2021128039 황종문

Voyant Tools and Descriptive Metadata: A Case Study in How Automation Can Compliment Expertise Knowledge

Voyant Tools와 설명적 메타데이터: 자동화가 전문 지식을 보완하는 방법의 사례 연구

Kate Gregorya , Lauren Geigerb and Preston Salisburyb a Congressional and Political Research Center, Mississippi State University Libraries, Mississippi State, MS, USA; bCollection Management Services, Mississippi State University Libraries, Mississippi State, MS, USA

1. 서론

무료 오픈소스 애플리케이션 Voyant-Tools가 인문학과 사회과학 연구자, 문헌 연구 전문가를 도울 수 있는 방법을 설명하기 위한 논문으로, 텍스트 마이닝 등 메타데이터의 분석을 할 때 Voyant- tools가 어떤 방식으로 사용될 수 있는지를 보여준다.

2. 연구 문제 및 방법

연구자들은 Voyant-tools를 이용하여 미국 미시시피주 상원의원 Alan Nunnelee 의원실에서 보낸 801건의 서신 모음(이하 'Nunnelee 컬렉션')을 텍스트마이닝하여 정량적으로 분석하였다.

제공받은 자료는 PDF또는 MS-Word자료 였기 때문에 직접 텍스트를 추출하거나, OCR을 이용하여 텍스트를 추출하였다. 텍스트를 분석하기 위한 사전단계로 구두점이나 어간을 제거하고 Stopword 목록을 생성하였다.

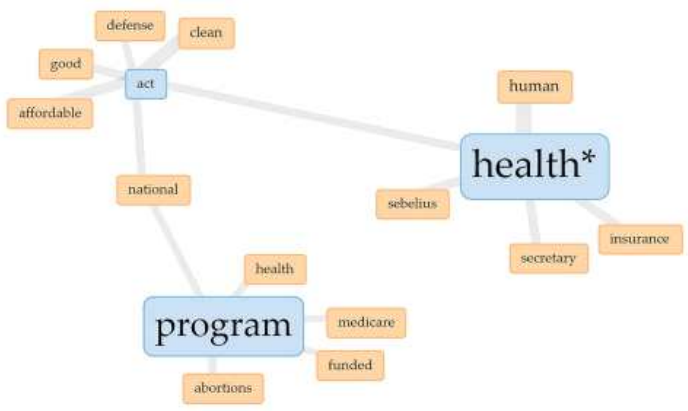
"U.S.," "Rep," "D.C.," "Mr.," "H.R.,"등과 같은 약어와 "Dear," "Thank," "Letter", "Sincerely" 등 공식적인 서신과 직접적으로 관련된 단어와 "Washington" 및 우편번호 "20515", 선출직 공직자에게 공식적으로 호칭하는 데 사용된 "Honorable"과 "including," "believe," "continue," "provide," , "support" 와 같이 광범위하게 사용되는 동사와 "time," "information", "issue." "Office," "department", "federal"과 같은 명사를 Stopword 목록에 포함하였다. 이는 워드클라우드(Cirrus)에서 컬렉션의 주제를 보다 실질적으로 분석을 하기 위한 목적이다.

3. 연구 결과 및 해석

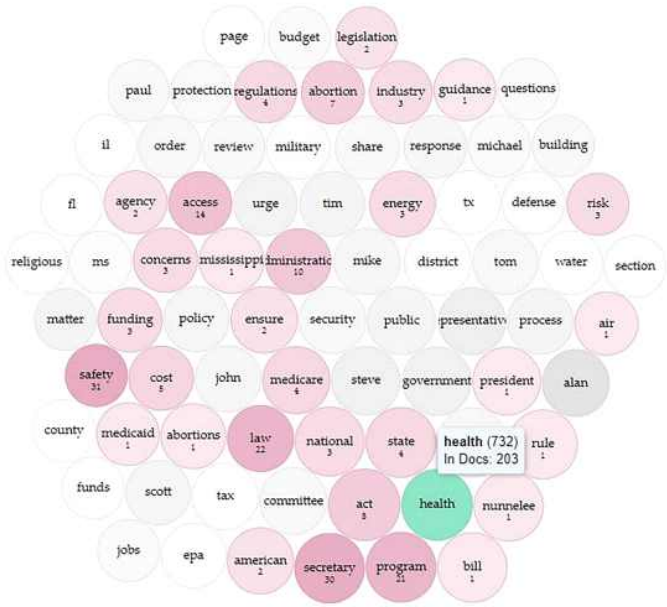
빈출 어휘 간 관계를 이용하여 분석한 키워드 짝은 아래와 같다

- Energy policy(에너지 정책) - United States(미국)
- Health insurance(건강 보험) - United States(미국)
- National Security(국가 안보) - United States(미국)
- United States. Congress. House – 미국. Congress. House
- Patient Protection and Affordable Care Act (PPACA, 환자 보호 및 의료비 부담 경감법)

