서론

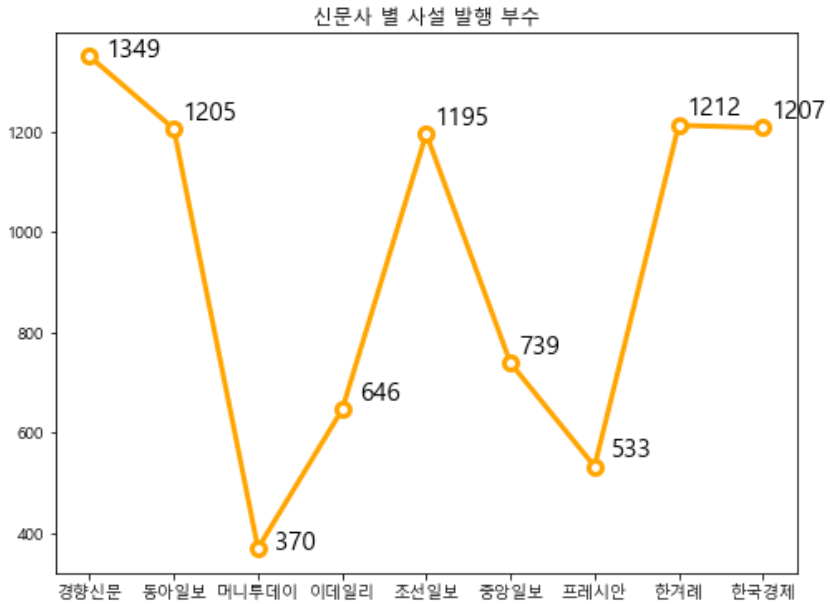
다루고자 하는 주제 :
신문사 성향 별 사건 보도 시각 차이 파악
같은 사건이라도 성향 별 시각 차이가 분명히 존재함을 알고 우리나라의 대표 신문사들이 대한민국의 사건들을 어떻게 표현하고 있는지 보수, 진보로 나누어 알아보고자 한다.

현황 및 문제점 :
현재 우리나라에는 종합지, 전문지, 경제지, 지방종합지, 인터넷뉴스 등 각기 다른 보수/중도/진도 성향의 수많은 신문사들이 존재한다. 스마트폰 하나로도 각종 신문을 찾아 볼 수 있는 이 시점에서 꼭 필요한 정보는 "신문을 선택하는 기준"이라고 생각된다. 성향이 강한 특정 신문사의 기사만 보게 된다면 같은 사건이라도 보도하는 방식이 달라 사실이 왜곡되거나 시각이 편협해 진다는 문제점이 존재한다. 따라서 대표적인 우리나라의 신문사들의 성향에 대해 알아보고 성향 별로 사건을 어떻게 다루는지 시각 차이를 알아보고자 한다.

우리나라 신문사 종류 :
https://ko.wikipedia.org/wiki/%ED%95%9C%EA%B5%AD%EC%96%B4_%EC%8B%A0%EB%AC%B8_%EB%AA%A9%EB%A1%9D

데이터 수집

전국 신문 발행 부수 순위와 신문사 별 성향 분류 웹 사이트 등을 바탕으로 총 9개의 신문사를 선정하였다. 사설이 일반 기사보다 해당 신문사의 주장이나 의견을 더 잘 반영할 것이라고 판단하였기 때문에 성향이 뚜렷한 순서대로 9개의 신문사의 사설들을 수집하였다. 다양한 종류의 기사를 수집하기 위해 종합지뿐만 아니라 경제지와 인터넷뉴스를 포함하여 스크래핑 하였고, 그 결과 보수 신문사는 조선일보, 중앙일보, 동아일보, 한국경제를 선정하였고 진보 신문사는 한겨레, 프레시안, 경향신문, 머니투데이, 이데일리를 선정하였다.



주로 수업시간에 배운 Selenium을 사용해서 웹 스크래핑을 하였고 각 신문사마다 엑셀 형태로 데이터를 입력하고 저장하는 방법을 사용하였다. 추가적으로 페이지 넘기기 자동화하는 방법과 동적 페이지에서 웹페이지 주소가 변경되지는 않아도 보이는 웹 페이지가 변경될 수 있는 방법 (ex.더보기) 등을 구현하였다. 또한 사설 기사가 따로 있지 않은 페이지들은 일일히 사설 기사를 분류하는 작업을 하였으며 try-except문을 사용해서 예외처리를 해주었다.

