2018 – 가을학기 - 온라인데이터수집과 분석

**Homework2**

2018321084 안주영

**목적**: 서로 다른 4개의 신문사 (경향, 동아, 조선, 한겨레)에서 작성한 ‘양진호 회장’ 사건 관련 기사를 분석하여 해당 사건을 바라보는 신문사별 관점과 시각이 차이가 나는지를 알아보는 것

**1. 수집방법**

각 신문사별로 url이 저장되어있는 txt파일을 읽어서 신문사와 url이 저장되어있는 데이터프레임을 생성하였다. 기존에 수업시간에 사용하였던 네이버 뉴스의 기사를 가져오는 함수를 사용하여서 데이터프레임에 새로운 컬럼에 기사내용이 저장되도록 하였다.

기존에 수업시간에 사용하였던 함수의 경우 requests 모듈의 에러로 인하여 기사의 내용을 전부 긁어오지 못하는 문제가 발생하여서 urllib.request모듈을 사용하여서 기사내용을 불러올 수 있도록 함수를 수정하였다.

**2. 전처리방법**

정규식 모듈인 regex모듈을 사용하여서 텍스트 전처리를 수행하였다. 네트워크 분석에서 문장단위로 단어 간의 연관성을 알아보기 위해서 필요하기 때문에 문자와 공백 그리고 문장을 나누는 단위인 마침표와 물음표 느낌표를 제외한 텍스트를 제거하였다.

추가적으로 ‘양진호’라는 인물은 언급하는데 ‘양진호’ , ‘양 회장’ , ‘양회장’이라는 단어를 혼용하고 있기 때문에 해당 단어들을 모두 ‘양진호’로 바꾸어 주었다.

한글 텍스트분석이기 때문에 대소문자 변경을 따로 해주지 않았고 POS-tagging과정에서 자동으로 token화 되기 때문에 토큰화 또한 따로 해주지 않았다.

**3. 분석방법**

텍스트 전처리가 된 텍스트 파일들을 활용하여서 신문사와 텍스트 내용을 컬럼으로 갖는 데이터프레임을 생성하고 분석에 사용하였다.

*3-1. 빈도 분석*

텍스트와 출력할 단어의 개수를 입력 받는 cntWord함수를 생성하였다. cntWord함수는 입력으로 텍스트와 숫자 n을 받아서 해당 텍스트에서 가장 높은 빈도를 보이는 n개의 단어와 각 단어 별 빈도를 리스트 형태로 출력해주는 함수이다. 또한 명사와 형용사 동사를 제외한 품사의 경우 제거하였다.

사용자 사전을 등록하기 위해서 kornounextractor 모듈을 사용해서 2번이상 등장하는 단어를 사용자 사전으로 등록하였다. ‘갑질’이라는 단어가 갑지로 등록되는 문제가 있어서 수정해주는 코드를 사용하였다.

불용어사전에는 신문사의 이름과 기자 포탈의 이름인 네이버 등을 포함시켰다. 또한 1글자 단어의 경우 뜻을 갖지 않는 경우가 많기 때문에 단어의 길이가 1인 단어들도 제거하였다.

각 신문사별로 뉴스기사 텍스트를 모두 합하여 빈도분석을 시행하였고 각 신문사별로 높은 출현 빈도를 보이는 10개의 단어를 출력해서 비교하였다.

*3-2. 네트워크 분석*

빈도분석과 전체적으로 유사하게 진행하였다. twitter가 아닌 komoran을 사용하여서 POS-Tagging을 시행하였다. 빈도분석과 동일한 방식으로 사용자사전과 불용어사전을 지정하였다.

수업시간에 활용하였던 do\_kr\_sna함수를 사용하기 위해서 kr\_sna.py파일을 동일한 폴더에 복사하여서 활용하였다. do\_kr\_sna함수의 경우 기본값으로 상위빈도를 보이는 20개의 단어를 활용하여서 네트워크의 노드를 구성하게 되는데 실제 네트워크 분석을 시행하기 위해서는 빈출단어를 사용하는 것 보다 원하는 노드를 지정하여서 네트워크 분석을 시행하는 경우가 더 효과적이므로 노드를 지정해주었다.

우선 사건당사자의 이름은 ‘양진호’를 노드에 추가하였고 ‘갑질’과 연관된 단어와 ‘위디스크’라는 업체와 관련된 단어, 그리고 수사진행과정과 관련된 단어를 노드에 추가하였다.

netwk함수는 뉴스기사와 단어를 입력으로 받아서 해당 뉴스기사에서 해당 단어와 함께 출현하는 단어들을 리스트형태로 출력해주는 함수로 구성하였다. 품사의 경우 명사인 경우만 사용하였다. 빈도수가 높은 단어부터 내림차순으로 출력해주고 단어와 함께 빈도수도 함께 출력해준다.

