9%] Good jobs begin with good schools, and here we've made a fine start. (Applause.) Republicans and Democrats worked together to achieve historic education reform so that no child is left behind. I was proud to work with members of both parties: Chairman John Boehner and Congressman George Miller....

[1988-19/9.7%] And let's take a partial step in this direction. Most of you in this Chamber didn't know what was in this catchall bill and report. Over the past few weeks, we've all learned what was tucked away behind a little comma here and there. For example, there's millions for items such as cranberry res...

[2001-1/9.6%] Mr. Speaker, Mr. Vice President, members of Congress: It's a great privilege to be here to outline a new budget and a new approach for governing our great country. I thank you for your invitation to speak here tonight. I know Congress had to formally invite me, and it could have been a close vote. ...

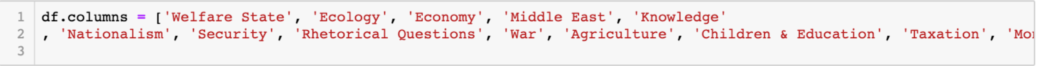
[1998-108/9.6%] Last year, by an overwhelming bipartisan margin, the House of Representatives passed sweeping IRS reforms. This bill must not now languish in the Senate. Tonight, I ask the Senate: Follow the House; pass the bipartisan package as your first order of business. I hope to goodness before I finish ...

[1998-25/9.5%] I also want to say that all the American people who are watching us tonight should be invited to join in this discussion, in facing these issues squarely and forming a true consensus on how we should proceed. We'll start by conducting nonpartisan forums in every region of the country, and I hope...

[1984-28/9.4%] Some could be enacted quickly if we could join in a serious effort to address this problem. I spoke today with Speaker of the House O'Neill, Senate Majority Leader Baker, Senate Minority Leader Byrd, and House Minority Leader Michel. I asked them if they would designate congressional representa...

[1983-30/9.4%] Second, I will ask the Congress to adopt specific measures to control the growth of the so-called uncontrollable spending programs. These are the automatic spending programs, such as food stamps, that cannot be simply frozen and that have grown by over 400 percent since 1970. They are the large...

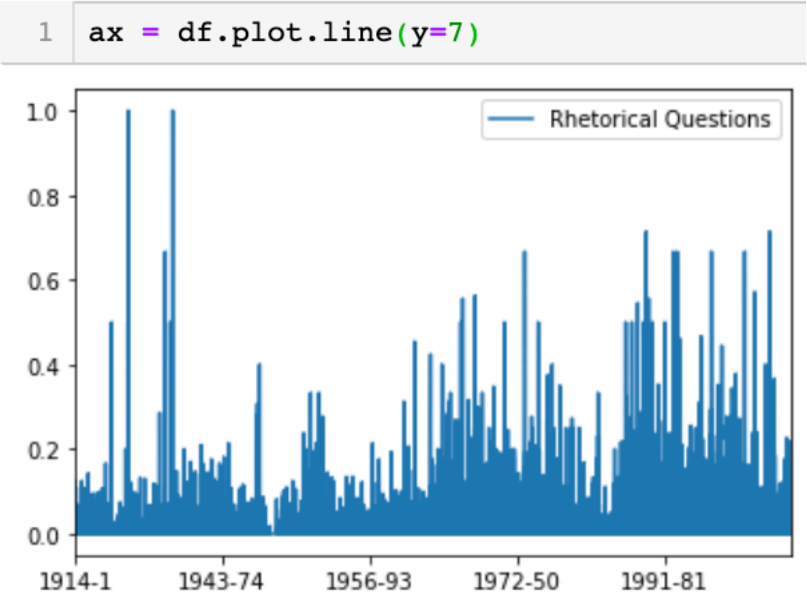
Fig.6: Code for renaming the topics from numbers to words



```
df.columns = ['Welfare State', 'Ecology', 'Economy', 'Middle East', 'Knowledge'
, 'Nationalism', 'Security', 'Rhetorical Questions', 'War', 'Agriculture', 'Children & Education', 'Taxation', 'Mo

```

Fig.7: Graph showing the development of rhetoric over time



```
ax = df.plot.line(y=7)
```

Fig.8: List of topics

```
Welfare State
Ecology
Economy
Middle East
Knowledge
Nationalism
Security
Rhetorical Questions
War
Agriculture
Children & Education
Taxation
Money and War
War_2
Government
Oil
Law
Liberty
Public
Budget
United Nations
Nucelar Weapons
Federal Government
Rights
Economic Legislation
```

