1. 프로젝트 개요
   1. 주제 선정 배경

대립구조를 가지는 진보, 보수의 성향을 분석하기 위함.

* 1. 프로젝트 주제

정치적 성향에 따른 미국 진보 보수 간의 사용하는 어휘 차이점 분석

* 1. 프로젝트 진행 계획

진보, 보수의 성향을 띠는 언론사의 기사를 데이터로 사용하여 양간의 사용 어휘 의 차이점 및 경향을 분석할 것이다. 기사 데이터를 전처리하고, tf-idf와 LDA를 이용하여 topic modeling할 것이다. 그 결과로 나온 topic들을 언론사 별로 어떤 어휘를 사용하고 어떤 주제를 다루는지 분석하여 벤다이어그램으로 시각화할 것이다.

* 1. 역할 분담

Pre-processing: 라우라, 아이린, 허수민

Topic modeling: 김희선, 박채린, 이은주

* 1. 주간 계획표

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1st week | 2nd week | 3rd week | 4th week | 5th week |
| proposal |  |  |  |  |  |
| pre-processing |  |  |  |  |  |
| topic modeling |  |  |  |  |  |
| evaluation |  |  |  |  |  |
| analysis |  |  |  |  |  |
| final report |  |  |  |  |  |

* 1. 데이터 셋

Figure 1을 참고하여 liberal(Atlantic, Buzzfeed News, Vox), conservative(Breitbart, New York Post, National Review, Fox News)를 선정하였다.

(출처: <https://www.kaggle.com/snapcrack/all-the-news>)

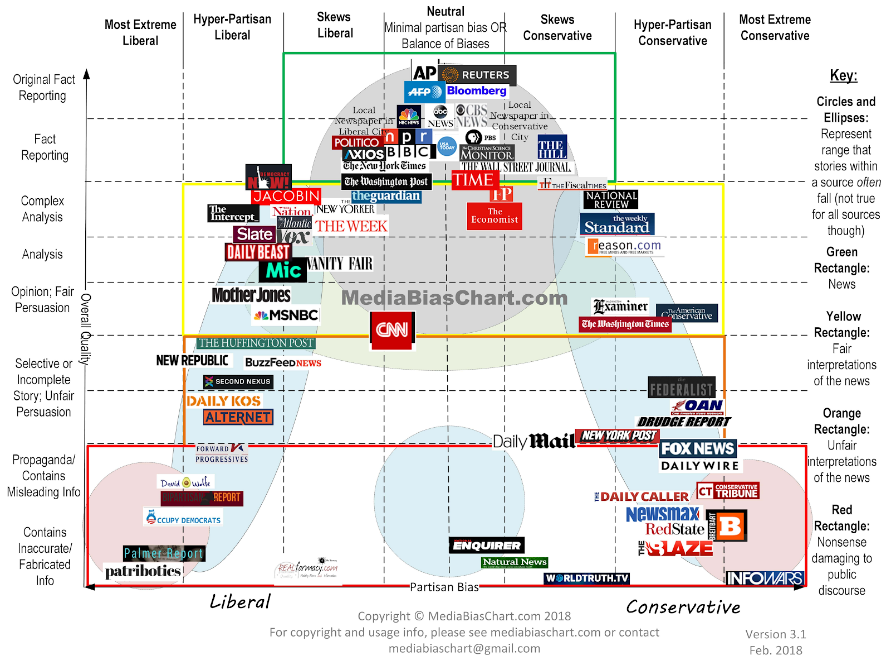


Figure 1. 미국 언론 진보, 보수 성향 차트

* 1. Tool

Language: python3, Anaconda

Ide: jupyter notebook

1. 프로젝트 진행 과정
   1. 데이터 전처리

* 진보, 보수 성향을 중심으로 기사 분할

원본 데이터 셋 “article1.csv”, “article2.csv”, “article3.csv”는 총 10개의 컬럼으로 구성되어 있다. 이 중 publication 컬럼 값을 Bias Chart를 참고해서 만든 진보, 봅수 언론사 목록과 비교하여 진보 16980개, 보수 51831개를 중 각각 16000개를 사용하여 “liberal16000.csv”, “conserv16000.csv”를 생성하였다.

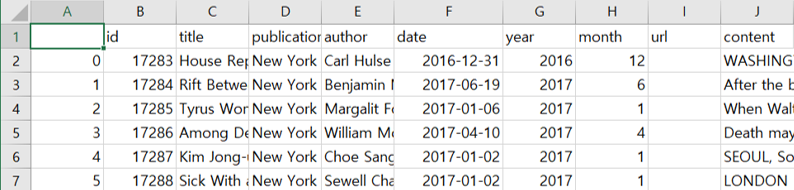


Figure 2. 원본 데이터셋 형식

* + - Text operation

“liberal16000.csv”, “conserv16000.csv”의 content column(기사내용)을 문서 별로 리스트에 저장하고 모두 소문자로 변경하였다. nltk.tokenize의 word\_tokenize, punctucation 함수를 사용해서 구두점을 제거하고 단어로 잘랐다. 인용부호와 “-”를 삭제하기 위해서 단어 형식을 [a-z0-9]+로 정규화한 후  nltk.corpus의 english stopwords를 제거하고 porterstemmer로 남은 단어들을 어근의 형태로 변경하였다. 변경된 단어 사이에 공백을, 하나의 기사가 끝나면 개행문자를 추가하여 “new\_liberal.txt”, “new\_conserv.txt” 파일로 저장하였다.

