FINAL REPORT

**TOPIC MODELING BY AMERICAN NEWS**



|  |  |
| --- | --- |
| **과목** | 자연어처리 및 정보검색 |
| **담당교수** | 정재은 교수님 |
| **팀명** | HOLA TEAM |
| **팀원** | Irene Llop Escrich  허수민  김희선  박채린  이은주  Laura Nienajadlo |
| **제출일** | 2019.06.14 |

**- 목 차 -**

1. **프로젝트 개요**3
2. 주제 선정 배경3
3. 프로젝트 주제3
4. 프로젝트 진행 계획 3
5. 역할 분담 3
6. 주간 계획표 3
7. 데이터셋 4
8. Tools 4
9. **프로젝트 진행 과정**5
10. 데이터 전처리5
11. LDA6
12. **결과 분석 및 결론**7
13. Topic 개수에 따른 단어 분포의 변화 7
14. 두 진영의 어휘 차이 분석 9
15. 한계점 12
16. 프로젝트 후기 및 결론 13
17. **프로젝트 개요**
    1. **주제 선정 배경**

대립구조를 가지는 진보, 보수의 성향이 언론에서 어떻게 차이가 나는지 알고자 하였다.

* 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 어휘의 차이점 및 경향을 분석할 것이다. 기사 데이터를 전처리하고, TF-IDF와 LDA를 이용하여 topic modeling 할 것이다. 그 결과로 나온 topic들을 언론사 별로 어떤 어휘를 사용하고 어떤 주제를 다루는지 분석하여 벤다이어그램으로 시각화 할 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**

미국의 다양한 언론사 별 기사 32000개(진보 16000 + 보수 16000)

* + - * + liberal(Atlantic, Buzzfeed News, Vox)
        + conservative(Breitbart, New York Post, National Review, Fox News)

Figure 1을 참고하여 진보, 보수적 성향을 띠는 미국 언론사를 선정하였다.

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

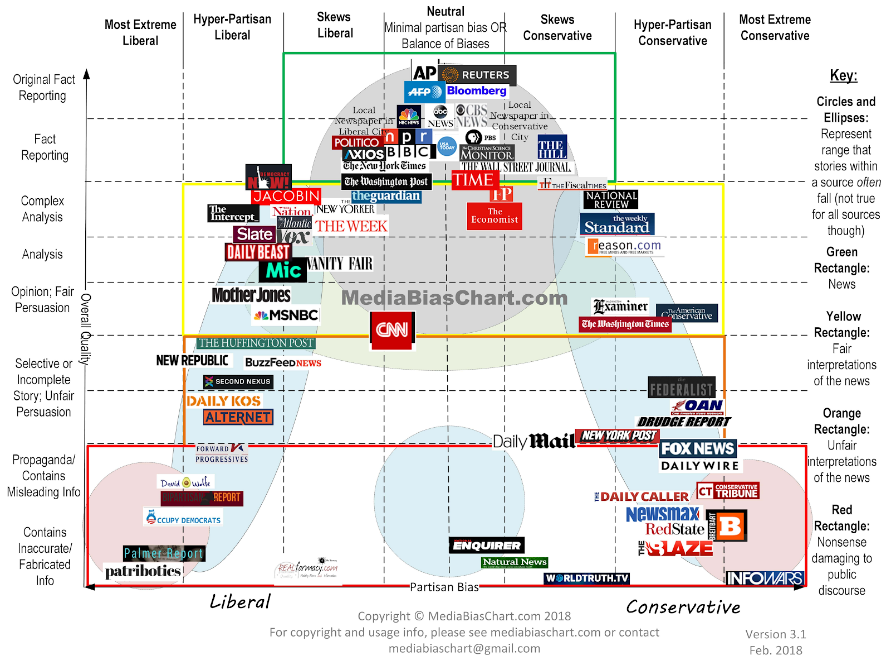


Figure 1. 미국 언론 진보, 보수 성향 차트

* 1. **Tools**

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 개행 문자를 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)

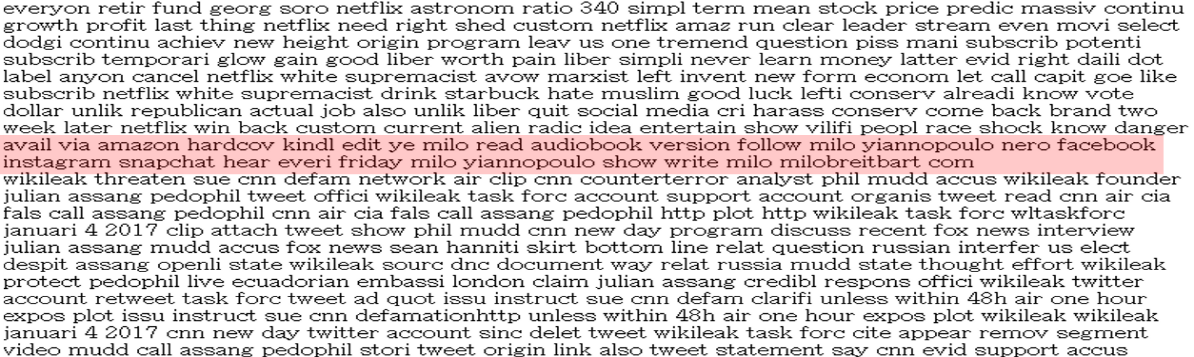


Figure 3. 전처리 결과 파일에 포함된 보수 언론사 Breitbart 기자 SNS홍보문

* + - 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.

1. **결과 분석 및 결론**
2. **Topic 개수에 따른 단어 분포의 변화**
   * Topic N이 커질수록, Topic을 이루는 단어들의 디테일이 커지나, 같은 단어가 여러 Topic에서 등장하는 문제가 발생한다. (Figure 5)
   * 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.