(1) Collocate link – 주제어 간 연결 관계

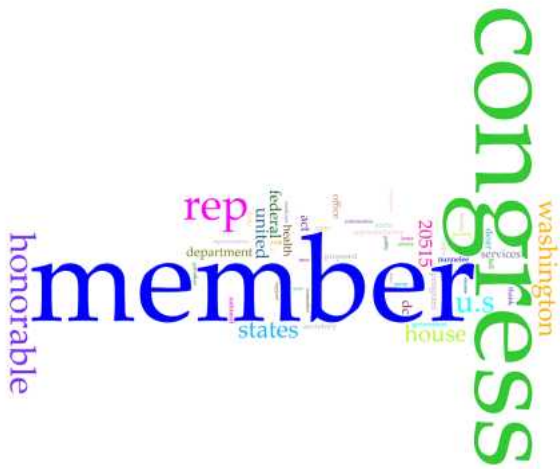


(2) Terms berry



(3) 전처리 전 Cirrus

(4) 전처리 후 Cirrus



(5) Summary – 전체 주요 어휘 및 챕터별 주요 어휘

Most frequent words in the corpus: act (789), health (732), program (657), state (656), secretary (640)

Distinctive words (compared to the rest of the corpus):

1. 01-18-13 Letter to Presid...: russian (14), children (12), adoption (6), families (8), russia (4)
2. 010612 Recess Appointment...: appointments (7), cordray (3), appointment (3), senate (6), recess (2)
3. 011513 Letter to Sec...: abedini (13), iran (10), abedini's (3), iran's (4), revolutionary (3)
4. 011912 Solyndra Bonuses...: solyndra (5), bonuses (4), tx (7), solyndra's (3), oh (4)
5. 012612 Iran Sanctions...: sanctions (9), iran (8), nuclear (8), tx (10), ny (7)
6. 02 17 2012 MCs Letter...: treaty (14), 03 (9), 01 (10), ratification (5), sea (5)
7. 020612 Letter to HHS...: abortifacients (4), religious (7), affiliated (3), conscience (4), coverage (4)
8. 020612 Rokita Freshman...: device (10), tax (14), medical (8), innovation (4), jobs (6)
9. 021712 Flake Letter to...: treaty (14), 03 (9), 01 (10), ratification (5), sea (5)
10. 030912 Letter Obama Suppo...: israel (11), iran (7), nuclear (5), israeli (4), developments (2)
11. 041511 FDA Artificial...: john (12), michael (9), james (6), jr (6), david (6)
12. 05 17 2011 full debate...: az (5), tariff (3), tax (6), jim (6), tx (5)
13. 05.10.13 Nunnelee RD...: vacancy (5), rural (10), development (11), approval (5), project (5)
14. 072312-HHS_TANF_Waiver_Le...: 1115 (3), waivers (3), welfare (3), section (6), anf (2)
15. 090911 LTR to President...: immigration (8), illegal (7), amnesty (5), enforce (4), laws (4)
16. 102313_FINAL_CMS Letter...: cancer (15), oncology (4), rates (8), hospital (6), payment (7)
17. 1099 Repeal Support HR 4: reporting (5), business (8), ppaca (4), 1099 (3), owners (3)
18. 11 19 12 Benghazi WH...: situation (5), room (4), sure (3), benghazi (2), gave (2)
19. 111611 Keystone XL Letter...: xl (4), keystone (4), project (5), energy (6), security (7)
20. 111711 EPA E15 Letter: engines (11), ethanol (9), fuel (7), els (5), gasoline (6)

4. 결론 및 제언

Voyant-Tools는 텍스트 자료의 패턴을 찾고 전체 말뭉치 또는 특정 부분을 설명하는 데 도움을 준다. Voyant-Tools를 사용하면 컬렉션을 설명하는 데 필요한 시간을 줄일 수 있다. 주제 제목과 키워드를 만드는 데 가장 유용한 도구는 Cirrus, Summary, Terms Berry, Collocate 도구였다. 개별 용어와 말뭉치가 서로 어떻게 연관되어 있는지 확인하기 위해 Cirrus와 Terms Berry를 사용했고, 어떤 용어가 가장 비중이 크고 어떤 용어를 강조해야 하는지 파악하는 데는 Summary와 Collocate-Links가 도움이 되었다.

Kate Gregory, Lauren Geiger & Preston Salisbury (2022) Voyant Tools and Descriptive Metadata: A Case Study in How Automation Can Compliment Expertise Knowledge, Journal of Library Metadata, 22:1–2, 1–16, DOI: 10.1080/19386389.2022.2030635