* + - 기자, 언론사 이름 및 SNS 계정 정보 삭제

기사의 하단부에 기사를 작성한 기자 및 언론사의 정보, 관련 SNS 계정 홍보문이 반복해서 언급되어 기사의 내용과는 관계가 없는 단어들이 전처리 결과의 많은 부분을 차지하였다. 이를 해결하기 위해서 원본 데이터셋의 author , publications 칼럼을 tokenize하고 해당 기자 이름 및 언론사가 기사 원문에서 재언급될 경우 삭제하였다. 또한 이름 및 언론사가 포함된 계정, “@”가 포함된 단어를 삭제하였다.(Figure 3)

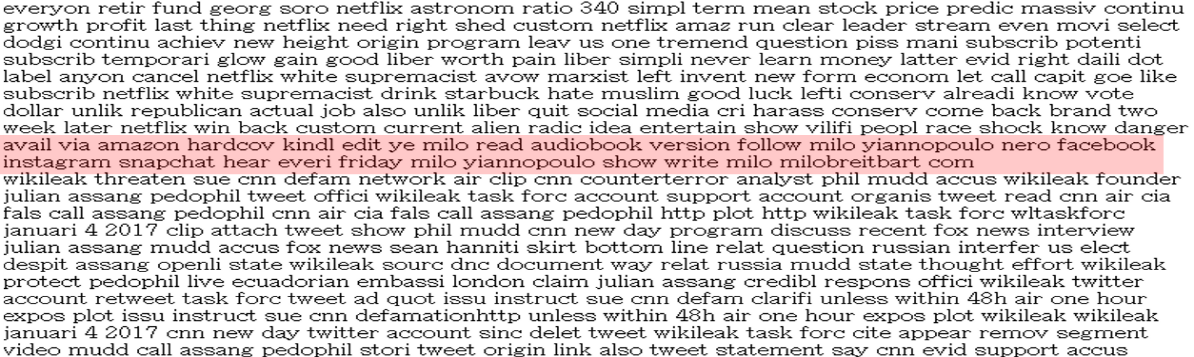


Figure 3. 전처리 결과 파일에 포함된 보수 언론사 Breitbart 기자 SNS홍보문

* + - TF-IDF에서 걸러지지 않는 단어 제거

LDA 모델 테스트 결과를 참고하여 추가로 찾아진 Stopwords(“nosense.txt” 참고)를 삭제하고, Tf-Idf에서 걸러지지 않는 단어 중 진보와 보수 언론 모두에서 비슷한 빈도로 언급되는 단어를 제거하였다. “new\_liberal.txt”, “new\_conserv.txt”에서 등장하는 단어의 개수를 세어 각각 전체 단어 수 대비 퍼센트를 구하였다. 각 파일에서 함께 등장하는 단어를 찾아서 빈도 차이가 0.000001% 이하일 경우 제거하였다. 최종적으로 테스트에 사용할 “saved\_liberal.txt”, “saved\_conserv.txt” 파일을 생성하였다.

* 1. Topic modeling(LDA) 과정
     + 전처리된 Liberal News와 Conservative News 데이터로 만든 Bag of Words으로 TF-IDF vectorization을 하고, gensim의 LDA 모델 생성 함수 LdaMulticore를 이용해서 LDA 모델을 생성하였다.
  2. 결과 분석 및 결론
* Topic 개수에 따른 단어 분포의 변화