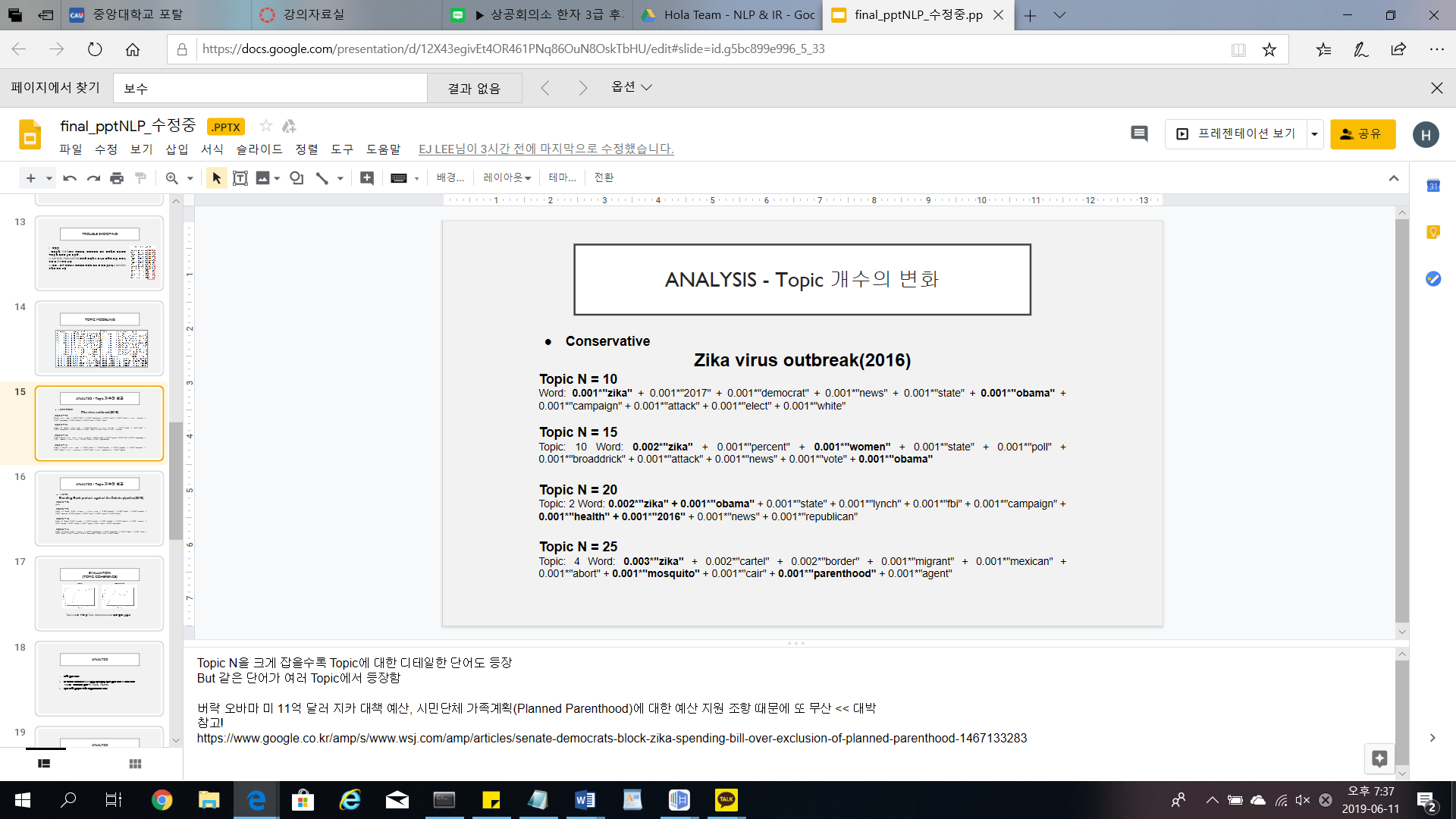


Figure 5. conservative 뉴스에서, 토픽 수를 다르게 했을 때 zika와 관련된 토픽을 이루는 단어들

* Figure 6에서 N=15일 때, Pipeline과 관련된 토픽이 명확하게 드러났다. Pipelin, tribe의 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.
* 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.



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

* + 즉, 위와 같은 여러 개의 N으로 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 나타났기 때문에 N=15일 때의 보수 및 진보의 특성을 파악하고자 했다.

