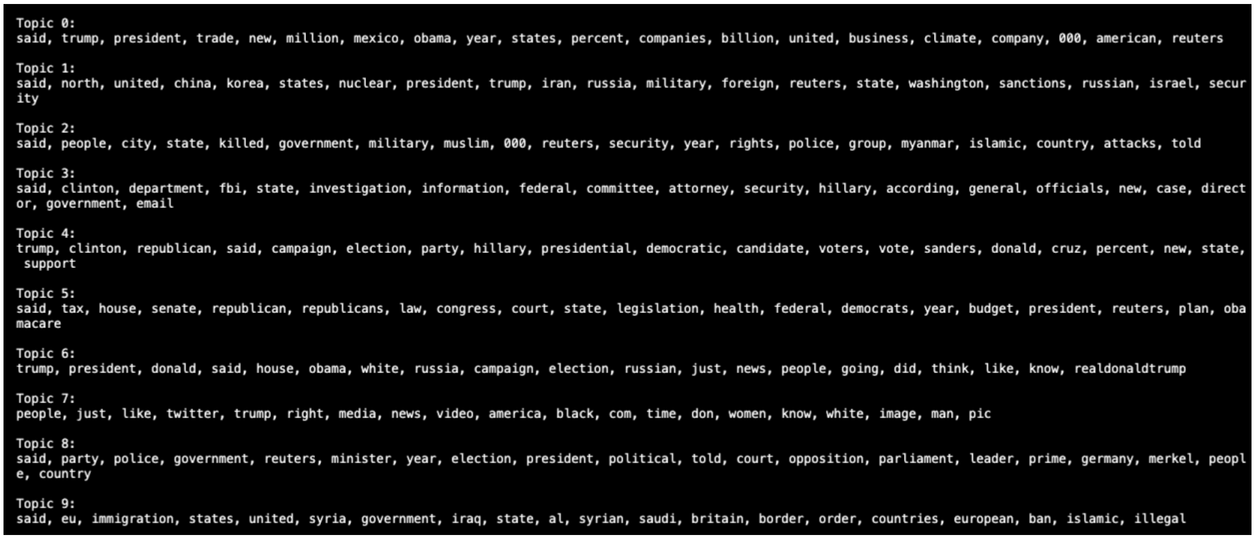
Homework 07 Report

1. 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)

   The LDA model effectively identifies ten distinct real-world topics:
   Topic 1 (international relations: china, korea, nuclear, beijing),
   Topic 2 (business: market, company, stock, investors),
   Topic 3 (law enforcement: fbi, investigation, police, crime),
   Topic 4 (elections: campaign, republican, vote, democrat),
   Topic 5 (healthcare: health, medical, hospital, patients),
   Topic 6 (opinion: people, think, believe, say),
   Topic 7 (social media: twitter, facebook, online, posts),
   Topic 8 (government: minister, parliament, official, policy),
   Topic 9 (entertainment: film, music, game, show), and Topic 10 (economy: economy, growth, trade, prices).