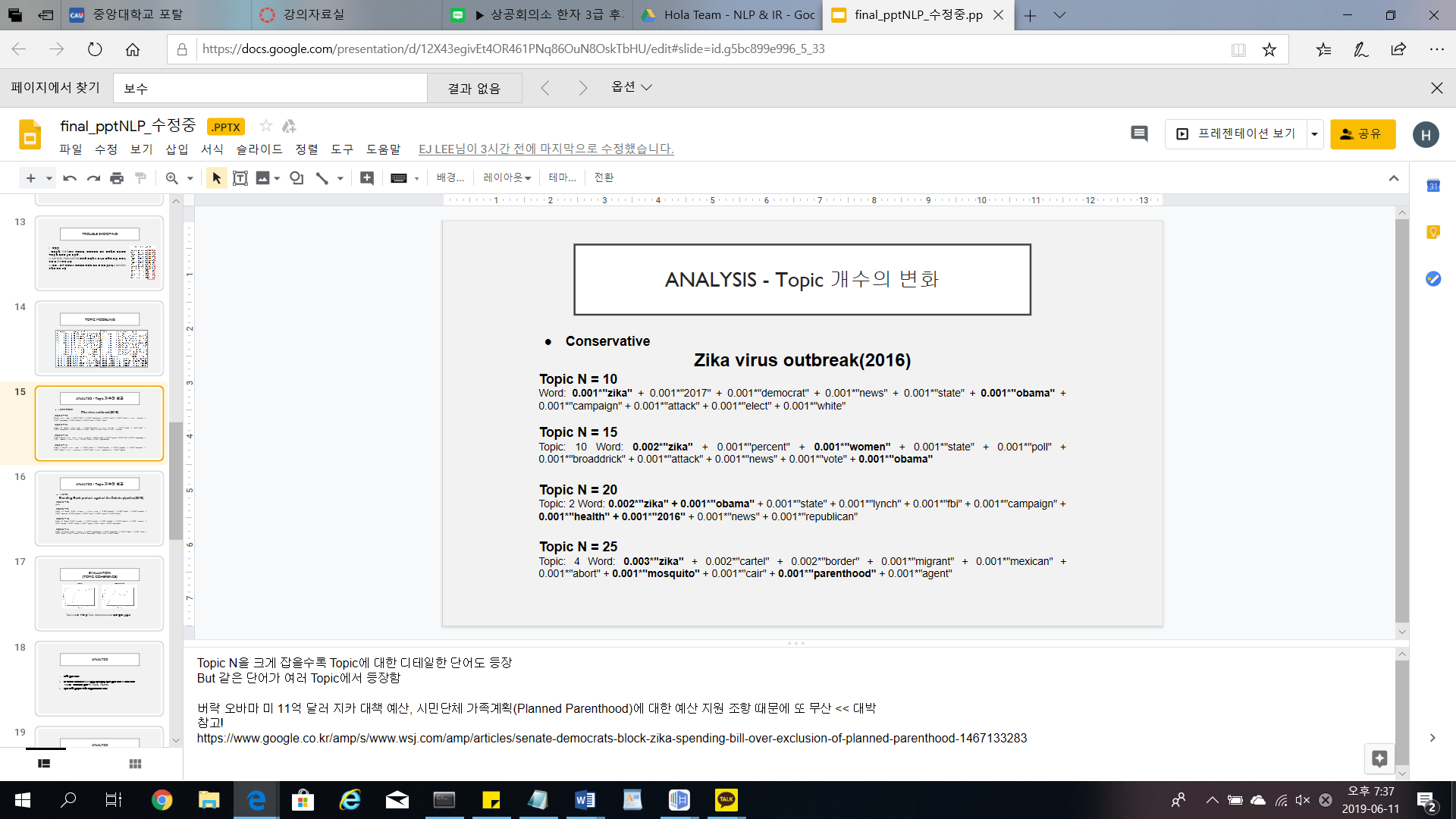


Figure 5. conservative 뉴스에서, 토픽 수를 다르게 했을 때 zika와 관련된 토픽을 이루는 단어들

* + Topic N이 커질수록, Topic을 이루는 단어들의 디테일이 커지나, 같은 단어가 여러 Topic에서 등장하는 문제가 발생한다.
  + N=10일 때, zika, obama가 등장하였지만 직접적으로 드러나는 토픽을 알 수 없었다. 하지만 N=25일 때, zika, obama, mosquito, parenthood의 단어를 통해 ‘버락 오바마 미 11억 달러 지카 대책 예산, 시민단체 가족계획(Planned Parenthood)에 대한 예산 지원 조항 때문에 또 무산’이라는 토픽을 정확하게 파악할 수 있었다.