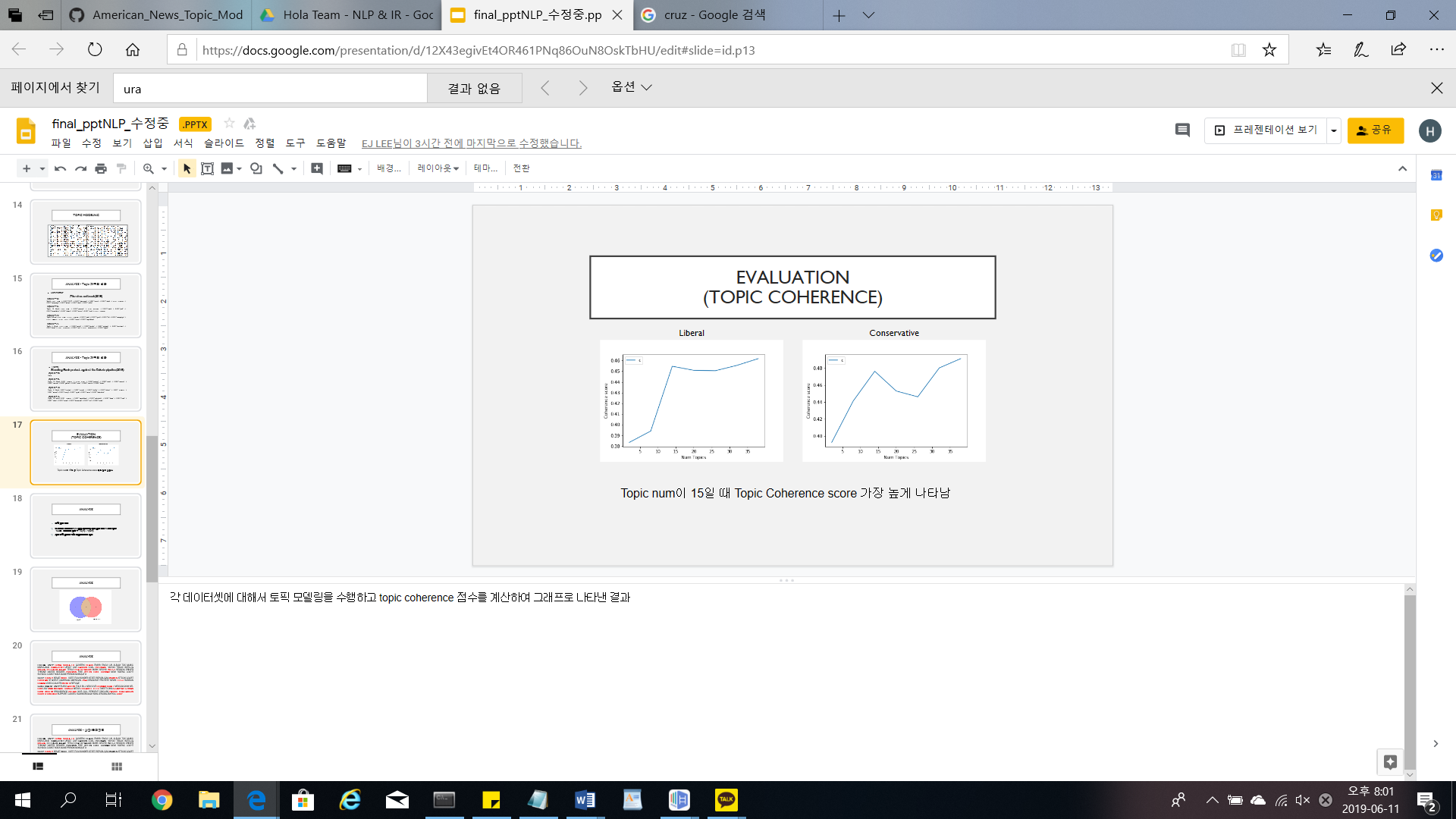


Figure 7. Topic Coherence Score

1. **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.

