Foundations of AI
Yunyu Guo
Assignment 7

Collab link with all the outputs:
https://colab.research.google.com/drive/1oVitFuzstCvtR3T_WBK-1LzDasEAIlpR?usp=sharing

a. I recommend converting the documents to lowercase and filtering out stopwords. **Start with k = 10 topics. Fit an LDA object to the set of all news text. Then, examine the top n words from each topic (choose a reasonable n such as 10 or 20). How well do the topics represent real-world topics?** (One sentence)

```
1. Topic Usage Patterns in True vs Fake News:
--------------------------------------------------

Topic 0:
Top words: trump, clinton, said, republican, campaign, hillary, president, party, election
Usage in Fake News: 0.184
Usage in True News: 0.085
Bias (+ = fake, - = true): 0.368

Topic 1:
Top words: trump, people, one, president, like, donald, twitter, news, white
Usage in Fake News: 0.386
Usage in True News: 0.024
Bias (+ = fake, - = true): 0.881

Topic 2:
Top words: said, house, bill, would, senate, republicans, republican, may, president
Usage in Fake News: 0.033
Usage in True News: 0.109
Bias (+ = fake, - = true): -0.538

Topic 3:
Top words: said, trump, court, president, house, state, former, clinton, department
Usage in Fake News: 0.102
Usage in True News: 0.154
Bias (+ = fake, - = true): -0.202

Topic 4:
Top words: police, gun, said, shooting, old, media, man, officers, year
Usage in Fake News: 0.088
Usage in True News: 0.023
Bias (+ = fake, - = true): 0.578
```

```
Topic 5:
Top words: said, party, government, state, election, would, political, coalition, reuters
Usage in Fake News: 0.029
Usage in True News: 0.106
Bias (+ = fake, - = true): -0.569

Topic 6:
Top words: said, trump, united, president, states, north, korea, russia, iran
Usage in Fake News: 0.036
Usage in True News: 0.188
Bias (+ = fake, - = true): -0.677

Topic 7:
Top words: said, china, chinese, national, taiwan, cuba, beijing, obama, president
Usage in Fake News: 0.026
Usage in True News: 0.032
Bias (+ = fake, - = true): -0.111

Topic 8:
Top words: said, government, reuters, people, security, state, military, killed, police
Usage in Fake News: 0.030
Usage in True News: 0.127
Bias (+ = fake, - = true): -0.613

Topic 9:
Top words: said, would, tax, new, percent, trump, year, million, states
Usage in Fake News: 0.086
Usage in True News: 0.150
Bias (+ = fake, - = true): -0.271
```

```
2. Topic Coherence and Distinctiveness:
------------------------------------------------

Topic 0:
Coherence score: 677.732
Average word overlap: 2.3 words

Topic 1:
Coherence score: 458.280
Average word overlap: 1.1 words

Topic 2:
Coherence score: 633.213
Average word overlap: 1.9 words

Topic 3:
Coherence score: 510.875
Average word overlap: 2.3 words

Topic 4:
Coherence score: 370.581
Average word overlap: 1.1 words

Topic 5:
Coherence score: 396.233
Average word overlap: 1.8 words

Topic 6:
Coherence score: 635.317
Average word overlap: 2.0 words

Topic 7:
Coherence score: 446.216
Average word overlap: 1.4 words
```

The topics represent real-world topics well because: each topic captures distinct, interpretable themes, the topics show strong coherence (all scores > 370), low word overlap between topics (1.1-2.3 words on average), reflect known characteristics of true news (focus on policy, international affairs) and fake news (focus on sensation, social media)