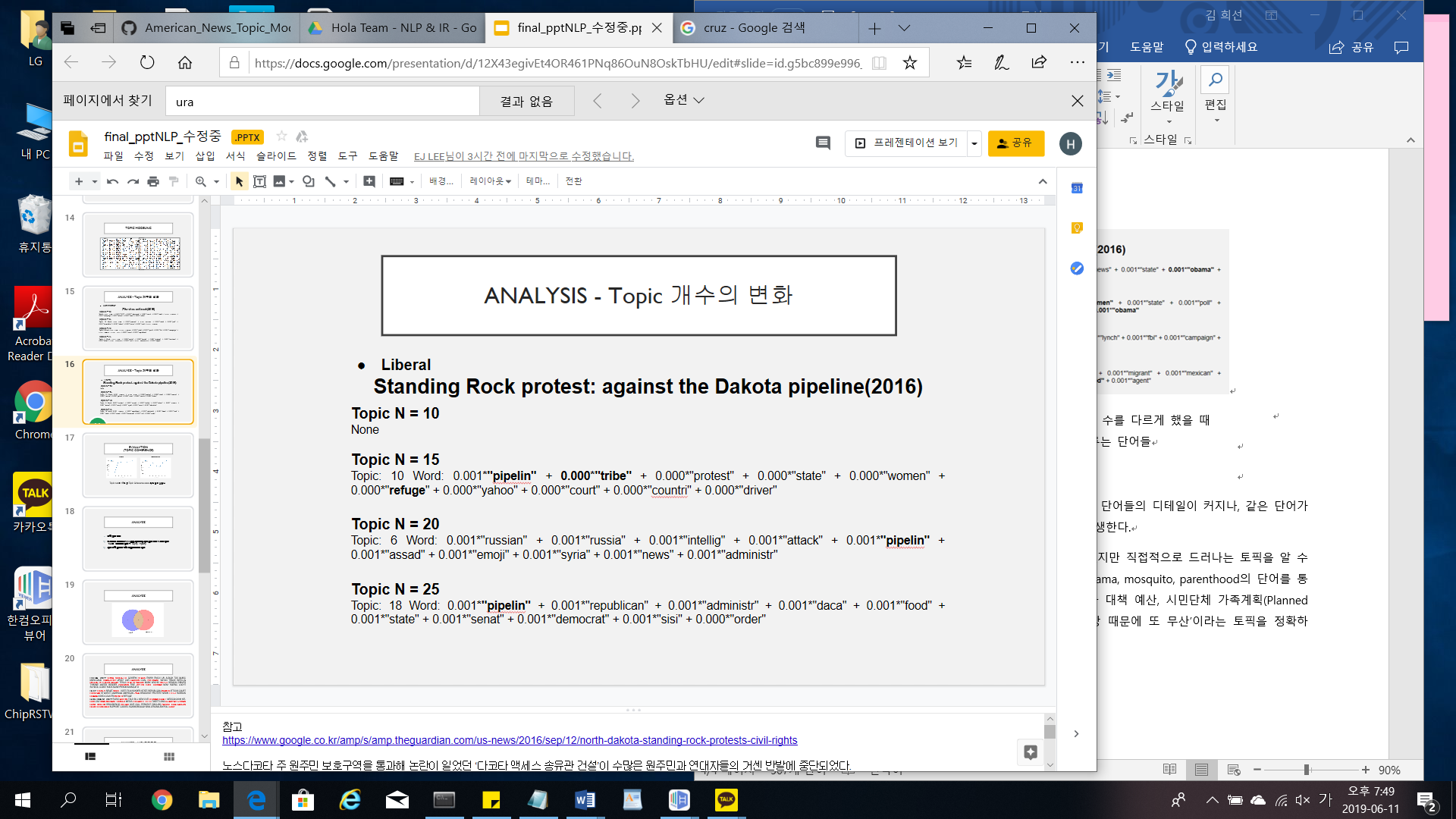


Figure 6. Liberal News에서, 토픽 수를 다르게 했을 때 pipeline과 관련된 토픽을 이루는 단어들

* N=15일 때, Pipeline과 관련된 토픽이 명확하게 드러났다. Pipelin, tribe 의 단어를 통해 ‘노스다코타 주 원주민 보호구역을 통과해 논란이 일었던 ‘다코타 액세스 송유관(pipeline) 건설’이 수많은 원주민(tribe)과 연대자들의 거센 반발에 중단되었다.’다는 기사 내용을 추측할 수 있다.

(<https://www.google.co.kr/amp/s/amp.theguardian.com/us-news/2016/sep/12/north-dakota-standing-rock-protests-civil-rights>)

* N=20일 때, 러시아-시리아, 아사드 단어와 pipeline이 등장하면서 관련성이 적어졌다.(근데 앞에서는 N이 커질수록 토픽 잘나왔다고 했음)
* N=25일 때, daca(미성년 입국자 추방 유예, 미성년자의 국외추방을 유예하는 민법 제도)가 pipeline과 등장하면서 daca가 사건과 직접적인 관련은 없지만 피난민 단어와 관련이 있는 것으로 보인다.
  + 즉, 위와 같은 여러 개의 N개로 모델을 실행시키면서 대체적으로 N이 높아질수록 Topic의 정확도가 높아지는 결과를 보였다. 그렇지만 한 단어가 여러 토픽에서 동시에 등장하는 경우도 증가하며 정확도를 감소시켰다. 이러한 점들을 종합하여 가장 토픽을 잘 나타내는 N은 15인 것으로 판단했다. 또한 Figure 7을 통해 N=15일 때 Topic Coherence Score가 가장 높게 나타나 N=15일 때의 보수 및 진보의 특성을 파악하고자 했다.