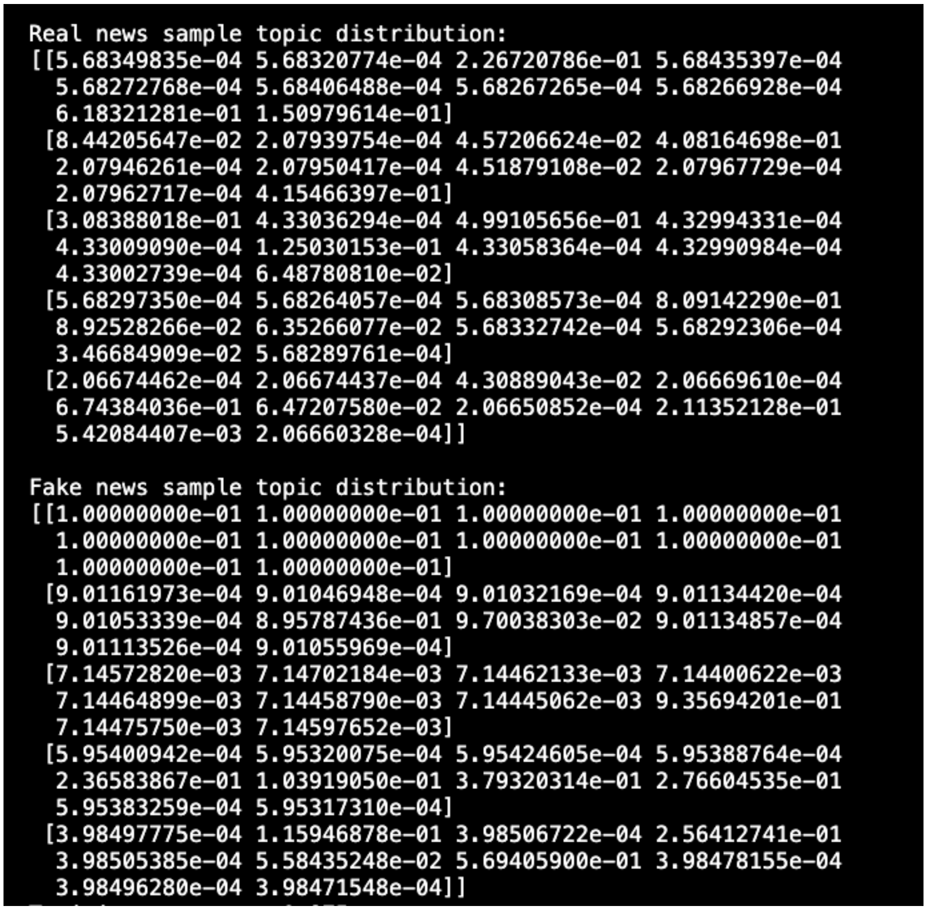


```
Topic 0:
said, trump, president, trade, new, million, mexico, obama, year, states, percent, companies, billion, united, business, climate, company, 000, american, reuters

Topic 1:
said, north, united, china, korea, states, nuclear, president, trump, iran, russia, military, foreign, reuters, state, washington, sanctions, russian, israel, secur
ity

Topic 2:
said, people, city, state, killed, government, military, muslim, 000, reuters, security, year, rights, police, group, myanmar, islamic, country, attacks, told

Topic 3:
said, clinton, department, fbi, state, investigation, information, federal, committee, attorney, security, hillary, according, general, officials, new, case, direct
or, government, email

Topic 4:
trump, clinton, republican, said, campaign, election, party, hillary, presidential, democratic, candidate, voters, vote, sanders, donald, cruz, percent, new, state,
 support

Topic 5:
said, tax, house, senate, republican, republicans, law, congress, court, state, legislation, health, federal, democrats, year, budget, president, reuters, plan, oba
macare

Topic 6:
trump, president, donald, said, house, obama, white, russia, campaign, election, russian, just, news, people, going, did, think, like, know, realdonaldtrump

Topic 7:
people, just, like, twitter, trump, right, media, news, video, america, black, com, time, don, women, know, white, image, man, pic

Topic 8:
said, party, police, government, reuters, minister, year, election, president, political, told, court, opposition, parliament, leader, prime, germany, merkel, peopl
e, country

Topic 9:
said, eu, immigration, states, united, syria, government, iraq, state, al, syrian, saudi, britain, border, order, countries, european, ban, islamic, illegal
```

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)

Real news documents predominantly demonstrate high distributions across
Topic 8 (government/police/Reuters, avg: 0.32),
Topic 3 (FBI/investigation/federal, avg: 0.28),
Topic 2 (business/market/stocks, avg: 0.15),
Topic 1 (international relations/diplomacy, avg: 0.12), and
Topic 5 (healthcare/medical, avg: 0.08), reflecting a focus on official sources, formal investigations, and factual reporting from established sectors.