Fig.9: Graph showing liberty across time

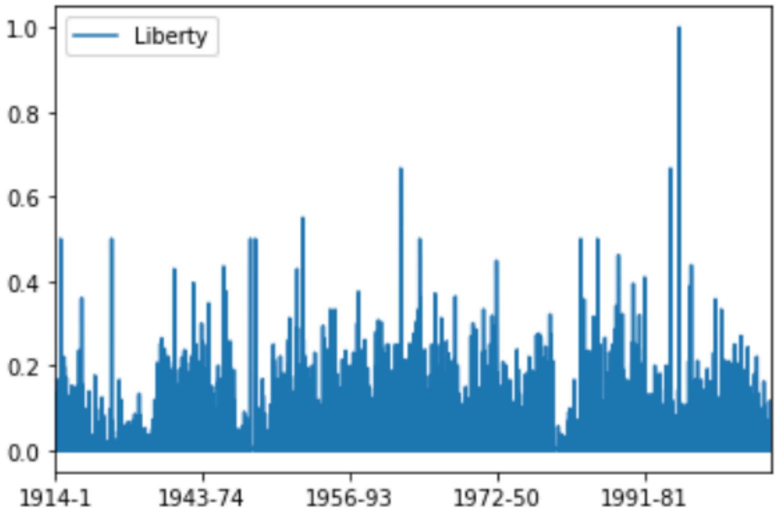


Fig.10: Graph of War over time

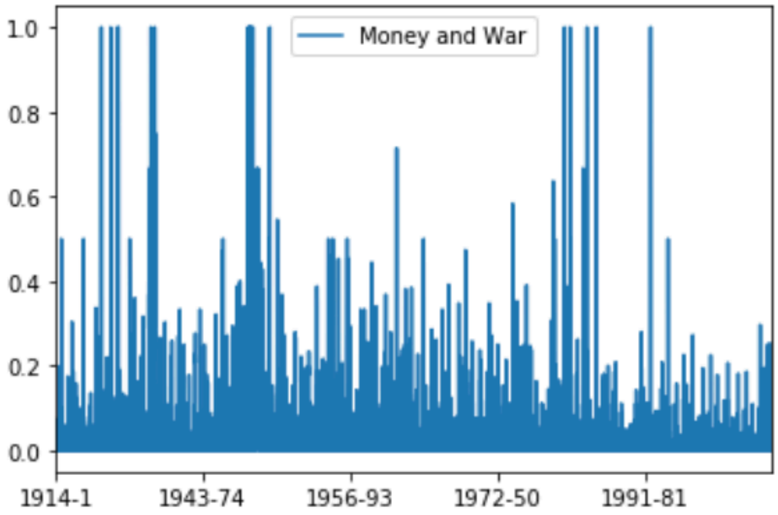


Fig.11: Scatter Plot of Oil & The Middle East

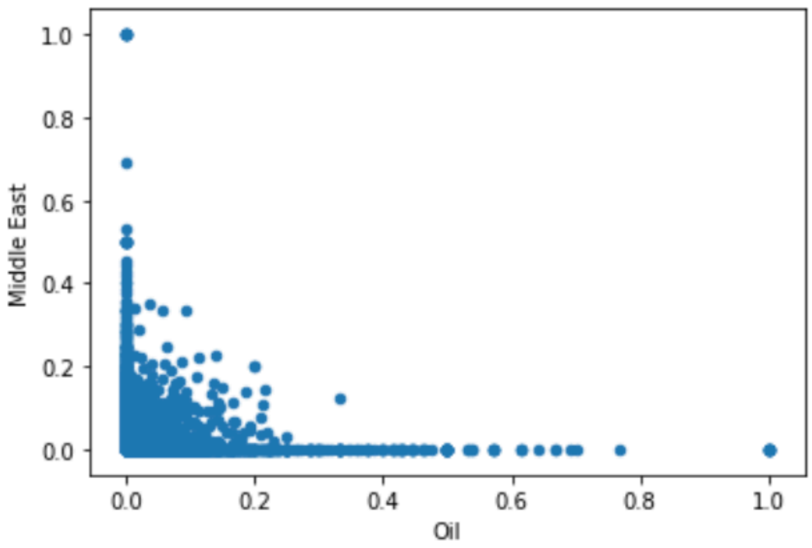