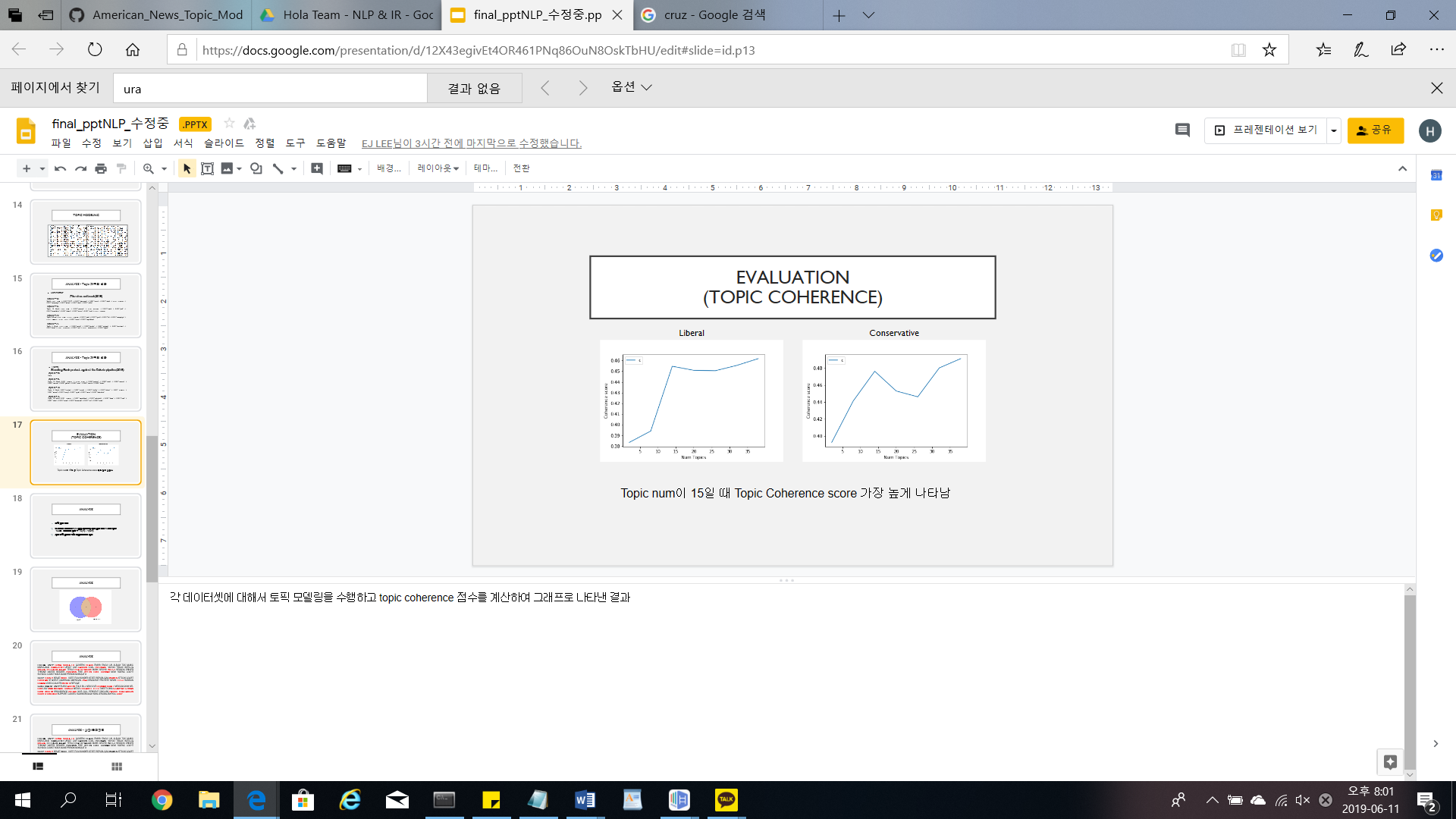


Figure 7. Topic Coherence Score

* 두 진영의 어휘 차이 분석
  + 각 진영의 모델에서 Topic마다 가중치가 가장 높은 단어 10개를 뽑아 총 150개의 단어에서 중복되는 단어를 제거하여 liberty, conservative party에서 각 각 75개, 71개의 단어를 추출했다.
  + 위의 단어를 이용하여 벤다이어그램을 생성하고, 사용 어휘를 비교했다(Figure 8). 하지만 모든 단어를 벤다이어그램에 넣으면 보기에 좋지 않아 단어의 개수로 벤다이어그램을 생성하고, 자세한 단어는 Figure 9에 정리하였다.