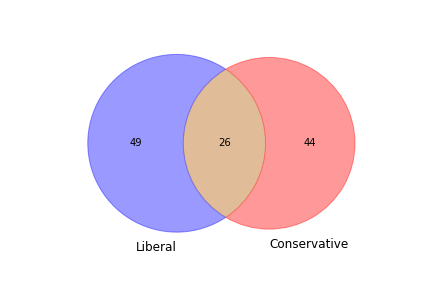


Figure 8. 각 진영에서 사용된 단어의 시각화

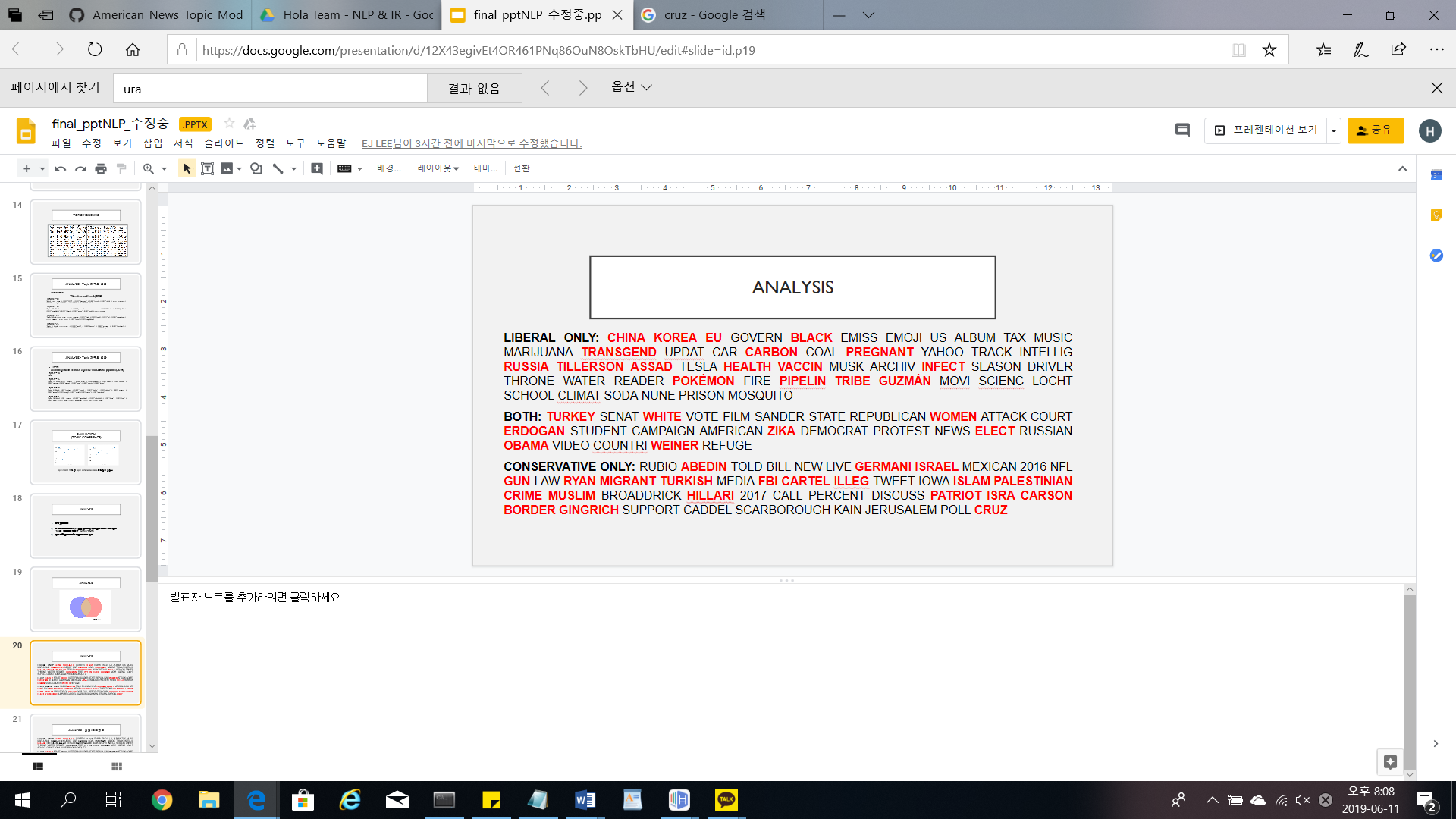


Figure 9. 각 진영에서 사용된 단어

* + 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, Isra이, 공통적으로는 Turkey가 등장하였다. Liberal에서는 Conservative보다 분쟁이 적은 나라에 대한 단어가 등장하였다. (Figure 10)

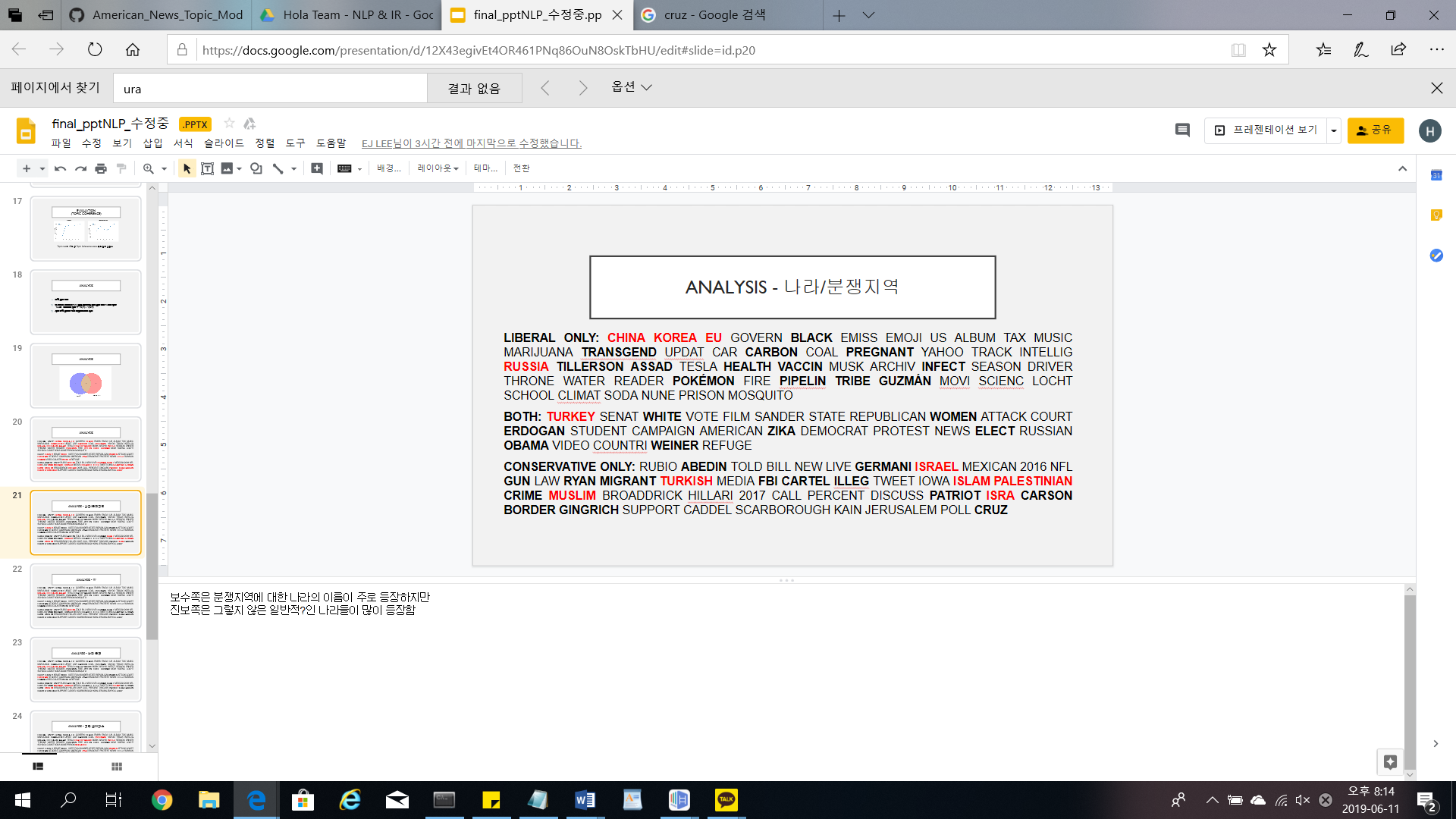


Figure 10. 나라와 관련된 단어들

* + 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)