2. Randomly select 5 real news examples and 5 fake news examples, and examine the topic distributions for each document. Which topics are prevalent in the real news

documents? (One sentence) Which topics are prevalent in the fake news documents? (One sentence)

```
Analysis of 5 Selected Documents from Each Category:

Fake News Examples:
--------------------------------------------------

Fake Document 1:
Preprocessed terms: ability achieve activist activists adjourned administration adviser along although america...
Topic distribution:
Topic 0: 0.802
Topic 1: 0.000
Topic 2: 0.053
Topic 3: 0.000
Topic 4: 0.000
Topic 5: 0.000
Topic 6: 0.000
Topic 7: 0.000
Topic 8: 0.142
Topic 9: 0.000
--------------------------------------------------

Fake Document 2:
Preprocessed terms: 2016it aboard accounts acting advanced advisor afternoon also anytime apparently...
Topic distribution:
Topic 0: 0.345
Topic 1: 0.558
Topic 2: 0.000
Topic 3: 0.000
Topic 4: 0.000
Topic 5: 0.000
Topic 6: 0.094
Topic 7: 0.000
Topic 8: 0.000
Topic 9: 0.000
--------------------------------------------------

Fake Document 3:
Preprocessed terms: abuses accent according add allowed already ambushed american answer appalled...
Topic distribution:
Topic 0: 0.461
Topic 1: 0.335
Topic 2: 0.001
Topic 3: 0.063
Topic 4: 0.001
Topic 5: 0.001
Fake Document 4:
Preprocessed terms: 2017finally 2017it 2017pence 2017vice according adults alarming alone american americans...
Topic distribution:
Topic 0: 0.036
Topic 1: 0.194
Topic 2: 0.000
Topic 3: 0.000
Topic 4: 0.000
Topic 5: 0.000
Topic 6: 0.000
Topic 7: 0.768
Topic 8: 0.000
Topic 9: 0.000
--------------------------------------------------

Fake Document 5:
Preprocessed terms: action alleges also amazon americans anti case categorized center ceo...
Topic distribution:
Topic 0: 0.001
Topic 1: 0.298
Topic 2: 0.001
Topic 3: 0.062
Topic 4: 0.482
```

In the fake news documents, political content (Topic 0) and social media content (Topic 1) are most prevalent.

```
True Document 1:
Preprocessed terms: acting agency allegations allowed among announced answer anti arrested assembly...
Topic distribution:
Topic 0: 0.001
Topic 1: 0.001
Topic 2: 0.001
Topic 3: 0.238
Topic 4: 0.001
Topic 5: 0.704
Topic 6: 0.001
Topic 7: 0.001
Topic 8: 0.054
Topic 9: 0.001
--------------------------------------------------

True Document 2:
Preprocessed terms: abdication abortion added administration advance aerospace affect agricultural ahead aimed...
Topic distribution:
Topic 0: 0.298
Topic 1: 0.000
Topic 2: 0.186
Topic 3: 0.507
Topic 4: 0.008
Topic 5: 0.000
Topic 6: 0.000
Topic 7: 0.000
Topic 8: 0.000
Topic 9: 0.000

True Document 3:
Preprocessed terms: abuses administration appear asian authorities bans burma called committed countries...
Topic distribution:
Topic 0: 0.001
Topic 1: 0.001
Topic 2: 0.077
Topic 3: 0.129
Topic 4: 0.001
Topic 5: 0.001
Topic 6: 0.185
Topic 7: 0.001
Topic 8: 0.601
Topic 9: 0.001
--------------------------------------------------

True Document 4:
Preprocessed terms: abroad accord action added african ago agreed aimed aligned amendments...
Topic distribution:
Topic 0: 0.000
Topic 1: 0.000
Topic 2: 0.103
Topic 3: 0.000
Topic 4: 0.000
Topic 5: 0.206
Topic 6: 0.688
Topic 7: 0.000
Topic 8: 0.000
Topic 9: 0.000
--------------------------------------------------

True Document 5:
Preprocessed terms: asked audit audits bid called came campaign candidate cbs clarify...
Topic distribution:
Topic 0: 0.382
Topic 1: 0.001
Topic 2: 0.001
Topic 3: 0.525
Topic 4: 0.001
Topic 5: 0.001
Topic 6: 0.001
Topic 7: 0.001
Topic 8: 0.001
Topic 9: 0.089
```