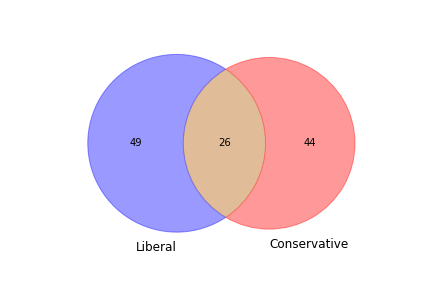
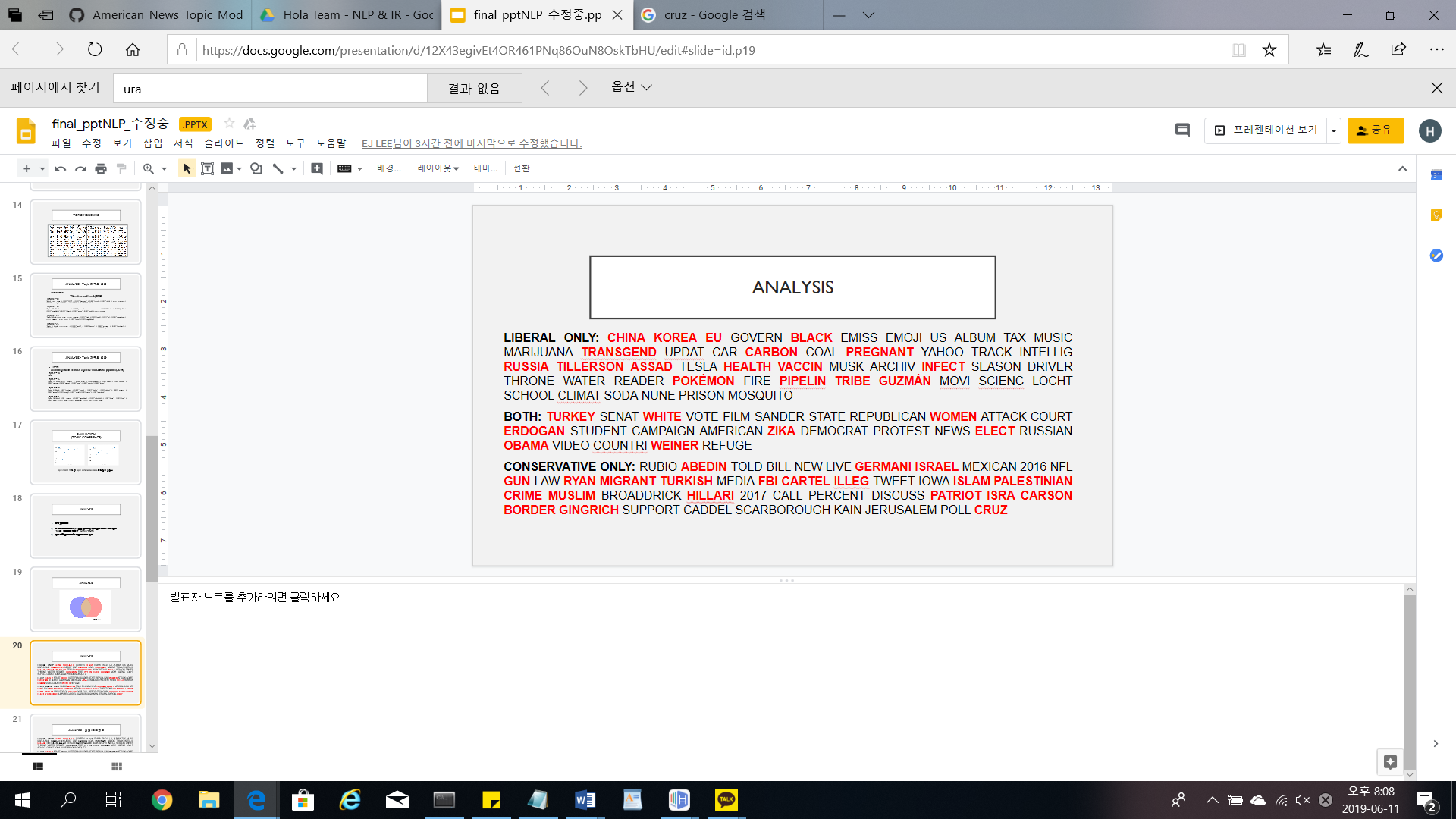


Figure 8. 각 진영에서 사용된 단어의 시각화

Figure 9. 각 진영에서 사용된 단어



* + 단어들 중 의미가 있다고 판단되는 단어들을 빨간색으로 표시하여, 어떤 차이가 있는지 분석했다. 아래에서 자세하게 살펴보고자 한다.
  + liberal에서는 China, Korea, EU, Russia가, Conservative에서는 Israel, Islam, Palestinian, Muslim, Isra, 공통적으로 Turkey가 등장하였다. liberal에서는 conservative보다 분쟁이 적은 나라에 대한 단어가 등장하였다. (Figure 10)