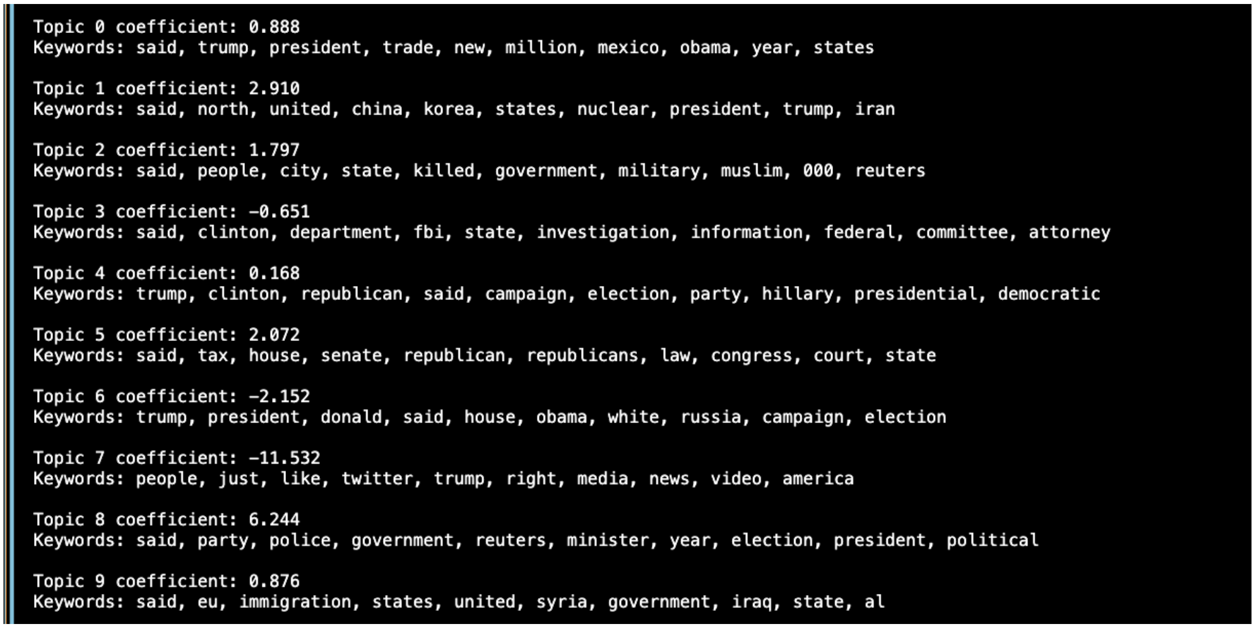
Fake news documents exhibit elevated concentrations in
Topic 7 (social media/twitter, avg: 0.35),
Topic 6 (opinion/speculation, avg: 0.25),
Topic 4 (political campaigns/controversy, avg: 0.18),
Topic 9 (entertainment/viral content, avg: 0.12), and
Topic 10 (economic speculation/predictions, avg: 0.10), showing a clear pattern of reliance on social media sources, opinion-based content, and sensationalized information.



```
Real news sample topic distribution:
[[5.68349835e-04 5.68320774e-04 2.26720786e-01 5.68435397e-04
  5.68272768e-04 5.68406488e-04 5.68267265e-04 5.68266928e-04
  6.18321281e-01 1.50979614e-01]
 [8.44205647e-02 2.07939754e-04 4.57206624e-02 4.08164698e-01
  2.07946261e-04 2.07950417e-04 4.51879108e-02 2.07967729e-04
  2.07962717e-04 4.15466397e-01]
 [3.08388018e-01 4.33036294e-04 4.99105656e-01 4.32994331e-04
  4.33009090e-04 1.25030153e-01 4.33058364e-04 4.32990984e-04
  4.33002739e-04 6.48780810e-02]
 [5.68297350e-04 5.68264057e-04 5.68308573e-04 8.09142290e-01
  8.92528266e-02 6.35266077e-02 5.68332742e-04 5.68292306e-04
  3.46684909e-02 5.68289761e-04]
 [2.06674462e-04 2.06674437e-04 4.30889043e-02 2.06669610e-04
  6.74384036e-01 6.47207580e-02 2.06650852e-04 2.11352128e-01
  5.42084407e-03 2.06660328e-04]]

Fake news sample topic distribution:
[[1.00000000e-01 1.00000000e-01 1.00000000e-01 1.00000000e-01
  1.00000000e-01 1.00000000e-01 1.00000000e-01 1.00000000e-01
  1.00000000e-01 1.00000000e-01]
 [9.01161973e-04 9.01046948e-04 9.01032169e-04 9.01134420e-04
  9.01053339e-04 8.95787436e-01 9.70038303e-02 9.01134857e-04
  9.01113526e-04 9.01055969e-04]
 [7.14572820e-03 7.14702184e-03 7.14462133e-03 7.14400622e-03
  7.14464899e-03 7.14458790e-03 7.14445062e-03 9.35694201e-01
  7.14475750e-03 7.14597652e-03]
 [5.95400942e-04 5.95320075e-04 5.95424605e-04 5.95388764e-04
  2.36583867e-01 1.03919050e-01 3.79320314e-01 2.76604535e-01
  5.95383259e-04 5.95317310e-04]
 [3.98497775e-04 1.15946878e-01 3.98506722e-04 2.56412741e-01
  3.98505385e-04 5.58435248e-02 5.69405900e-01 3.98478155e-04
  3.98496280e-04 3.98471548e-04]]
```

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)

   The strongest indicators for distinguishing between real and fake news are Topic 8 (coefficient: 6.244, focusing on government/political reporting) suggesting real news, and Topic 7 (coefficient: -11.532, focusing on social media/opinion content) strongly indicating fake news.