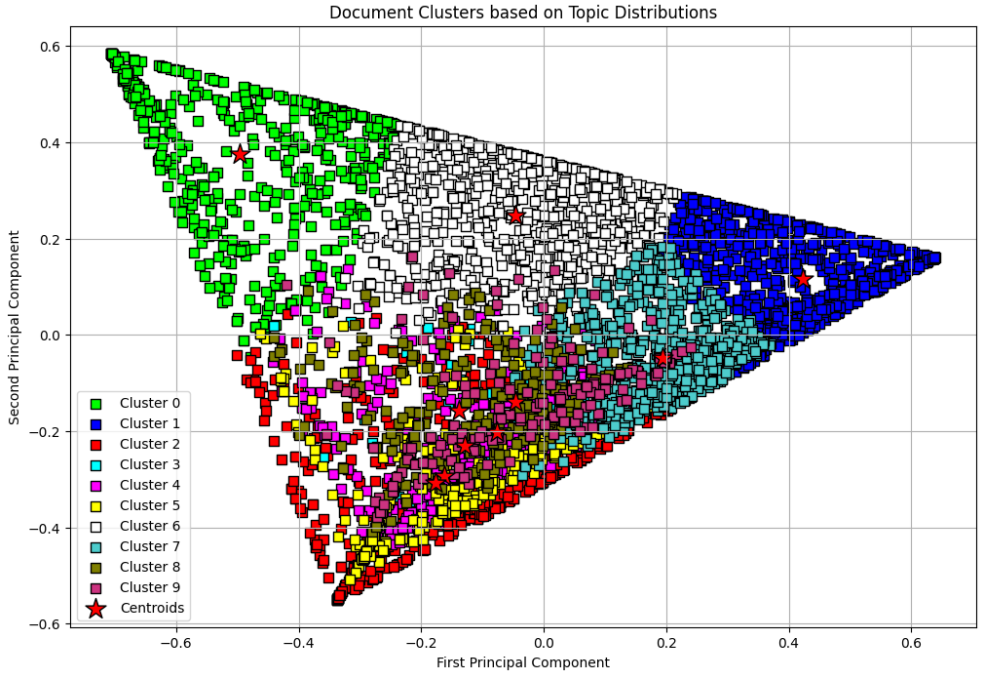
In the true news documents, they contain diverse topic distribution than fake news. There is a focus on institutional/governmental topics, it also has multiple significant topics per document.

3. Use the LDA vectors for the documents as features in a Logistic Regression classifier to predict whether each document is real news or fake news. According to the resulting coefficients from the regression, which topics are most useful in determining whether something is real news or fake news? (One sentence)

```
--------------------------------------------------------------------------------
Topic 1: -10.345 (indicates FAKE news)
Top words: trump, people, one, president, like, donald, twitter, news, white

Topic 4: -4.181 (indicates FAKE news)
Top words: police, gun, said, shooting, old, media, man, officers, year

Topic 6: 3.499 (indicates TRUE news)
Top words: said, trump, united, president, states, north, korea, russia, iran

Topic 2: 3.252 (indicates TRUE news)
Top words: said, house, bill, would, senate, republicans, republican, may, president

Topic 8: 3.236 (indicates TRUE news)
Top words: said, government, reuters, people, security, state, military, killed, police

Topic 5: 2.637 (indicates TRUE news)
Top words: said, party, government, state, election, would, political, coalition, reuters

Topic 3: 1.325 (indicates TRUE news)
Top words: said, trump, court, president, house, state, former, clinton, department

Topic 9: 1.221 (indicates TRUE news)
Top words: said, would, tax, new, percent, trump, year, million, states

Topic 0: -0.753 (indicates FAKE news)
Top words: trump, clinton, said, republican, campaign, hillary, president, party, election

Topic 7: 0.346 (indicates TRUE news)
Top words: said, china, chinese, national, taiwan, cuba, beijing, obama, president
```

Social media(-10.345) and sensational content (-4.181) are strong predictors of fake news. International affairs (3.499) and legislative content (3.252) strongly indicate true news.