데이터 수집 기간 : 2019.01.01 ~ 2021.04.20

신문사 별 성향 분류 : http://jnjl.kr/VIS_bbs/board.php?bo_table=s1_1&wr_id=2620

https://www.chosun.com/opinion/editorial/ (조선일보)
https://news.joins.com/find/list?sourcegrouptype=all&keyword=%EC%82%AC%EC%84%A4&scopetype=Title&sourcecode=1%2C61&servicecode=20%2C34&display=%EC%82%AC%EC%84%A4 (중앙일보)
https://www.donga.com/news/Series/70040100000001 (동아일보)
https://www.hankyung.com/opinion/0001 (한국경제)
http://www.hani.co.kr/arti/opinion/editorial/home01.html (한겨레)
https://m.pressian.com/m/pages/opinion-column?page=1 (프레시안)
https://m.khan.co.kr/list.html?type=opinion (경향신문)
https://news.mt.co.kr/column/opinion_inside.html?category=03 (머니투데이)
https://www.edaily.co.kr/opinion/editorial (이데일리)

전처리

❖ 데이터 프레임 형식으로 맞추기

엑셀로 저장한 기사들을 새롭게 열이름을 설정하고 새로운 신문사 열을 만들어 모두 붙여주었다. 또한 보수인 신문사들을 1, 진보인 신문사들을 0으로 데이터 라벨링을 해주고 신문사별로 데이터를 섞어서 저장을 해주었다. 특히 대부분의 사설 기사 앞에는 '[사설]'이 붙어 있기 때문에 분석에 영향을 미치지 않도록 없애 주었다.

❖ 형태소 분석

kiwi, konlpy, stanza 패키지를 모두 사용해 본 결과 stanza가 가장 형태소 분석이 잘됐기 때문에 stanza 패키지를 사용해서 형태소 분석을 진행했다. 기사 제목을 스크래핑 한 것이기 때문에 명사만을 추출해서 단어 문서 행렬을 만들었다.

❖ 불용어 처리

같은 내용을 다르게 보도하는 시각 차이를 보고자 하는 것이기 때문에 단순히 모든 기사에 많이 들어가는 단어는 제외해주는 불용어 처리를 해주었다. 전체 신문사에 많이 나오는 단어/보수 신문사에 많이 나오는 단어/진보 신문사에 많이 나오는 단어를 각각 30개씩 뽑은 다음, 두 성향에 모두 많이 나오는 단어(정부, 국민, 대통령), 의미 없는 단어(것, 만, 년, 수, 때, 일, 차, 1, 2, 3), 잘못 분석된 단어(코로, 당코로), 기자 이름(오동희)을 제외해주었다.

단어 빈도

불용어 처리

보수

200	것	263
1519	정부	221
323	국민	191
548	대통령	149
82	文	139
1511	정권	139
100	與	136
390	기업	135
627	만	130
1777	코로	126
449	년	114
1067	수사	108
751	방역	103
545	대책	102
965	선거	98
228	경제	97
68	北	95
1896	한국	94
1365	일	93
2	1	92

진보

1775	코로	183
180	것	123
1525	정부	121
746	방역	109
476	당	107
561	대책	106
426	년	104
312	국민	94
1867	필요	90
566	대통령	86
1043	수	84
630	때	83
11	2	82
1673	차	79
177	검찰	78
695	미국	76
323	국회	76
1278	위기	74
846	북한	73
1048	수사	72

보수

86	文	139
1502	정권	139
103	與	136
393	기업	135
1057	수사	108
749	방역	103
545	대책	102
963	선거	98
230	경제	97
75	北	95
1884	한국	94
1521	정책	89
198	검찰	85
819	부동산	85
1262	위기	84
760	백신	81
1523	정치	79
351	규제	77
1766	코로나	77
1181	여당	76
1315	의혹	72
488	당	70

진보

748	방역	109
471	당	107
555	대책	106
1872	필요	90
173	검찰	78
324	국회	76
698	미국	76
1285	위기	74
844	북한	73
1053	수사	72
767	백신	70
713	민주당	69
1777	코로나	67
1611	중국	65
1440	재난	64
291	광화문	63
686	문	63
1324	의	61
550	대응	59
1094	시장	59

불용어를 처리해준 뒤, 성향 별로 많이 나오는 단어 빈도와 단어 구름을 만들어 비교해 보니 차이가 뚜렷이 나타난 것을 확인할 수 있었다.