**4. 분석결과**

*4-1. 빈도 분석 결과*

[표 1] 신문사별 10개 최빈 단어와 빈도수

|  |
| --- |
| Chosun  [('양진호', 57), ('폭행', 31), ('위디스크', 26), ('했다', 23), ('직원', 22), ('경찰', 20), ('위반', 17), ('영상', 14), ('있다', 14), ('한국미래기술', 13)]  Kyunghyang  [('직원', 32), ('양진호', 27), ('했다', 25), ('폭행', 23), ('직장', 21), ('괴롭힘', 20), ('상사', 18), ('폭언', 16), ('갑질', 16), ('있다', 15)]  Hankyoreh  [('웹하드', 57), ('업체', 54), ('필터링', 32), ('양진호', 30), ('했다', 26), ('불법', 24), ('직원', 20), ('영상물', 19), ('위디스크', 19), ('폭행', 17)]  Donga  [('양진호', 67), ('했다', 51), ('직원', 32), ('폭행', 26), ('위디스크', 19), ('교수', 18), ('사무실', 14), ('영상', 13), ('압수수색', 13), ('경찰', 12)] |

워드 클라우드의 경우 단어 수를 40개까지 늘려서 그려보았다.

|  |  |
| --- | --- |
| **<조선일보>** | **<동아일보>** |
| **<한겨례>** | **<경향신문>** |

[그림 1] 각 신문사별 워드클라우드(상위 40개 단어 사용)

*4-2. 네트워크 분석 결과*

노드로 사용한 단어리스트는 가해자 이름인 ['양진호'], 갑질과 관련된 단어 ['갑질','괴롭힘','상사'], 업체와 관련된 단어 ['위디스크','영상','음란물', '불법'], 수사진행 사항과 관련된 단어 ['수사', '압수수색', '검찰', '경찰','폭행'] 를 사용하였다.

[표 2] ‘양진호’와 자주 같이 사용된 단어 와 빈도수

|  |
| --- |
| Chosun  [['폭행', 21], ['위디스크', 16], ['경찰', 10], ['영상', 5], ['불법', 4], ['압수수색', 4], ['음란물', 4], ['수사', 2]]  Kyunghyang  [['폭행', 9], ['갑질', 6], ['경찰', 4], ['위디스크', 3], ['영상', 3], ['괴롭힘', 2], ['상사', 1], ['압수수색', 1], ['수사', 1], ['음란물', 1], ['불법', 1]]  Hankyoreh  [['폭행', 10], ['위디스크', 9], ['불법', 8], ['경찰', 7], ['음란물', 4], ['수사', 4], ['영상', 4], ['압수수색', 1], ['검찰', 1]]  Donga  [['폭행', 21], ['위디스크', 11], ['영상', 9], ['압수수색', 9], ['경찰', 9], ['수사', 6], ['불법', 5], ['검찰', 4], ['음란물', 3]] |

지정한 노드들을 사용해서 각 노드들간의 네트워크를 시각화 해 보았다. 생성된 네트워크를 통해 각 언론사별로 사용한 단어들 간 어떠한 연관성이 있는지 확인 할 수 있다.

|  |  |
| --- | --- |
| **<조선일보>** | **<동아일보>** |
| **<한겨례>** | **<경향신문>** |

[그림 2] 신문사별 단어간 네트워크

R의 gcor() 함수를 사용해서 각 신문사별 네트워크의 상관관계를 구해보았다.

[표 3] 신문사 네트워크간 상관관계

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Netwk1** | 경향신문 | 경향신문 | 경향신문 | 한겨례 | 한겨례 | 조선일보 |
| **Netwk2** | 한겨례 | 조선일보 | 동아일보 | 조선일보 | 동아일보 | 동아일보 |
| **Corr** | 0.4440803 | **0.4064618** | 0.4811751 | 0.6806986 | **0.8217923** | 0.7435025 |

**5. 결과해석**

기존에 동아일보와 조선일보의 경우 보수언론으로, 한겨례와 경향신문의 경우 진보언론으로 많이 알려져 있다. 따라서 이번 양진호 회장의 폭행사건에 대해서 보수언론의 경우 친 기업적인 색채를 갖고 주로 사실관계를 보도할 것이라고 예상되고, 진보언론의 경우 오너의 지나친 갑질에 대해서 사회적인 비판을 보여줄 것으로 예상되었다.