4. Pick real news or fake news, whichever is more interesting to you. Then, use the LDA vectors for those news documents to cluster them. You can use KMeans clustering with a reasonable value for K (if you don't have strong feelings for a particular K, I recommend 10). Then, select 5 news documents from each resulting cluster. Do the clusters correspond to anything? (One sentence)




```
Cluster 0:
--------------------------------------------------

Fake Document 609
Title:  These Hilarious Donald Trump Impressions Prove That Other Candidates Hate Him Just As Much As We Do (VIDEOS)
Text preview: republican front runner donald trump does pretty amazing job making himself look like complete moron but given much material that impersonati
Top topics:
Topic 0: 0.704
Topic 1: 0.293
Topic 2: 0.000
------------------------------

Fake Document 1733
Title: MEET LEFTIST ANALYSTS At Research Firm Who Created ABC/Washington Post Poll Showing Hillary With 12 Point Lead [VIDEO]
Text preview: many americans were shocked the latest abc washington post goal seeking report aka poll that shows hillary opening point lead with likely vot
Top topics:
Topic 0: 0.466
Topic 9: 0.178
Topic 6: 0.145
------------------------------

Fake Document 3870
Title:  Bernie Sanders Open To Being Clinton's V.P. (VIDEO)
Text preview: hillary clinton the path the democratic nominee for president the united states the convention nears clinton and sanders are still battling f
Top topics:
Topic 0: 0.950
Topic 6: 0.047
Topic 9: 0.000
```

```
Cluster 1:
--------------------------------------------------

Fake Document 3500
Title:  EXPLOSIVE REPORT: Fmr. NSA Analyst Reveals Team Trump Working With Russians To Punish Reporters
Text preview: all know how hostile the trump white house the press they say that anything they don like fake news and they have even gone far ban critical
Top topics:
Topic 1: 0.650
Topic 3: 0.174
Topic 0: 0.145
------------------------------

Fake Document 3766
Title:  Leave It To Sesame Street To Teach America What A Jerk Donald Trump Truly Is (VIDEO)
Text preview: for everyone other than maybe his investment partners his wife and his white nationalist supporters most people look donald trump with utter
Top topics:
Topic 1: 0.984
Topic 9: 0.012
Topic 0: 0.000
------------------------------

Fake Document 3471
Title:  National Security Expert Warns Of The DIRE Dangers Of Trump's Plan For Intel Community (VIDEO)
Text preview: for the first time our lives are entering era where have elected president who downright dangerous donald trump petty childish ignorant thin
Top topics:
Topic 1: 0.848
Topic 3: 0.149
Topic 4: 0.000
```

```
Cluster 2:
--------------------------------------------------

Fake Document 4358
Title: YOU'RE FIRED! WHY THE WHITE HOUSE Just Fired A Senior National Security Aide
Text preview: national security council aide craig deare was dismissed friday after was learned that harshly criticized the president and his top aides dea
Top topics:
Topic 3: 0.560
Topic 6: 0.198
Topic 8: 0.129
------------------------------

Fake Document 3571
Title:  Donald Trump Is An Illegitimate President-Elect Because Russia Helped Him Steal Election
Text preview: russian interference our election should immediately disqualify donald trump from the presidency friday cia assessment found that the consens
Top topics:
Topic 0: 0.464
Topic 3: 0.442
Topic 1: 0.093
------------------------------

Fake Document 4878
Title: BOOM! JOHN SUNUNU: "Bothers Me That Mueller Is Hiring "Blatantly Biased Lawyers" [Video]
Text preview: former governor john sununu let alison camarata have when she asked about the russia investigation and robert mueller sununu said that trump
Top topics:
Topic 3: 0.741
Topic 1: 0.145
```