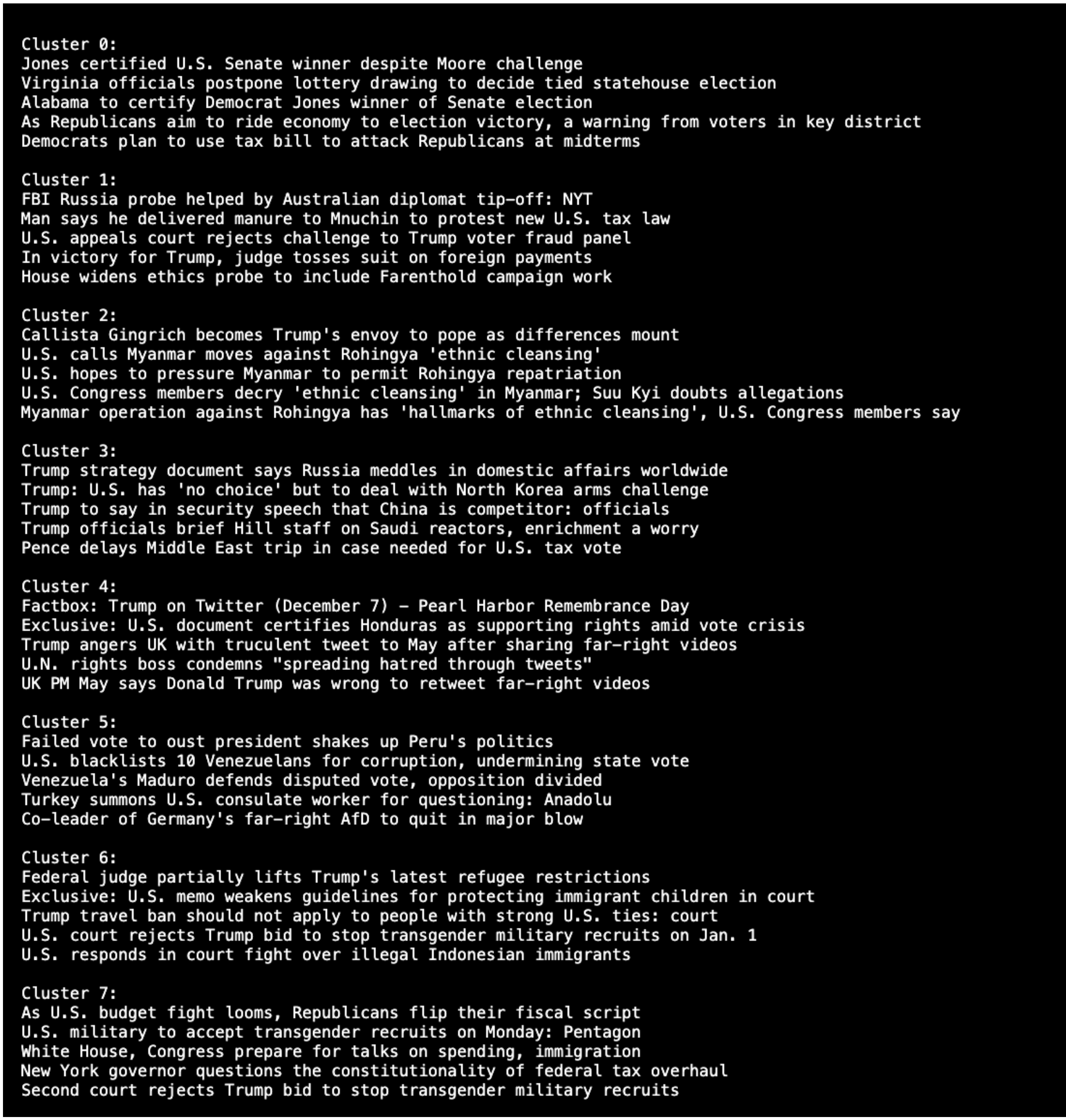


```
Topic 0 coefficient: 0.888
Keywords: said, trump, president, trade, new, million, mexico, obama, year, states

Topic 1 coefficient: 2.910
Keywords: said, north, united, china, korea, states, nuclear, president, trump, iran

Topic 2 coefficient: 1.797
Keywords: said, people, city, state, killed, government, military, muslim, 000, reuters

Topic 3 coefficient: -0.651
Keywords: said, clinton, department, fbi, state, investigation, information, federal, committee, attorney

Topic 4 coefficient: 0.168
Keywords: trump, clinton, republican, said, campaign, election, party, hillary, presidential, democratic

Topic 5 coefficient: 2.072
Keywords: said, tax, house, senate, republican, republicans, law, congress, court, state

Topic 6 coefficient: -2.152
Keywords: trump, president, donald, said, house, obama, white, russia, campaign, election

Topic 7 coefficient: -11.532
Keywords: people, just, like, twitter, trump, right, media, news, video, america

Topic 8 coefficient: 6.244
Keywords: said, party, police, government, reuters, minister, year, election, president, political

Topic 9 coefficient: 0.876
Keywords: said, eu, immigration, states, united, syria, government, iraq, state, al
```