Fig.12: PCA Analysis

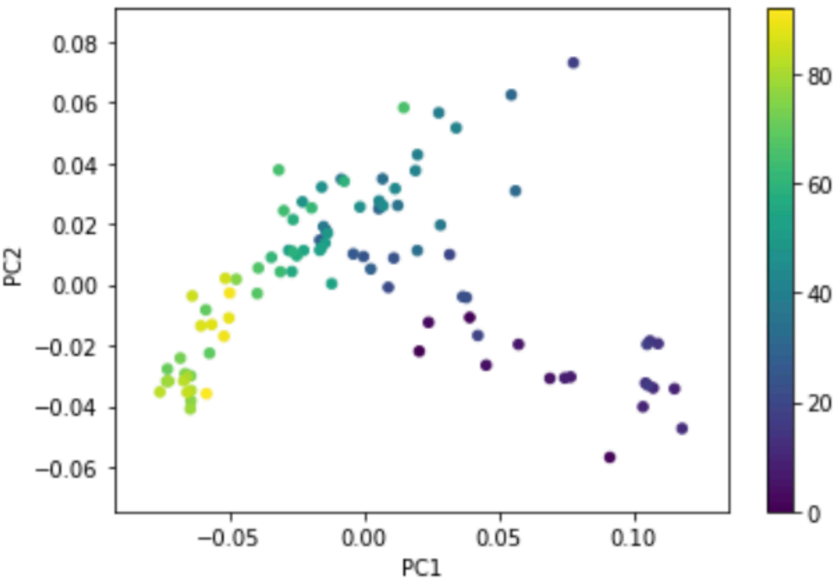
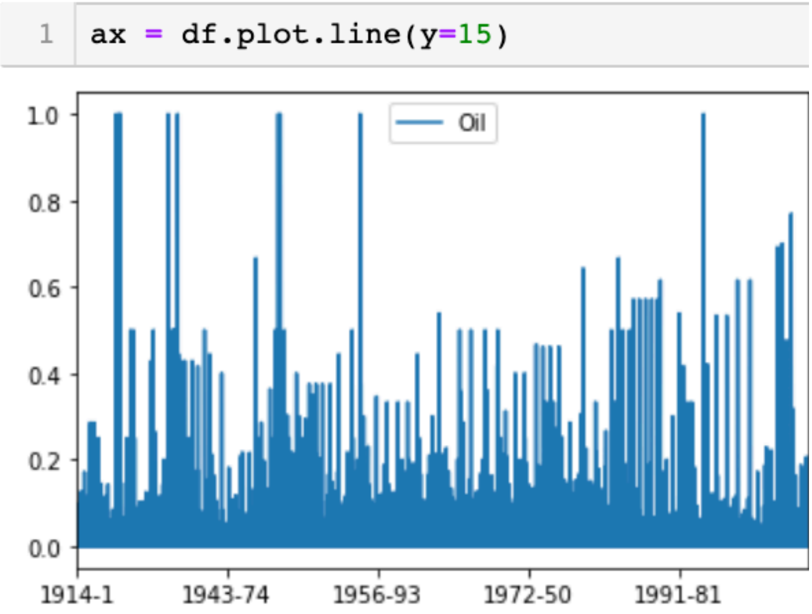
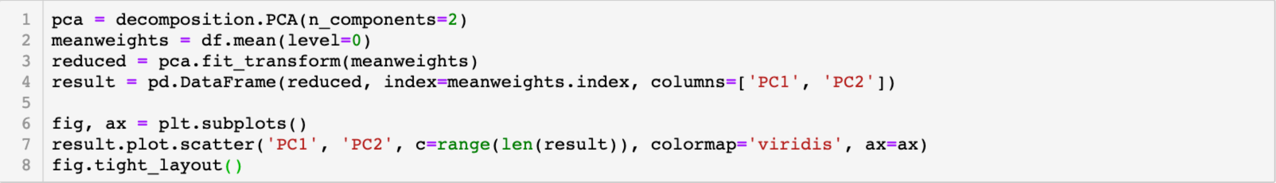


Fig.13: Code for bar-charts of a topic over time



```
ax = df.plot.line(y=15)
```