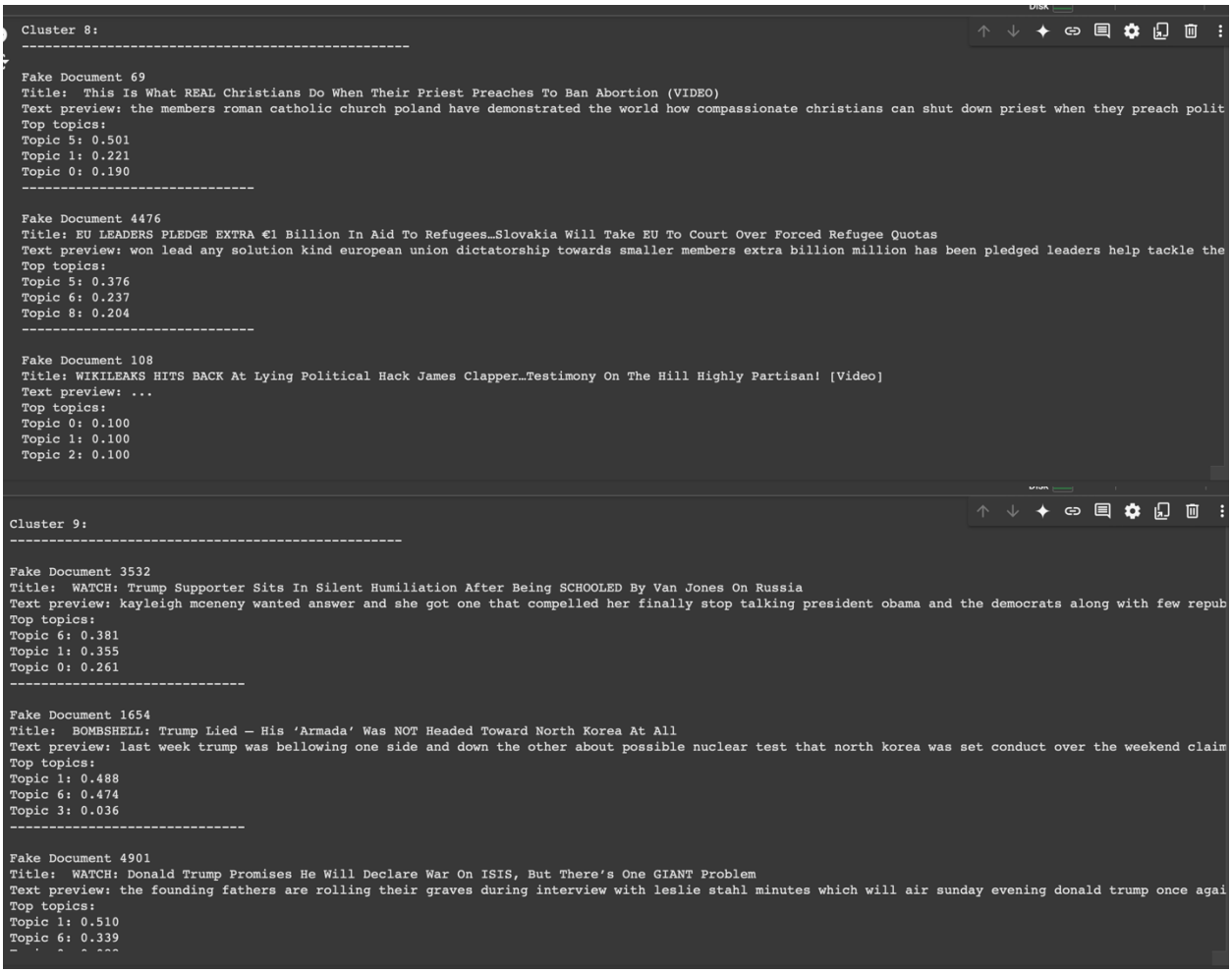
```
Cluster 3:
--------------------------------------------------

Fake Document 162
Title: McCain's Mad World and The Cancer of Conflict
Text preview: 21st century wire says some devastating news befell john sidney mccain iii recently his staff announced that the senator had been diagnosed w
Top topics:
Topic 6: 0.237
Topic 4: 0.231
Topic 8: 0.216
------------------------------

Fake Document 602
Title: HUGE SECURITY LAPSE: International Flight Passengers Skip Customs
Text preview: skipping customs jfk might seem like great thing for passengers arriving international flight but those passengers might putting all american
Top topics:
Topic 8: 0.558
Topic 1: 0.379
Topic 3: 0.061
------------------------------

Fake Document 1135
Title: WHILE OBAMA TRIES TO DISARM AMERICANS, Israeli Ministers Encourage Citizens To Carry To Eliminate Enemy
Text preview: the destruction america full swing pay billions deliver unchecked muslims refugees the shores america meanwhile real men israel are arming de
Top topics:
Topic 8: 0.416
Topic 6: 0.257
```

```
Cluster 4:
--------------------------------------------------

Fake Document 505
Title:  Target Just Made A Major Move That's Going To Blow Conservative Minds (VIDEO)
Text preview: conservatives who are notoriously anti labor are going outraged the move target just made target will raising their national minimum wage sta
Top topics:
Topic 9: 0.669
Topic 1: 0.257
Topic 7: 0.058
------------------------------

Fake Document 982
Title: EXPLODING AFRICAN REFUGEE POPULATION STRESSING WELFARE SYSTEM IN MINNESOTA Are Sending Millions Of Dollars Back To Africa
Text preview: not just the our open southern borders need concerned about our state department who seems hell bent populating our states with refugees from
Top topics:
Topic 9: 0.528
Topic 1: 0.317
Topic 8: 0.152
------------------------------

Fake Document 4296
Title: "FAIR SHARE" FAIL: Trump's Taxes Show He Paid Almost TWICE The Rate Of Socialist Bernie Sanders Who Owns 3 Homes, Preaches Equality [VIDEO]
Text preview: watch tucker carlson point out that trump paid tax rate while socialist fair share bernie only paid tax rate...
Top topics:
Topic 9: 0.468
Topic 1: 0.331
```

```
Cluster 5:
--------------------------------------------------

Fake Document 1244
Title: WATCH: TWO TX SCHOOL WORKERS FIRED For Refusing To Call 6 Year Old Girl A "Boy" : "One day, she wanted to be a girl, the next day she wanted to be a
Text preview: because year old girls are consumed with their sexual identities right the parents should getting regular visits social services and the fire
Top topics:
Topic 4: 0.551
Topic 9: 0.239
Topic 1: 0.205
------------------------------

Fake Document 1623
Title:  Court Just Gave Cops Permission To Murder Dogs