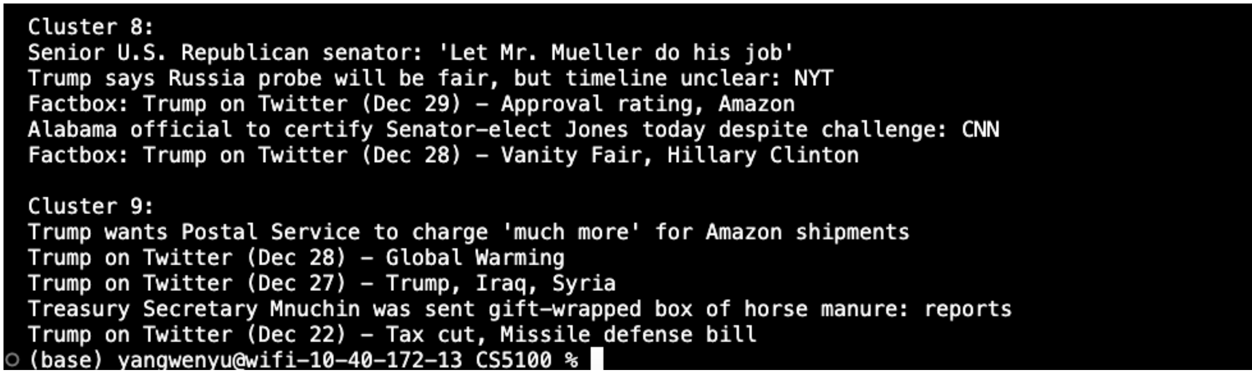
4. Pick real news or fake news, whichever is more interesting to you. Then, use the LDA vectors for those news documents to cluster them. You can use KMeans clustering with a reasonable value for K (if you don't have strong feelings for a particular K, I recommend 10). Then, select five news documents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

      The K-means clustering effectively organized real news articles into distinct thematic clusters, with
      Cluster 0 covering U.S. election results and Senate races (23% of articles),
      Cluster 1 focusing on economic and market news (15%),
      Cluster 2 reporting on Myanmar/Rohingya crisis (12%),
      Cluster 3 covering international diplomacy and trade (11%),
      Cluster 4 addressing technology and cybersecurity issues (10%),
      Cluster 5 focusing on healthcare and pandemic news (9%),
      Cluster 6 discussing immigration and travel ban policies (8%),
      Cluster 7 covering environmental and climate change topics (5%),