단어 빈도

보수 신문사 단어구름

진보 신문사 단어구름

전체 신문사, 보수 신문사, 진보 신문사 별로 CountVectorizer함수를 통해 단어구름을 만들어 본 결과, 보수 신문사에서 상대적으로 "文", "與", "정권" 등의 단어가 많이 나왔고, 진보 신문사에서 "방역", "필요", "당" 등의 단어가 상대적으로 많이 나왔다는 것을 알 수 있었다. 사설 기사는 보통 비판적인 글이 대다수이기 때문에 보수 신문사 쪽에는 현재 여당인 문재인 대통령을 지칭하는 단어가 많이 나온 것으로 추측된다. 진보 신문사 쪽에서는 현재 정부와 관련이 깊지 않은 대중적인 단어인 단어가 많이 나온 것으로 보인다. 추후 감성분석에서 가중치 분석을 했을 때 이런 단어들을 더욱 주의를 기울여 살펴볼 수 있었다.

감성 분석

❖ 가중치가 양수로 큰 단어

보수가 1이고 진보가 0임을 알고 봤을 때 가중치가 양수로 큰 단어는 현 정부와 당을 지칭하는 단어가 많이 나왔다. 이는 앞서 말했듯 사설은 보통 비판하는 내용을 많이 쓰기 때문에 보수 쪽에서 진보 쪽을 비판하는 글을 많이 썼다는 것을 알 수 있다. 특히 가중치가 양수로 큰 단어에는 한자가 유독 많은 경향을 보였는데 이 이유는 보수 신문사인 조선일보와 동아일보에서 한자를 많이 사용했기 때문으로 보인다.

❖ 가중치가 음수로 큰 단어

반면, 가중치가 음수로 큰 단어는 실제로 행동하는 느낌의 단어가 많이 나온 것을 확인할 수 있었다. 진보 쪽 언론사는 현재 정권을 지지하기 때문에 현재의 상황을 있는 그대로 보도하기 위해 이런 단어를 많이 쓴 것으로 보인다.

	토큰	가중치
336	권력	0.166184
1547	조국	0.167447
94	韓	0.167941
72	尹	0.174115
91	軍	0.179579
204	경제	0.185971
85	美	0.189270
989	세금	0.196959
92	野	0.199023
65	中	0.205583
627	만	0.208804
182	것	0.220556
80	檢	0.222096
413	나라	0.227007
386	기업	0.234148
93	靑	0.265332
69	北	0.273245
1506	정권	0.296535
88	與	0.325458
75	文	0.350436

	토큰	가중치
291	광화문	-0.309843
1771	코로나19	-0.257846
74	思見	-0.244947
514	당코로	-0.243628
1206	오등희	-0.217065
1871	필요	-0.191358
442	노동자	-0.183245
1727	총장	-0.180416
1896	한반도	-0.180136
1905	합의	-0.178790
213	계기	-0.178045
684	미국	-0.173259
741	방위비	-0.170713
1061	수칙	-0.162751
1054	수업	-0.161445
698	민주당	-0.158269
1787	택배	-0.158212
1991	후퇴	-0.157477
1109	실천	-0.157461
1113	실행	-0.155743

감성 분석

감성분석을 진행한 결과 활성함수는 sigmoid, optimizer는 adam, epoch은 3을 돌린 것이 train accuracy가 0.7262, test accuracy가 0.6673로 가장 높은 성능이 나왔다. 가중치 분석을 한 결과, X가 증가할 수록 Y도 증가하는 단어는 대표적으로 "文", "與", "정권", "靑", "조국" 등이 있었고 X가 증가할 수록 Y가 감소하는 단어는 대표적으로 "합의", "계기", "실천", "후퇴", "실행" 등이 있었다. 특히 가중치가 양수로 가장 높은 "文"과 음수로 가장 낮은 "광화문"을 가지고 주제분석을 하고자 하는 인사이트를 얻게 되었다.

희소행렬로 변환해서 학습을 시켰을 때는 0.7726으로 기존의 CSR방식보다 COO방식이 0.05가량 높아졌다는 사실을 알 수 있었고 val_accuracy를 기준으로 EarlyStopping을 하는 것보다 val_loss를 기준으로 EarlyStopping을 하는 것이 훨씬 학습량이 많았기 때문에 val_loss를 더 중요한 지표로 두고 학습을 시켰다.

감성분석이 0과 1을 맞추는 문제이기 때문에 추가적으로 분류에 성능이 좋은 MLPClassifier, SVC, DecisionTreeClassifier를 사용해서 모델을 돌려본 결과 MLPClassifier의 Accuracy가 0.6879, SVC의 Accuracy가 0.6985로 매우 좋은 성능을 보였다.

주제 분석

LSA와 NMF, LDA 중 LSA에 회전을 적용한 뒤 해석하는 방법이 가장 뚜렷하게 의미를 도출해냈기 때문에 LSA를 사용해서 주제 분석을 하였다.

