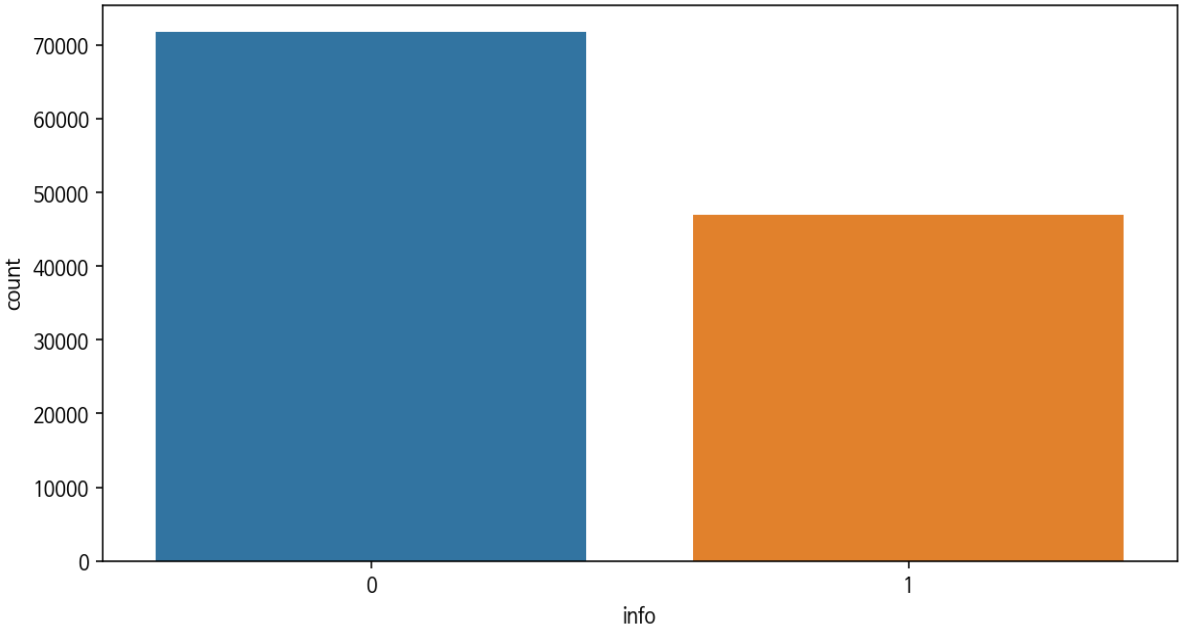
EDA

[코드공유실습] 진짜 뉴스와 가짜뉴스의 갯수



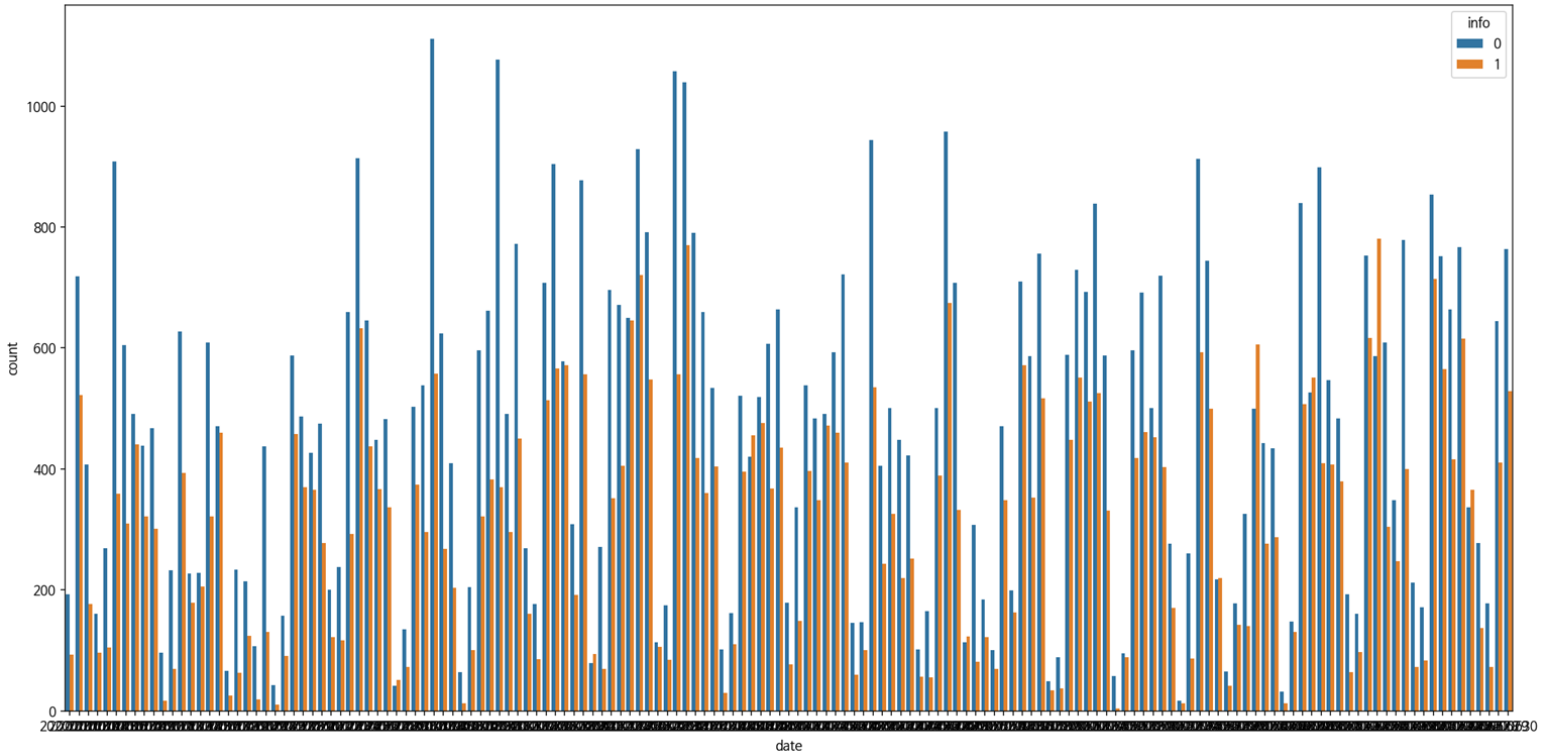
진짜 뉴스 개수 : 71813