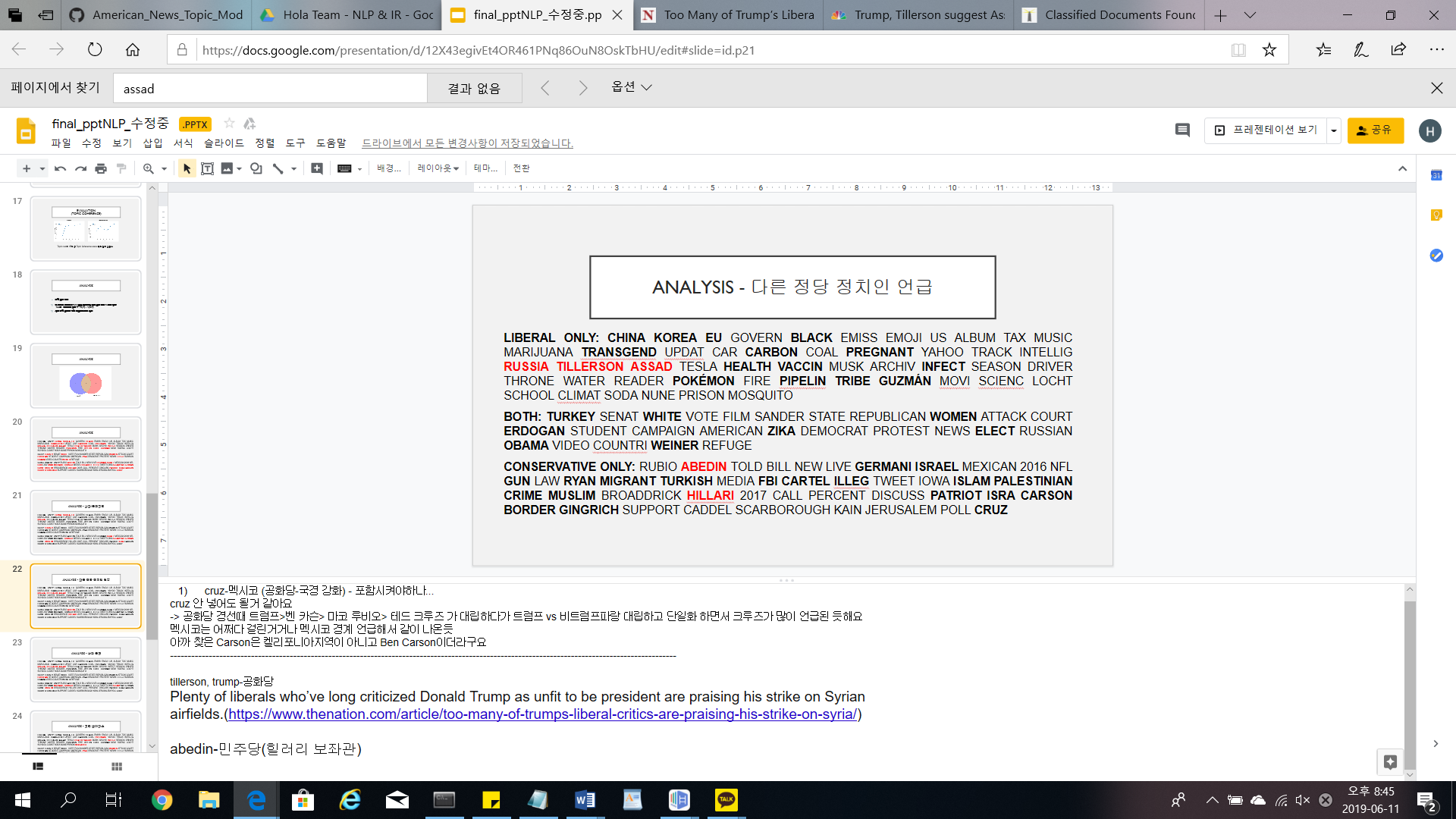


Figure 11. 다른 정당 소속 정치인의 언급

* + Liberal과 conservative에서 모두 난민문제에 관한 토픽이 발견되었다. liberal에서 발견된 Russia, Tillerson, Assad 단어를 통해, 러시아가 Assad를 지원하고 미국이 시리아 반군을 지원했던 시리아 내전에 관한 토픽을 보이고 있다.

Conservative에서 Migrant, Turkish을 통해, 터키 대통령 에도르안 대통령이 아사드 정부를 반대했고, 시리아 내전로 인해 난민이 급증했던 사건에 관한 토픽으로 짐작된다.