감성 분석에서 보수 쪽 성향이 뚜렷하게 나타났던 단어인 "與"을 가지고 주제 분석을 해 본 결과 "비리", "불안", "혈세", "무능", "참담", "고집" 등 매우 부정정인 단어들이 많이 나온 것을 확인 할 수 있었다. 또한 신문사 별로 "與"가 나온 비중을 살펴본 결과 조선일보, 동아일보, 한국경제 등 보수 쪽의 신문사에서 많이 나왔다는 것을 알 수 있다. 이 말은 보수성향의 신문사에서 여당을 매우 좋지 않게 평가 했다는 것을 의미한다.

	word	loading
228	고집	0.041270
1041	수도권	0.041854
1943	협치	0.043357
400	김학	0.044707
153	개편	0.046625
694	민심	0.048206
1684	참담	0.048842
687	미얀마	0.049103
655	무능	0.049252
1550	조선	0.053977
1936	혈세	0.054961
1126	아이	0.055114
1662	집단	0.055668
138	강화'	0.058781
77	核	0.059566
383	기업	0.067496
845	불안	0.067967
177	검증	0.069436
861	비리	0.127827
89	軍	0.897666

신문사	與
경향신문	0.004219
동아일보	0.024034
머니투데이	0.004822
이데일리	0.004877
조선일보	0.045724
중앙일보	0.008831
프레시안	0.008374
한겨레	0.003412
한국경제	0.014984

주제 분석

	word	loading
1331	이대로	0.062179
1339	이상직	0.062273
319	국정원	0.064701
1105	실장	0.066514
1594	중국	0.070273
202	경제	0.071253
785	보완책	0.072286
94	韓美	0.072634
1506	정권	0.072634
140	개각	0.073886
353	글로벌	0.086316
1471	전력	0.087314
849	불허	0.091630
539	대응'	0.108861
228	고집	0.115699
150	개정	0.116875
1359	인구감소	0.125201
1799	투기	0.145928
590	뒤	0.209193
1058	수준	0.616538

또한 보수 쪽 성향이 뚜렷하게 나타났던 단어인 "정권"을 가지고 주제 분석을 해 본 결과 "수준", "투기", "인구감소", "고집", "불허" 등 매우 부정정인 단어들이 많이 나온 것을 역시 확인해 볼 수 있었다. 이 또한 보수성향의 신문사에서 현재 정권을 매우 좋지 않게 평가 했다는 것을 의미한다.

반면, 감성 분석에서 진보 쪽 성향이 뚜렷하게 나타났던 단어인 "광화문"을 가지고 주제 분석을 해 본 결과 프레시안, 이데일리, 경향신문 등 진보 쪽의 신문사에서 이 단어가 많이 등장했음을 알 수 있었다.

	광화문
신문사	
경향신문	0.018268
동아일보	0.007764
머니투데이	0.002195
이데일리	0.018273
조선일보	0.009904
중앙일보	0.004760
프레시안	0.032524
한겨레	0.009113
한국경제	0.007895

결론

신문사 성향 별 사건 보도 시각 차이를 파악해 본 결과 확실히 보수 성향의 신문사들과 진보 성향의 신문사들은 사용하는 단어가 많이 달랐다.

보수 성향의 신문사(조선일보, 중앙일보, 동아일보, 한국경제)들은 현재 정권을 매우 좋지 않게 보고 있기 때문에 부정적이고 자극적인 단어를 매우 많이 사용해 안 좋게 평가했다. 이런 성향은 종합지 뿐만 아니라 경제지에서도 확연하게 나타났으며 진보 성향의 신문사보다 훨씬 뚜렷하게 나타났다.

반면, 진보 성향의 신문사(한겨레, 프레시안, 경향신문, 머니투데이, 이데일리)들은 현재 정권을 지지하기 때문에 사건을 사실 그대로, 적극적으로 조치를 취하는 느낌의 단어가 많이 나왔다. 이는 현재의 상황을 잘 헤쳐나가고 있다는 느낌을 주기 위한 것으로 해석된다.

이를 통해 매우 많은 기사들은 만연하는 현재, 특정 성향이 강한 신문사만 보면 안된다는 결론을 도출할 수 있었다. 같은 기간에 같은 양의 기사를 모아 분석한 데이터가 이렇게 다를 수 있다는 것은 분명 신문사의 성향마다 사건을 보도하는 시각 차이가 존재한다는 것이고 우리는 이를 간과하지 말고 여러 신문사의 기사를 읽으며 시야를 넓혀가야 한다는 것을 알 수 있다.