Text preview: the 6th district court ohio ruled monday that cop comes your home they are justified killing your dog your dog does much move mark and cheryl
Top topics:
Topic 4: 0.621
Topic 1: 0.172
Topic 5: 0.126
------------------------------

Fake Document 1282
Title: TRANSGENDER TARGET SUING Good Samaritan Who Saved Girl From Being Stabbed To Death [VIDEO]
Text preview: the timing couldn possibly any worse warning graphic videotarget smack dab the middle nightmare the news target suing this good samaritan cou
Top topics:
Topic 4: 0.487
Topic 1: 0.293
```

```
+ Text                                                    RAM Disk   Gemini

Cluster 6:
--------------------------------------------------

Fake Document 2642
Title:  70% Of Republican High Rollers Want Trump Out Of Their Party — And The 2016 Race
Text preview: donald trump has been absolute nightmare for the republican party establishment the last week has been especially horrible from his unbelieva
Top topics:
Topic 0: 0.624
Topic 1: 0.372
Topic 3: 0.000
------------------------------

Fake Document 315
Title:  Trump Explodes In Rage At Debate Results, Indicates He WON'T Accept An Election Loss (DETAILS)
Text preview: all know that donald trump skin very thin also narcissist the likes which the world has never seen the word leadership stage therefore the th
Top topics:
Topic 0: 0.569
Topic 1: 0.427
Topic 9: 0.000
------------------------------

Fake Document 338
Title: SOCIALIST Bernie Sanders Asks Trump's Pick For Education Sec If She'll Agree To FREE College…Gets Embarrassing Public SMACK DOWN [VIDEO]
Text preview: besty devos trump conservative choice for education secretary during her confirmation process leftist bernie sanders most important question
Top topics:
Topic 0: 0.350
Topic 1: 0.299
Topic 7: 0.216
```

```
Disk

Cluster 7:
--------------------------------------------------

Fake Document 677
Title: CONFUSED PROTESTERS Swarm Outside Trump NYC Fundraiser: 'Tax the Rich, Not Working People'
Text preview: the protesters nyc must confused they yelled tax the rich not working people aren the rich working people too many the rich worked like crazy
Top topics:
Topic 1: 0.431
Topic 9: 0.376
Topic 4: 0.188
------------------------------

Fake Document 2480
Title:  Why Doctor Who Made Bizarre Race Rant Against Michelle Obama Can't Be Fired
Text preview: denver colorado doctor under fire for bizarre racial rant about first lady michelle obama will probably keep her job anyway michelle herren w
Top topics:
Topic 1: 0.614
Topic 2: 0.296
Topic 3: 0.049
------------------------------

Fake Document 2393
Title: FIRE THIS FOX NEWS ANALYST: Trump Was "Allegedly Elected" [Video]
Text preview: our previous reports jehmu greene will never understand why fox news has crazy liberals like jehmu greene contribute any discussion she the r
Top topics:
Topic 1: 0.548
Topic 0: 0.156
Topic 4: 0.148
```