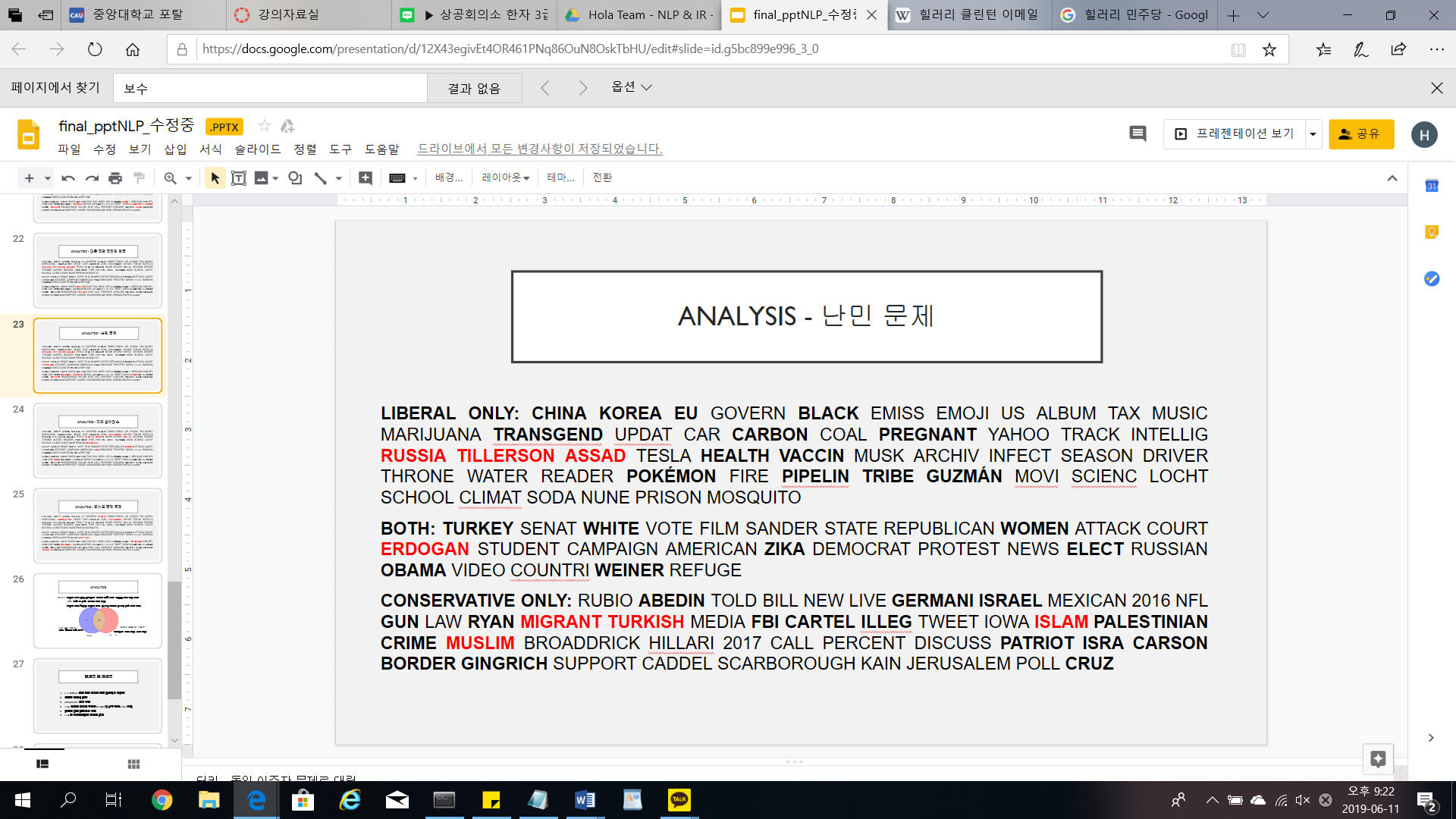


Figure 12. 난민 문제

* + Zika virus 오바마 케어 관련하여 좌우 대립이 있었다. 이와 관련된 토픽으로, Liberal에서는 ‘pregnant, health, vaccin, infect’가, Conservative에서는 ryan(오바마케어 관련 정치인)이 등장하였다. 즉 zika virus는 당시 중요한 이슈였다고 판단된다.

(<https://en.wikipedia.org/wiki/American_Health_Care_Act_of_2017>)

