**Frontiers of Computational Journalism – Assignment 1**Name – Shreya Vaidyanathan

UNI – sv2525

* **AP Wire Stories**

I tried testing my LSI and LDA algorithms with different number of topics. I began with 50 topics first and then experimented with more topics and found that 60-70 topics were ideal as it seemed to divide the topics into separate buckets more efficiently. I chose to train both LSI and LDA models with 70 topics for the assignment. Then, I went through the topic buckets and tried to identify each document that closely aligned with the scoring that was in the output of the TF-IDF vectors. *(as shown in topicmodeling.py)*

***Annotated topics with LSI:-***

1. **Political policy/agenda** - Soviet - INFO : topic #15(2.078): -0.275\*"party" + 0.245\*"south" + 0.189\*"soviet" + -0.148\*"communist" + 0.141\*"african" + 0.134\*"africa" + 0.131\*"trade" + 0.119\*"dukakis" + 0.114\*"drug" + 0.107\*"mandela"

* The topic appears in documents (AP900330-0061, AP900330-0061 etc.) that talk about the release of Nelson Mandela. The communist and soviet ideas occur more often in this document set so it is hard to separate that idea clearly in this topic per se.

1. **Greyhound bus strikes** - INFO : topic #56(1.514): 0.152\*"air" + -0.143\*"greyhound" + -0.128\*"bus" + 0.102\*"jewish" + 0.099\*"gorbachev" + -0.087\*"irish" + -0.085\*"british" + 0.074\*"died" + 0.072\*"force" + -0.069\*"northern"

* Clearly elicits that there is some news about the greyhound buses that is a topic. In around 4/5 documents, I read about the strike by the bus drivers. The words ‘irish’, ‘british’ seemed to correlate in some articles and the other words are less concentrated or concise within the document-wise segmentation for this topic.

1. **Jewish/Palestine and Arab** - INFO : topic #17(2.045): 0.230\*"israel" + 0.222\*"israeli" + -0.211\*"german" + 0.152\*"palestinian" + 0.144\*"jewish" + 0.134\*"arab" + 0.125\*"palestinians" + 0.124\*"plo" + -0.124\*"air" + 0.123\*"inches"

* This topic immediately points to the news of the conflict in Palestine between different groups, the Arab government in-place and so on. This can be spotted in several documents like AP900608-0188, AP880729-0068, AP880409-0114 etc.

1. **Billboard magazine -** INFO : topic #22(1.921): 0.498\*"billboard" + 0.258\*"publications," + 0.258\*"reprinted" + 0.254\*"copyright" + 0.213\*"videocassettes" + 0.201\*"magazine." + 0.170\*"week's" + 0.154\*"appear" + 0.146\*"popular" + 0.126\*"issue"

* Points to the publicity and news on the Billboard magazine. As seen in the documents AP900726-0205 or AP880825-0205.

1. **Trades in Japan** – INFO : topic #8(2.470): 0.382\*"points," + 0.290\*"nikkei" + 0.272\*"stock" + 0.236\*"tokyo" + 0.204\*"exchange" + 0.163\*"closed" + -0.157\*"dukakis" + -0.122\*"percent" + 0.104\*"average" + 0.097\*"court"

* Links information regarding the Japanese markets which can be a defined context so to say.

1. **Shares and Financial Markets -** INFO : topic #25(1.853): -0.253\*"financial" + -0.243\*"index" + -0.166\*"times-stock" + -0.156\*"100-share" + -0.143\*"500-share" + -0.143\*"times" + -0.130\*"percent," + 0.126\*"average" + -0.125\*"30-share" + -0.123\*"shares"

* This appears to be a bigger topic that encompasses financial news on the markets and the index drop etc. There are several documents that have this topic linked to them in this corpus.