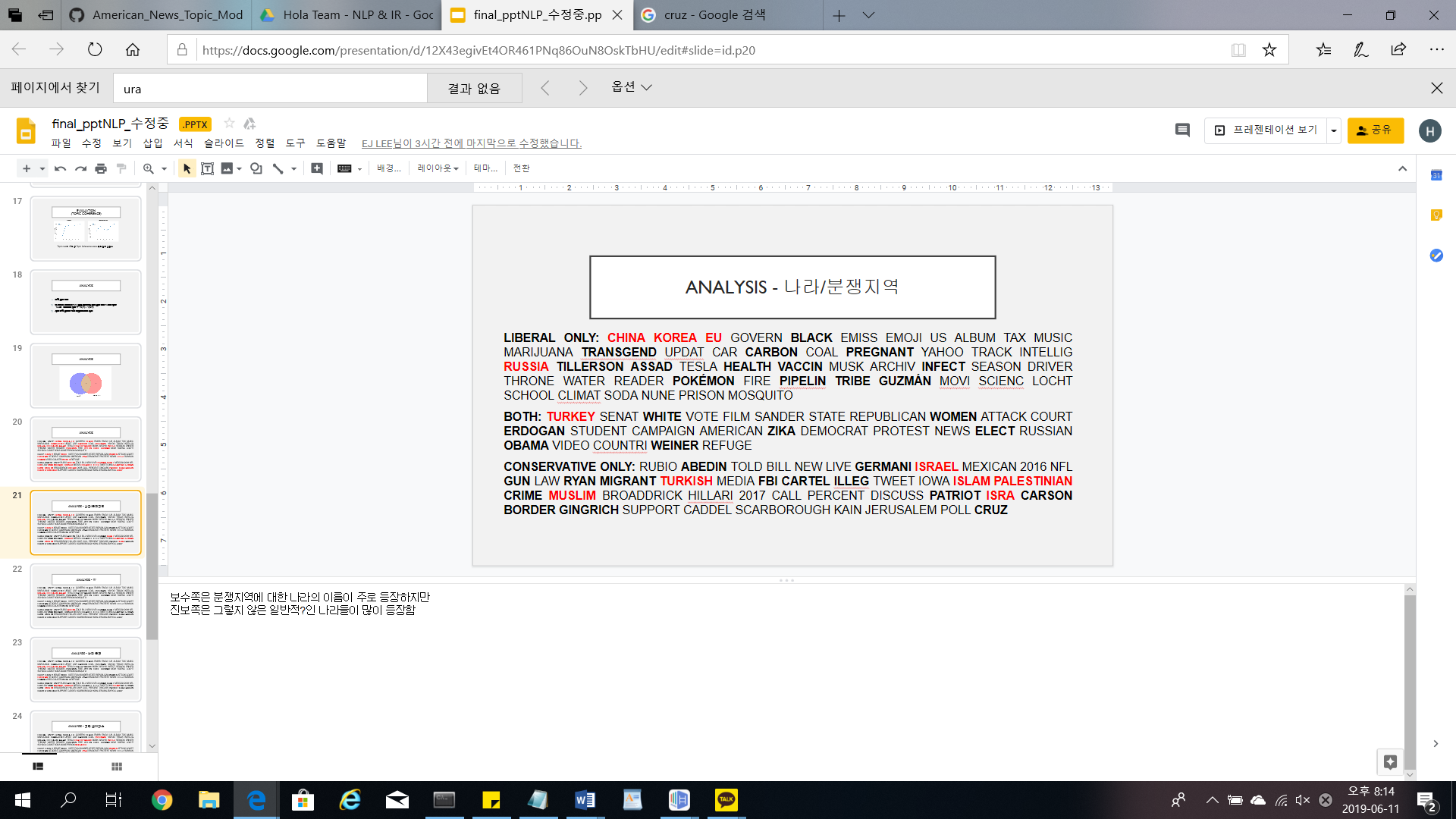


Figure 10. 나라와 관련된 단어들

* + Liberal에서 Tillerson, Assad, Russia가 등장하였다. 공화당인 Tillerson, Trump를 오랫동안 비판해왔던 liberal 진영이 Assad와 관련된 Trump의 행동을 칭찬했기 때문으로 판단된다. (<https://www.thenation.com/article/too-many-of-trumps-liberal-critics-are-praising-his-strike-on-syria/>)

또한 Conservative에서 Abedin과 Hillari가 등장하였다. 이는 민주당인 힐러리의 최측근 보좌관인 후마 애버딘의 전 남편 앤서니 위너의 노트북에서 상당수의 힐러리와 애버딘 간의 이메일이 발견됐던 이메일 스캔들과 관련된 것으로 추정된다. (https://ko.wikipedia.org/wiki/힐러리\_클린턴\_이메일\_논쟁)