*5-1. 빈도분석 결과 해석*

빈도분석을 통해서 보았을때 모든 신문사에서 비교적 공통적으로 ‘양진호’ ‘회장’ ‘직원’ ‘폭행’ ‘했다’ 라는 단어가 많이 사용되어서 ‘양진호 회장이 직원을 폭행했다.’ 라는 사실을 보도한 것을 알 수 있었다.

조선일보의 경우 다른 언론사와 다르게 보여주는 빈도 단어가 ‘경찰’, ‘위반’, ‘한국미래기술’ 등이 있었다. 한겨례의 경우 ‘업체’ ‘필터’ ‘불법’ ‘영상물’ 단어가 다른 언론사와 다르게 많이 사용되었고, ‘양진호’라는 당사자의 이름이 다른 언론사에 비해서 적게 사용된 것이 확인되었다. 실제로 네트워크 분석시에 ‘양진호’라는 단어의 빈도수가 낮아서 네트워크를 구성하는 노드의 개수를 조정해야 했다 경향신문의 경우 ‘갑질’ ‘괴롭힘’ ‘상사’ ‘폭언’등의 단어가 많이 사용되었다. 동아일보의 경우 ‘사무실’ ‘압수수색’ 등의 용어가 많이 사용되었고 특이하게 ‘교수’라는 단어가 많이 사용되었다

경향신문의 경우 다른 언론사와 다르게 단순 폭행사건이 아니라 상사의 갑질에 대한 사건으로 이번 사건을 해석하고 있는 것을 알 수 있었다. 조선일보와 동아일보의 경우는 실제 사실관계를 보도하는데 중점을 두었고, 위반사실에 대해서 경찰조사가 진행되고 있으며 사무실 압수수색까지 사건이 경과하고 있다는 사실을 알 수 있었다.

빈도분석을 통해 보았을 때 경향신문의 경우에는 사건을 통해 사회적 문제를 표현하고자 하는 표현을 많이 사용하였고, 한겨례의 경우는 ‘양진호’라는 이름보다 ‘웹하드’라는 업체의 속성을 많이 언급 한 것으로 보아 개개인의 문제보다는 기업의 문제에 초점을 맞춘 것으로 보인다. 보수언론 조선일보와 동아일보의 경우에는 대체적으로 사건의 진행과정을 알려주는 표현을 많이 사용하였다.

*5-2. 네트워크 분석 결과 해석*

네트워크 분석을 통해서 양진호라는 단어와 함께 사용된 단어를 살펴보았을 때 가장 큰 특징은 경향일보의 경우 ‘갑질’에 초점을 맞췄다는 것이 분명하게 드러났다. 워드 클라우드를 통해 ‘갑질119’라는 갑질 신고기관도 많이 언급되어 있는 사실을 볼 수 있다. 네트워크 그래프를 보았을 때 ‘갑질’과 ‘괴롭힘’이라는 단어의 노드의 연결이 경향신문의 네트워크에서만 존재하는 부분도 확인 할 수 있다.

상관관계 분석의 경우 보수언론인 조선일보와 동아일보, 진보언론인 한겨례와 경향신문이 같은 집단내에서는 높은 상관관계를 보이고 다른 집단 간에는 낮은 상관관계를 보일 것으로 예상되었다. 그러나 실제 상관관계를 보았을 때 한겨례의 경우는 같은 진보언론인 경향신문과의 연관성보다 보수 언론들과의 상관관계가 더 높은 것으로 확인되었다.

*5-3. 결론*

빈도분석을 통해 높은 빈도로 보이는 단어들을 비교해 보았을 때 조선일보, 동아일보의 경우 사건의 수사진행사항을 보도하는 것으로 보였고, 한겨례는 ‘위디스크’라는 업체의 문제에 대해 보도하는 것으로 보였다. 또한 경향신문은 직장내 갑질에 초점을 맞추는 것으로 예상 할 수 있었다. 그러나 네트워크 분석을 통해서 네트워크간의 상관관계를 보았을 때 실제로 경향신문의 경우는 다른 3개의 언론사들과 연관관계가 낮아 ‘갑질’에 초점을 맞춰 보도했다는 것을 확인 할 수 있었지만, 한겨례의 경우는 분석 전에 예상되었던 결과 그리고 빈도분석을 통해 예상하였던 결과와 다르게 조선일보, 동아일보 등 보수언론과 비슷한 기사를 작성했다는 사실을 파악 할 수 있었다.