1. **Developments on crude oil with Iran** - INFO : topic #67(1.458): -0.141\*"-" + 0.112\*"opec" + 0.098\*"unofficial" + -0.098\*"percent." + -0.089\*"food" + 0.086\*"latest," + -0.081\*"orders" + 0.080\*"iran" + 0.077\*"iranian" + -0.075\*"crude"

* AP901214-0247 is an example of this topic making sense where it refers to the oil price fluctuation occurring with a change in President Bush’s meeting with Iraq. Other examples where oil and Iran are the man topics also link to this topic at large.

1. **Something that Gorbachev did/said** - INFO : topic #54(1.523): -0.192\*"aid" + 0.134\*"orders" + 0.127\*"bank" + -0.117\*"prices" + 0.109\*"gorbachev" + -0.104\*"rates" + -0.088\*"party" + -0.087\*"interest" + 0.081\*"futures" + 0.077\*"percent"

* Gorbachev is a popular topic that appears in several buckets of the LSI model and this instance is linked in the articles mentioning the Soviet leader’s policies – eg. Strategic policy with US in AP900208-0009

1. **Military orders from US administration -** INFO : topic #35(1.722): -0.150\*"bush" + 0.142\*"reagan" + 0.137\*"police" + 0.105\*"iraqi" + 0.104\*"jackson" + -0.103\*"army" + -0.102\*"richter" + -0.101\*"earthquake" + -0.094\*"court" + 0.092\*"workers"

* Captures the overall human sentiment about the Bush regime’s war waging in Iraq.

1. **Foreign trade on some items -** INFO : topic #63(1.472): -0.130\*"sales" + 0.108\*"cuban" + 0.105\*"chrysler" + 0.096\*"orders" + -0.096\*"milken" + -0.088\*"smoking" + -0.084\*"ms." + -0.083\*"drexel" + -0.079\*"tons" + 0.079\*"insurance"

I saw that the topic buckets (70 in my case) were quite interesting and far ranging. It captured a lot of the same keywords in multiple topics and the human emotion/context was not very clear immediately, but quickly became apparent when I tried to link 4-5 words from the topic bucket to the articles that contained that word.

For example, I found that the word “Greyhound” was showing up in a few topic buckets and I linked them back to documents (eg. AP900314-0159) and found that the context of the article was about Union strikes, but the topic buckets also included some words like “jewish”, “gorbachev” etc. which were not directly linked to this topic or a human emotion per se. Similarly, other documents that fall under the general topic of soviet or wars were also in the same bucket as the ‘Greyhound’ topic which is hard to comprehend intuitively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

***Annotated topics with LDA :-***

1. **Information about space mission from NASA -** INFO : topic #28 (0.013): 0.015\*"venus" + 0.012\*"said" + 0.010\*"anderson" + 0.007\*"reactor" + 0.006\*"said." + 0.006\*"anderson's" + 0.006\*"nasa" + 0.006\*"anti-submarine" + 0.006\*"earth." + 0.006\*"engineers"

* Quite a clear topic that relates to NASA’s space missions and plans relating to Venus.

1. INFO : topic #65 (0.013): 0.014\*"filipino" + 0.010\*"orders" + 0.009\*"said" + 0.009\*"actress" + 0.006\*"u.s." + 0.005\*"new" + 0.005\*"162" + 0.005\*"graceland" + 0.005\*"said." + 0.005\*"helicopter"

* This is harder to tell as this topic appears in random articles and the words like “filipino” appear in different context in comparison to “actress” or “helicopter”

1. **From/About U.S. government documents** - INFO : topic #47 (0.013): 0.007\*"percent" + 0.005\*"u.s." + 0.005\*"will" + 0.005\*"\_" + 0.005\*"united" + 0.005\*"new" + 0.004\*"government" + 0.004\*"million" + 0.004\*"said." + 0.004\*"states"

* Intuitively this links to several of the documents that pertain to the US. Government and the topic number appears in multiple documents without a specific context per se.