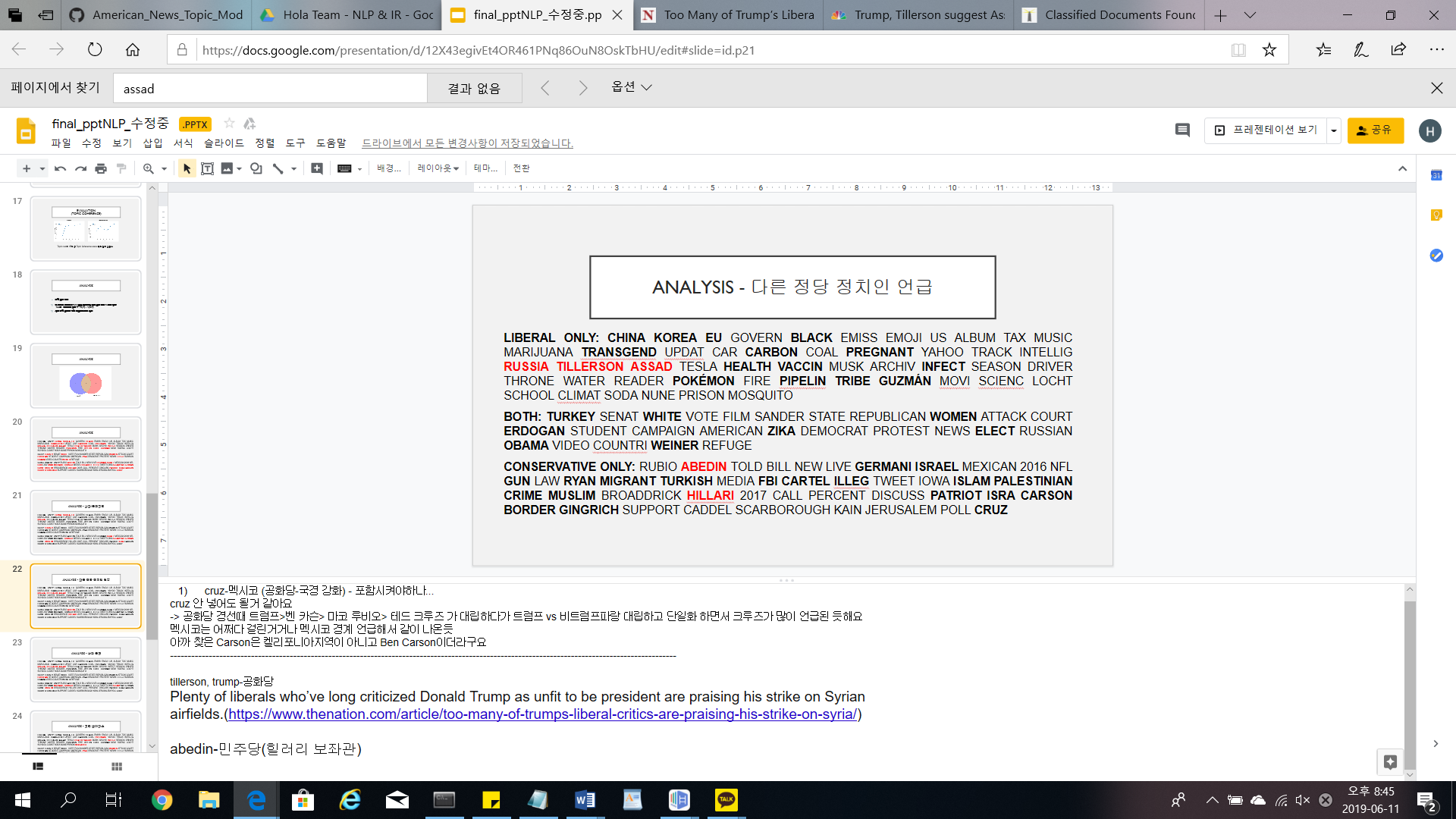


Figure 11. 다른 정당 소속 정치인의 언급

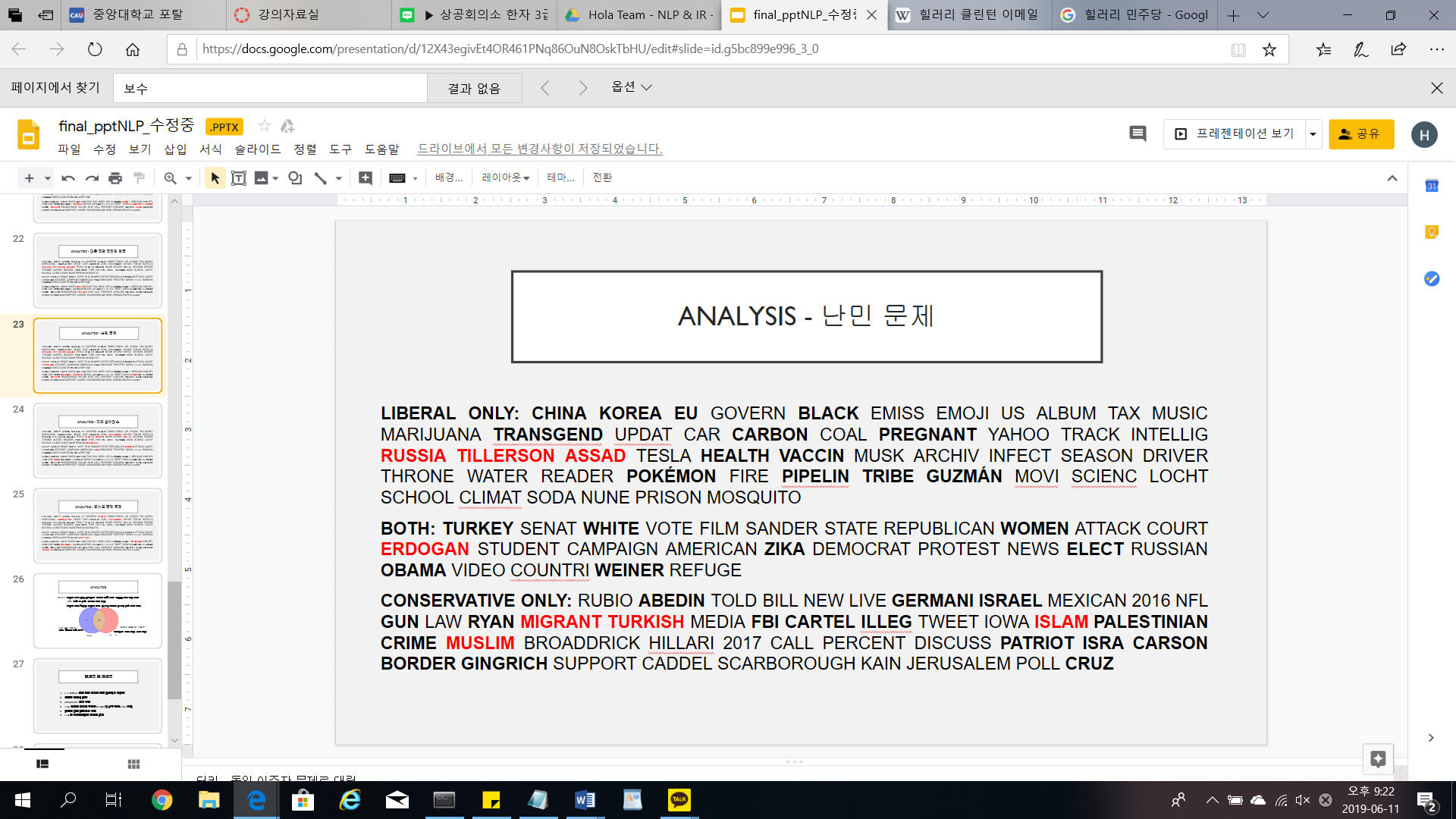


Figure 12. 난민 문제