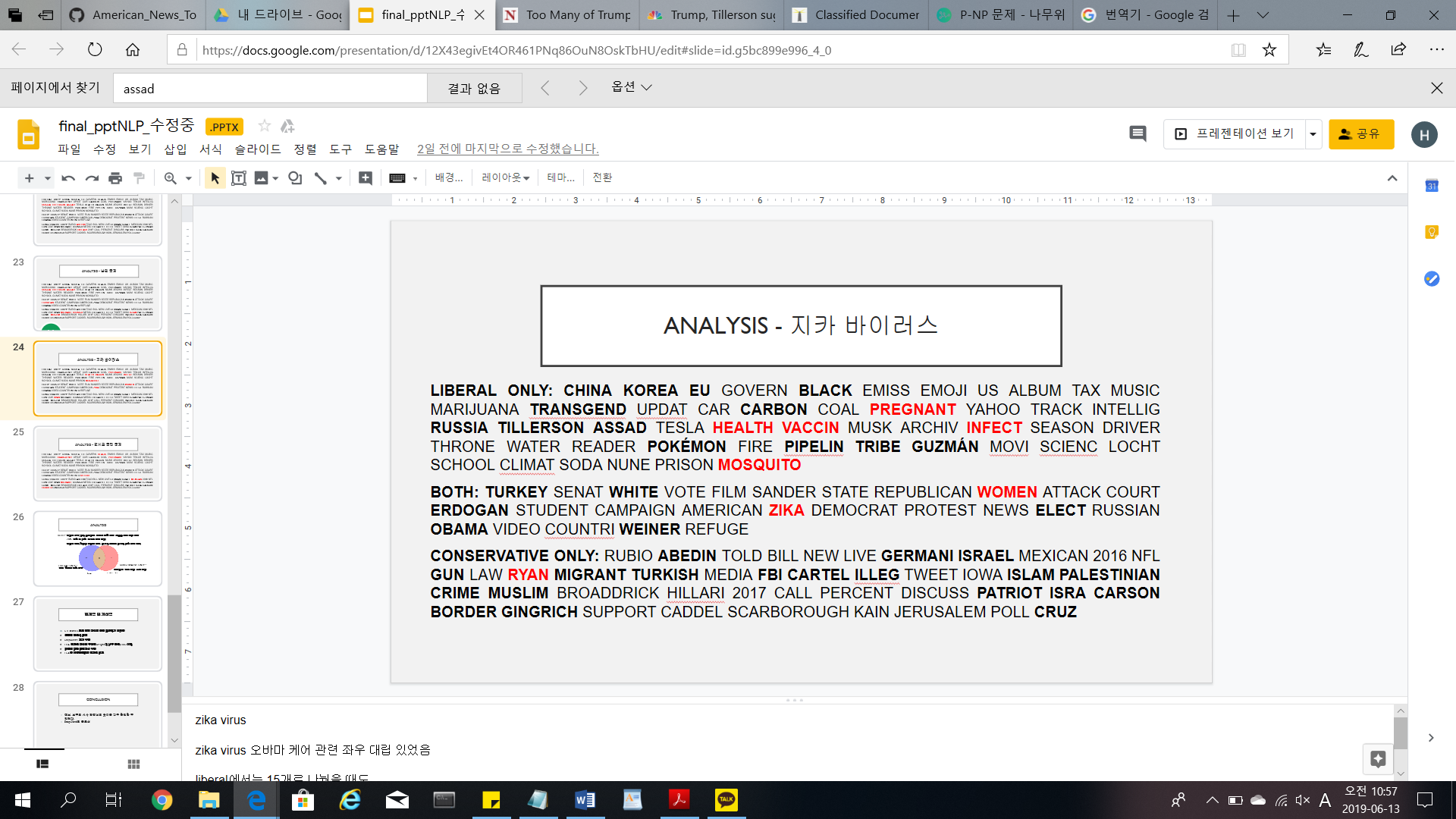


Figure 13. 지카 바이러스

* + 2016~2017년도 기사에서 지지율 변화 및 후보와 당의 공약 관련 기사가 많이 포함되어 있다. 공통적으로는 zika, women, white, refuge 등이 언급되었지만, liberal은 black, transgender, pregnant, climatee, carbonate, emission의 단어를 통해 인권 및 환경문제에 집중하고 있다는 것을 알 수 있다. Conservative는 migrant, mexican border, muslim를 통해 안보 및 난민 문제를 중요하게 여기는 것을 알 수 있다.

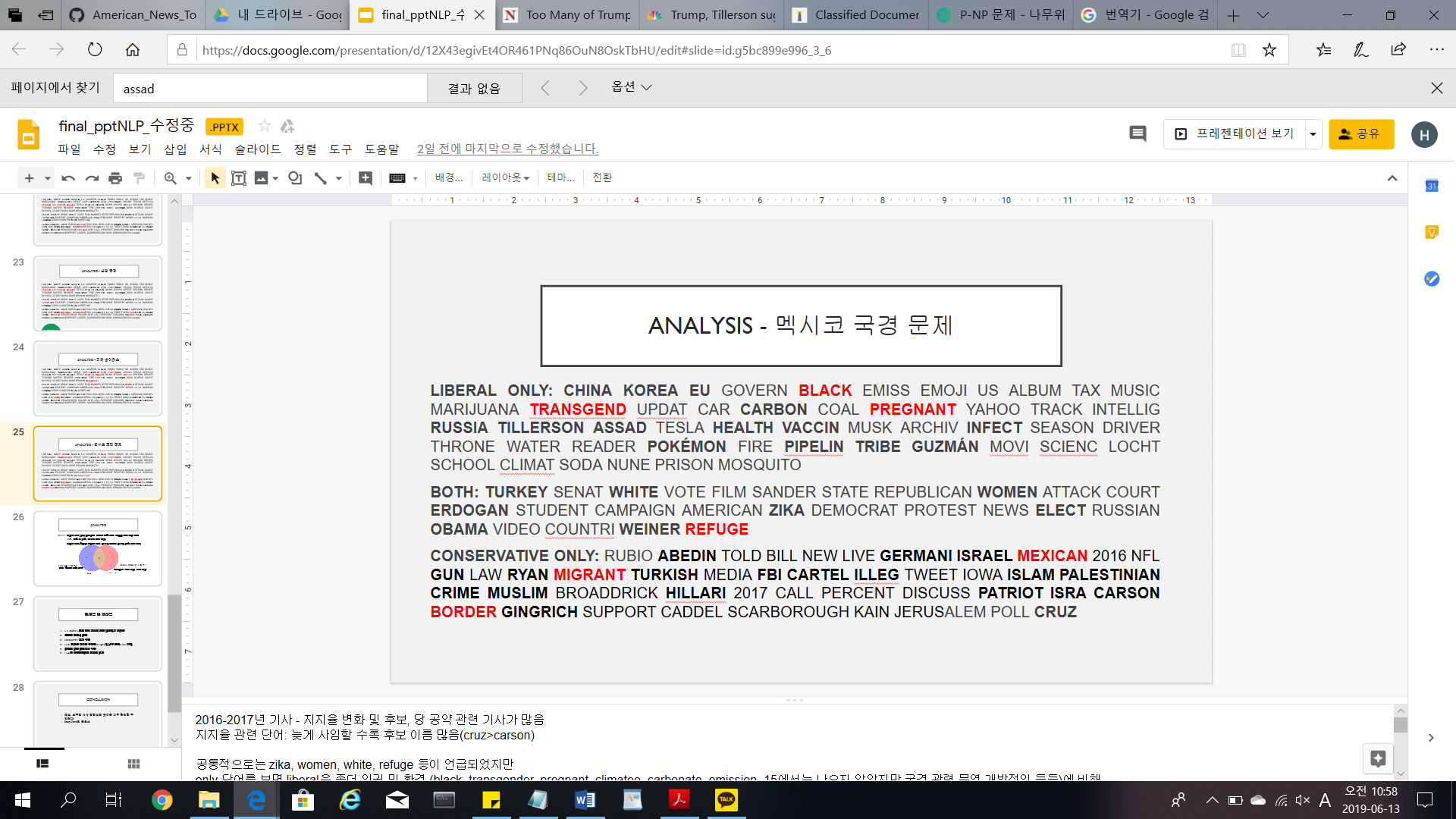
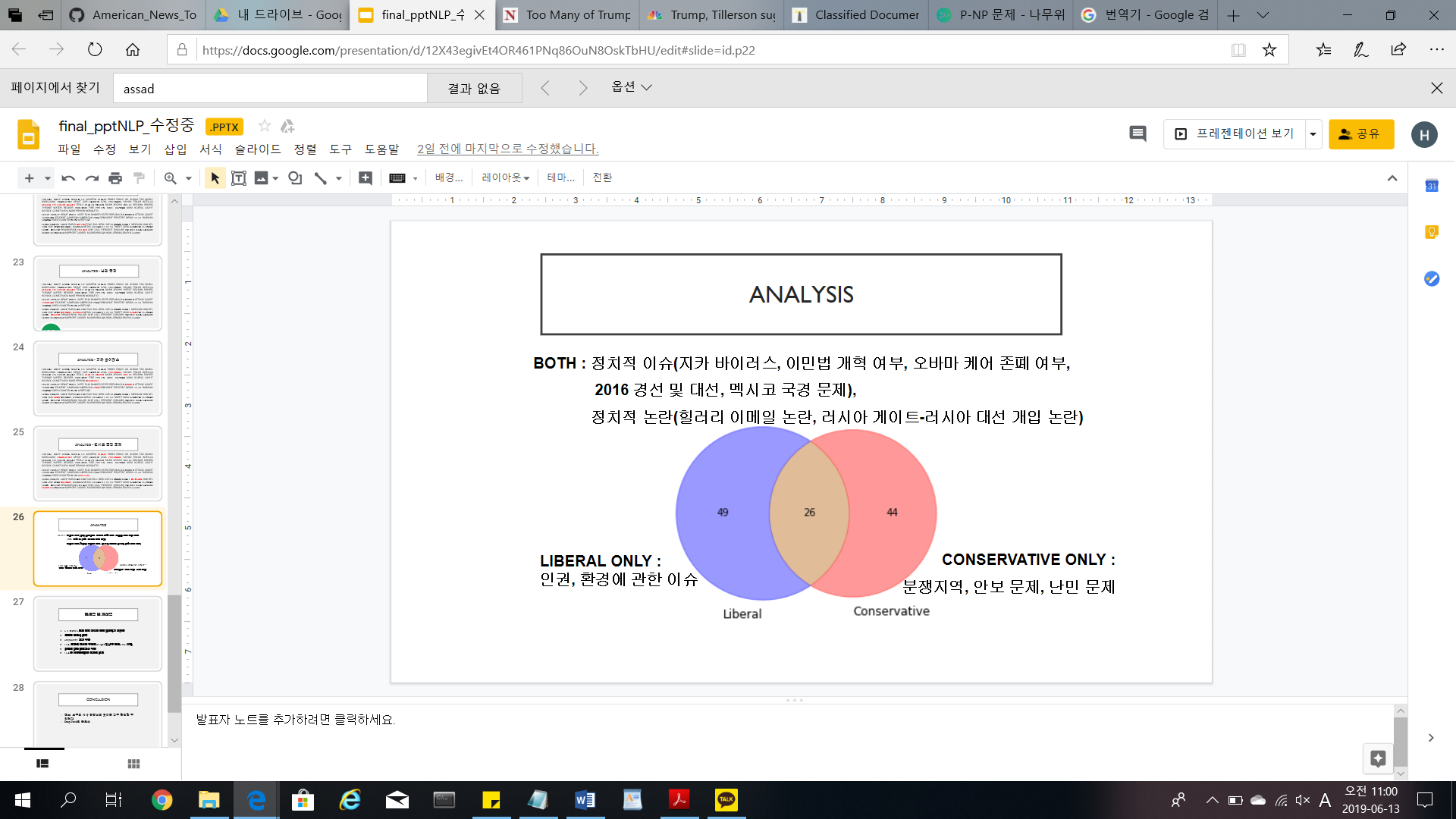