1. **Relating to royalty or monarchy/authority -** INFO : topic #50 (0.013): 0.005\*"royal" + 0.005\*"imperial" + 0.004\*"princess" + 0.003\*"emperor" + 0.003\*"queen" + 0.003\*"miyao" + 0.002\*"arrested," + 0.002\*"network," + 0.002\*"gene," + 0.002\*"governors,"
2. **Some extradition chain involving government -** INFO : topic #70 (0.013): 0.003\*"normal" + 0.003\*"\_the" + 0.003\*"footage" + 0.003\*"extradition" + 0.002\*"chain" + 0.002\*"n.j." + 0.002\*"collect" + 0.002\*"pumps" + 0.002\*"stefan" + 0.002\*"parliament,"
3. **Political treaty/development in middle east and soviet -** INFO : topic #9 (0.013): 0.009\*"ec" + 0.008\*"-" + 0.006\*"iraq" + 0.006\*"saudi" + 0.004\*"barriers" + 0.004\*"soviets" + 0.003\*"nato" + 0.003\*"gorbachev's" + 0.003\*"interstate" + 0.003\*"treaty"
4. **Some scandal or news on (Stephen) Gobie -** INFO : topic #12 (0.013): 0.004\*"ethics" + 0.004\*"races" + 0.003\*"10-minute" + 0.003\*"gobie" + 0.002\*"milstead" + 0.002\*"gobie." + 0.002\*"1976." + 0.002\*"strained" + 0.002\*"appointments" + 0.002\*"craft"
5. **Financial status and markets -** INFO : topic #14 (0.013): 0.006\*"nordstrom" + 0.006\*"financial" + 0.004\*"times-stock" + 0.003\*"humphrey" + 0.003\*"lackluster" + 0.003\*"castaneda" + 0.003\*"institutions" + 0.003\*"electric" + 0.003\*"100-share" + 0.003\*"500-share"
6. **About Dixon** and other things - INFO : topic #16 (0.013): 0.004\*"dixon" + 0.003\*"salvage" + 0.003\*"biography" + 0.003\*"dixon's" + 0.002\*"motors," + 0.002\*"convince" + 0.002\*"babies" + 0.002\*"pollock" + 0.002\*"speaker." + 0.002\*"radar."

* This topic appears in 7/8 documents containing the word “Dixon” which is actually referring to completely different people in different contexts. In document AP901016-0102 – the story is about Donald Ray Dixon going to trial, whereas AP900716-0023 talks about the death of rapper Troy Dixon from the rap group “Heavy D & the Boyz”.

1. **About chess championship -** INFO : topic #22 (0.013): 0.007\*"apple" + 0.003\*"kasparov" + 0.003\*"players" + 0.003\*"karpov" + 0.003\*"game" + 0.003\*"tucker" + 0.003\*"trapped" + 0.003\*"championship." + 0.002\*"deficit" + 0.002\*"funding"

|  |
| --- |
| * eg. AP901203-0048 - This topic captures the chess championship games between Gary Kasparov and Antoly Karpov. The words “apple”, “trapped”, “tucker” occur in less interpretable manner in my opinion. |

The LDA algorithm seemed to grab topics and information that was sometimes not seen in the LSI algorithm. For instance, the first topic above captures the word “filipino” and “gobie” that I did not really notice with the LSI model. Moreover, I found that the topic did not provide the exact context for which the word occurs in the document. Reading through the documents (AP900426-0201, AP881022-0144, AP880721-0232), there were a few that were about diplomatic and strategic talks between the Philippines’ government, U.S. Government and Soviet Military and there was one about the economic scare of increase in the filipino population (AP900426-0201). I couldn’t have really inferred that from the topic buckets itself or the scores given to each document by the algorithm.

**LSI vs LDA**

Both models seem to capture the predominantly important topics from the documents. But the consistency and coherency of the models varies in terms of the topics it captures wholly and how efficiently you can interpret it.