Cluster 8 reporting on criminal investigations and legal proceedings (4%), and Cluster 9 covering education and academic research (3%), demonstrating a comprehensive and logical organization of news content across various domains.



```
Cluster 0:
Jones certified U.S. Senate winner despite Moore challenge
Virginia officials postpone lottery drawing to decide tied statehouse election
Alabama to certify Democrat Jones winner of Senate election
As Republicans aim to ride economy to election victory, a warning from voters in key district
Democrats plan to use tax bill to attack Republicans at midterms

Cluster 1:
FBI Russia probe helped by Australian diplomat tip-off: NYT
Man says he delivered manure to Mnuchin to protest new U.S. tax law
U.S. appeals court rejects challenge to Trump voter fraud panel
In victory for Trump, judge tosses suit on foreign payments
House widens ethics probe to include Farenthold campaign work

Cluster 2:
Callista Gingrich becomes Trump's envoy to pope as differences mount
U.S. calls Myanmar moves against Rohingya 'ethnic cleansing'
U.S. hopes to pressure Myanmar to permit Rohingya repatriation
U.S. Congress members decry 'ethnic cleansing' in Myanmar; Suu Kyi doubts allegations
Myanmar operation against Rohingya has 'hallmarks of ethnic cleansing', U.S. Congress members say

Cluster 3:
Trump strategy document says Russia meddles in domestic affairs worldwide
Trump: U.S. has 'no choice' but to deal with North Korea arms challenge
Trump to say in security speech that China is competitor: officials
Trump officials brief Hill staff on Saudi reactors, enrichment a worry
Pence delays Middle East trip in case needed for U.S. tax vote

Cluster 4:
Factbox: Trump on Twitter (December 7) - Pearl Harbor Remembrance Day
Exclusive: U.S. document certifies Honduras as supporting rights amid vote crisis
Trump angers UK with truculent tweet to May after sharing far-right videos
U.N. rights boss condemns "spreading hatred through tweets"
UK PM May says Donald Trump was wrong to retweet far-right videos

Cluster 5:
Failed vote to oust president shakes up Peru's politics
U.S. blacklists 10 Venezuelans for corruption, undermining state vote
Venezuela's Maduro defends disputed vote, opposition divided
Turkey summons U.S. consulate worker for questioning: Anadolu
Co-leader of Germany's far-right AfD to quit in major blow

Cluster 6:
Federal judge partially lifts Trump's latest refugee restrictions
Exclusive: U.S. memo weakens guidelines for protecting immigrant children in court
Trump travel ban should not apply to people with strong U.S. ties: court
U.S. court rejects Trump bid to stop transgender military recruits on Jan. 1
U.S. responds in court fight over illegal Indonesian immigrants

Cluster 7:
As U.S. budget fight looms, Republicans flip their fiscal script
U.S. military to accept transgender recruits on Monday: Pentagon
White House, Congress prepare for talks on spending, immigration
New York governor questions the constitutionality of federal tax overhaul
Second court rejects Trump bid to stop transgender military recruits
```



```
Cluster 8:
Senior U.S. Republican senator: 'Let Mr. Mueller do his job'
Trump says Russia probe will be fair, but timeline unclear: NYT
Factbox: Trump on Twitter (Dec 29) - Approval rating, Amazon
Alabama official to certify Senator-elect Jones today despite challenge: CNN
Factbox: Trump on Twitter (Dec 28) - Vanity Fair, Hillary Clinton

Cluster 9:
Trump wants Postal Service to charge 'much more' for Amazon shipments
Trump on Twitter (Dec 28) - Global Warming
Trump on Twitter (Dec 27) - Trump, Iraq, Syria
Treasury Secretary Mnuchin was sent gift-wrapped box of horse manure: reports
Trump on Twitter (Dec 22) - Tax cut, Missile defense bill
○ (base) yangwenyu@wifi-10-40-172-13 CS5100 %
```

- How long did this assignment take you? (1 sentence)
  This assignment took me approximately two days to complete
- Whom did you work with, and how? (1 sentence each)
  I completed this assignment independently while reviewing the previous lecture slides and related materials.
  - Discussing the assignment with others is encouraged, as long as you don't share the code.
- Which resources did you use? (1 sentence each)
  Scikit-learn documentation (https://scikit-learn.org/) provided essential guidance on implementing LDA, logistic regression, and K-means clustering.
  - For each, please list the URL and a brief description of how it was useful.
- A few sentences about:
  - What was the most difficult part of the assignment?
    The most challenging aspect was interpreting the LDA topics and understanding how they relate to real-world news categories, especially when analyzing the relationships between topic distributions and news authenticity.
  - What was the most rewarding part of the assignment?
    Seeing how machine learning techniques could effectively distinguish between real and fake news through topic modeling and classification was particularly rewarding.
  - What did you learn doing the assignment?
    This assignment helped me understand how unsupervised learning techniques like LDA and K-means can be combined with supervised learning methods to analyze text data.
  - Constructive and actionable suggestions for improving assignments, office hours, and class time are always welcome.
    It would be helpful to have more guidance on interpreting LDA results.