Figure 14. 주요 공약 차이

* + 벤다이어그램을 통한 토픽 시각화



1. **한계점** 
   * Stemming으로 인해 단어가 일부 잘리거나 바뀌면서 모델링 된 결과를 분석하는데 어려움이 있었다.
   * stopwords를 제거했음에도 남아있는 stopwords로 인해 토픽을 파악하기 어려웠다. 데이터 셋에 맞는 stopwords를 customize한다면, 더 많은 분석이 가능했을 것으로 예상된다.
   * 모델링 결과로 주어진 단어 분포에서 weight값이 0.003이하로 작은 수에 맴돌고 있어 중요한 단어를 파악하기 어려웠다.
   * 미국 정치에 대한 배경지식이 부족해 토픽을 분석하는 데에 한계가 있었다.
2. **프로젝트 후기 및 결론**

* 진보, 보수의 시사 방향성 차이를 어느 정도 확인할 수 있었으나 살짝 아쉬움이 남는다. 좀 더 극명한 차이를 위해 중립적 입장은 배제하였는데 중립적인 언론사도 같이 모델링하여 결과 분석을 하였다면 또 다른 의미를 찾을 수 있지 않았을까 싶다. 또한 2016-17년도의 정치 흐름에 대한 진보, 보수의 어휘 차이를 알아보기에는 시간 범위가 광범위했다는 의견도 있었다. 기간을 한달로 축소시키거나, 특정 사건을 다루는 기사 데이터를 사용하여 분석했다면 더 좋은 결과를 얻는 프로젝트가 될 수 있었을 것 같다.

5주간 프로젝트를 진행하면서 소통과 분담의 중요성을 다시 느낄 수 있었던 의미있는 시간이였다. 프로젝트 결과가 뚜렷하게 나올 줄 알았는데 우리가 예상했던 것과는 다른 결과가 나와서 데이터 전처리에 예상보다 많은 시간과 정성을 쏟았다. 진보수의 어휘차이를 파악하기 위해서 stopword 제거는 매우 중요한 과정이며 양간의 중복되는 단어는 크게 의미를 가지지 못한다는 것을 알게 되었다.