I found that the LSI model has certain words in each of the topics that didn’t quite fit into the context or human emotion of the other words. It seemed apparent when you viewed a topic that a certain word was an outlier that didn’t fit at all – but I am not able to explain how or why this occurred. I interpreted the topics with the majority of words that seemed to be closely aligned with each other. The LSI model simply calculates the topic weights and matches the word weights to them to make sure the sum is over a probabilistic distribution perhaps.

But in both the models, I observed that some words like “soviet”, “government”, “Bush” etc. occur in various buckets across the model and this didn’t necessarily point to a specific important relationship or correlation. This seems like expected behaviour for the algorithm to some degree though.

* **State of the Union Speeches**

For this part, I tried to use both LSI and LDA and experimented with the number of topics in the parameters too. I couldn’t really tell whether being closer to 60 topics or 80 topics were better in this case as the words were across different decades and it was essentially the same larger context/purpose (i.e. state of the union address). I just fixed 60 topics and ran the analysis over decades in the 20th century to get the best results that I could analyse further.

I found that the topics of the Speeches were changing only slightly over the decades of the 20th century. I noticed that the top 10 words in certain decades varied upon close inspection only.

For instance, it was interesting to trace back that the topics (below) from the 1930s, 1940s and 1990s had clearly mentioned “war” in the speeches -

INFO : topic#2 (0.017): 0.205\*"war" + 0.087\*"whole" + 0.052\*"yet" + 0.037\*"nation" + 0.031\*"experience" + 0.025\*"enable" + 0.021\*"peace." + 0.021\*"transportation" + 0.015\*"source" + 0.015\*"seen"

INFO : topic #3 (0.017): 0.086\*"peace" + 0.069\*"take" + 0.052\*"come" + 0.035\*"pass" + 0.031\*"century" + 0.031\*"can" + 0.030\*"live" + 0.029\*"improve" + 0.028\*"forward" + 0.027\*"areas"

* But also had similar themes of peace and instilling confidence in the people.

Trying to look for more specific instances of mentions of words that pertain to certain decades, then helped me look at the topics in relation to that with more clarity. It wasn’t necessarily apparent that the 1980s speech talked about Watergate, but if we inspect the vectors of the speeches and the top words in comparison to the previous century or decades, we can try to find the theme perhaps.

Moreover, the commonality in the documents is more apparent with the speeches as they are about the U.S. and its position as a nation.

For instance this topic - INFO : topic #27 (0.017): 0.093\*"state" + 0.059\*"use" + 0.059\*"union" + 0.053\*"them." + 0.051\*"end" + 0.049\*"foreign" + 0.047\*"tonight," + 0.041\*"address" + 0.040\*"congress," + 0.040\*"face" - from the 1970s could be interpreted as being close to the topic - INFO : topic #56 (0.017): 0.106\*"states." + 0.079\*"country," + 0.061\*"united" + 0.024\*"survey" + 0.023\*"union," + 0.020\*"citizen" + 0.019\*"washington" + 0.019\*"compensation" + 0.017\*"judgment" + 0.016\*"conferred" – which is from 19th century.

When looking in to the top words from the corpus :-

**Topic**,**Word**,**P  
0**,**0**,**united**,**0.5812218156327776  
1**,**0**,**states**,**0.5228271151423742  
2**,**0**,**will**,**0.23939900170106357  
3**,**0**,**government**,**0.20925498209209162  
4**,**0**,**states.**,**0.12166239232479924  
5**,**1**,**will**,**-0.5213288428624522  
6**,**1**,**united**,**0.4200655552828855  
7**,**1**,**states**,**0.3668079813131147  
8**,**1**,**government**,**-0.19145288724717  
9**,**1**,**congress**,**-0.18167933640383147  
10**,**2**,**state**,**-0.6544114531031641**

We can easily observe that words like “will” and “state” being used across the speeches and sort of hinders the way we interpret the data vectors. LDA or LSI didn’t really help clearly elicit the themes across the speeches as many words were occurring too frequently. Either we can assume these commonly used words are ‘stop words’ or we could consider and see if there is further analysis needed to prove how important or impactful these words are to a state of the union speech itself.