K=10, KMeans, choose fake news:
Cluster 0:
Dominated by Topic 0 (campaign/election)
Heavy focus on Trump, Clinton, and election coverage
Cluster 1:
Dominated by Topic 1 (social media/Trump)
Very high Topic 1 scores (0.650-0.984)
Critical/negative coverage of Trump
Cluster 2:
High Topic 3 scores (legal/court)
investigations, legal cases. Many Russia investigation related.
Cluster 3:
Mix of Topics 8 (security), 4 (incidents), and 6 (international)
Focus on national security, military, and international threats
Cluster 4:
High Topic 9 scores (economic/tax)
about welfare, education, economics

Cluster 5:
Dominated by Topic 4 (police/incidents)
Sensational crime stories

- Your answers to the usual questions:
  - How long did this assignment take you? (1 sentence)
    2 days

  - Whom did you work with, and how? (1 sentence each)
    - Discussing the assignment with others is encouraged, as long as you don't share the code.
      On my own

  - Which resources did you use? (1 sentence each)
    - For each, please list the URL and a brief description of how it was useful.
      Data source: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
      Course code:
      https://colab.research.google.com/drive/1q8Cf9QUwd45aWkOyPVjljOVb2N4xdUj5?usp=sharing#scrollTo=-V0J04mIOc5m

      https://colab.research.google.com/drive/1paraIRBL87aYdG9dtR6x2fs0P4QBEmHe?usp=sharing#scrollTo=Wi6X76BoaEN9

      LatentDirichletAllocation:
      https://scikitlearn.org/1.5/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

      CountVectorizer:
      https://scikitlearn.org/1.5/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

  - A few sentences about:
    - What was the most difficult part of the assignment?
      Understand the logic of topic modeling:
      Raw Text → CountVectorizer → Word Counts → LDA → Topic Distributions

```
# CountVectorizer: Converts text to word counts
vec = CountVectorizer(stop_words=['the', 'a', ...])
X = vec.fit_transform(df['document'])

# LatentDirichletAllocation: Finds topics in word counts
lda = LatentDirichletAllocation(n_components=10)
doc_topics = lda.fit_transform(X)
```

- What was the most rewarding part of the assignment?

  Practice using cluster and LDA model

- What did you learn doing the assignment?
  How to handle the when run time is too long (limit the number of documents by random selections)
- Constructive and actionable suggestions for improving assignments, office hours, and class time are always welcome.