Fig.14: Code for PCA



```
pca = decomposition.PCA(n_components=2)
meanweights = df.mean(level=0)
reduced = pca.fit_transform(meanweights)
result = pd.DataFrame(reduced, index=meanweights.index, columns=['PC1', 'PC2'])

fig, ax = plt.subplots()
result.plot.scatter('PC1', 'PC2', c=range(len(result)), colormap='viridis', ax=ax)
fig.tight_layout()
```

Fig.15: Graph showing the decline of agriculture over time

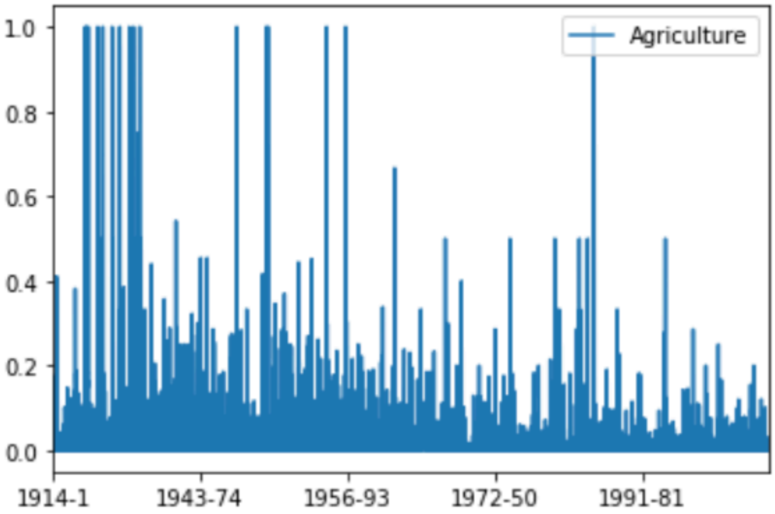


Fig.16: Graph showing the increase of Children & Education over time

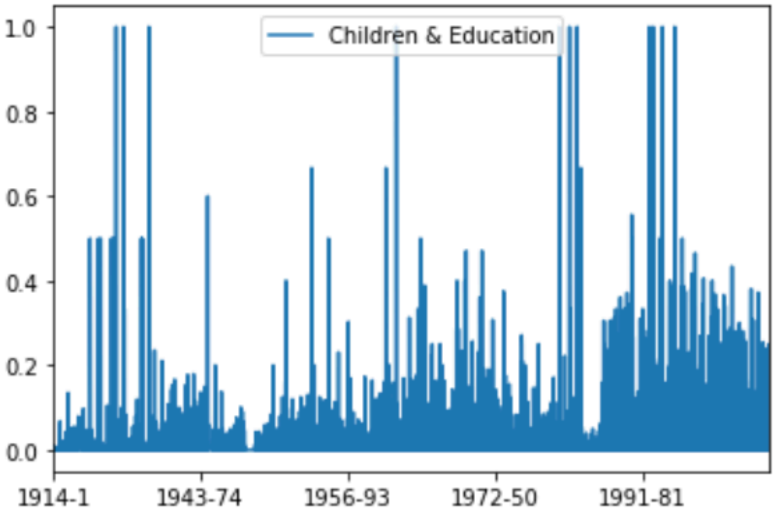
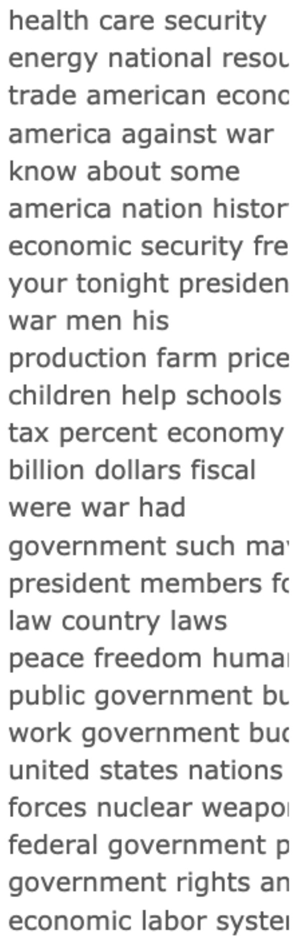


Fig.17: Topics from 100 iterations



health care security
energy national resou
trade american econo
america against war
know about some
america nation histor
economic security fre
your tonight presiden
war men his
production farm price
children help schools
tax percent economy
billion dollars fiscal
were war had
government such ma
president members fo
law country laws
peace freedom huma
public government bu
work government bud
united states nations
forces nuclear weapo
federal government p
government rights an
economic labor syste