가짜 뉴스 개수 : 46932

진짜 뉴스 비율 : 60.477%

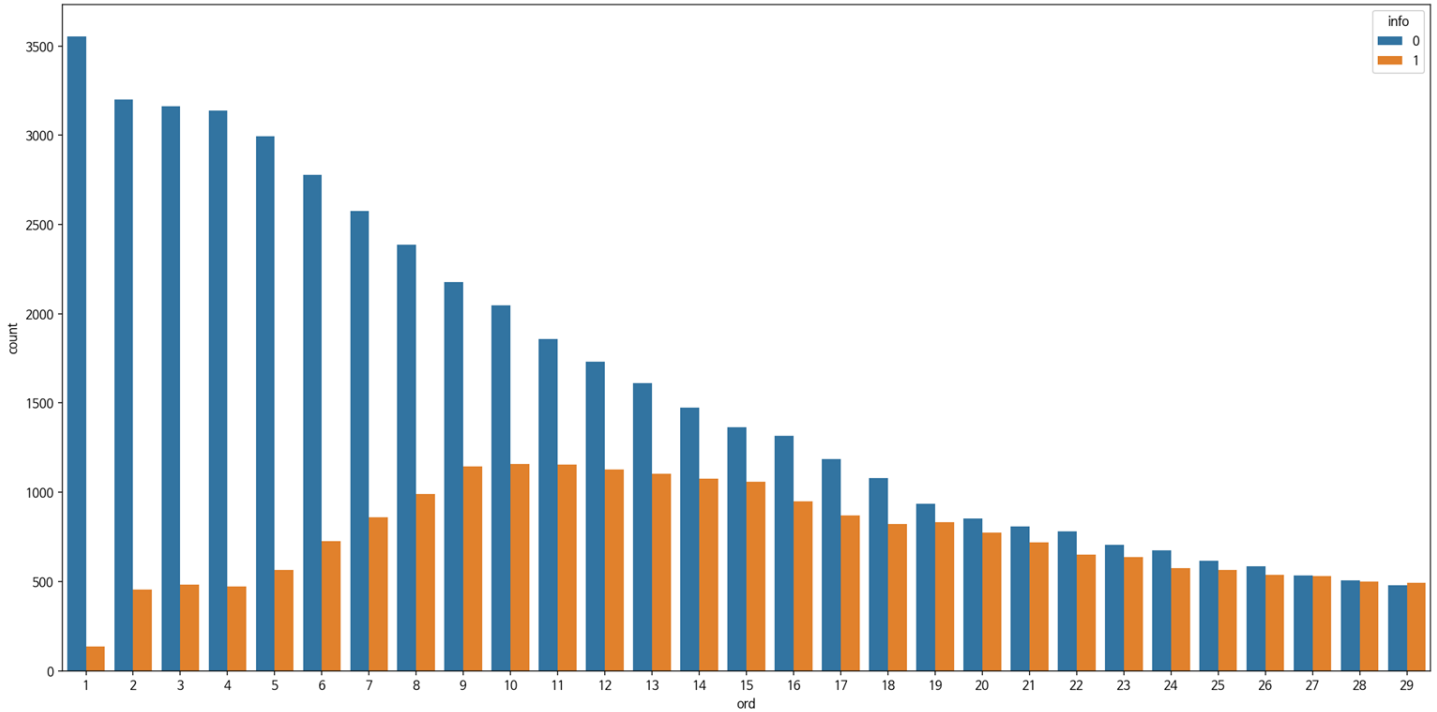
가짜 뉴스 비율 : 39.523%

[코드공유실습] 날짜별 진짜 뉴스와 가짜뉴스 갯수



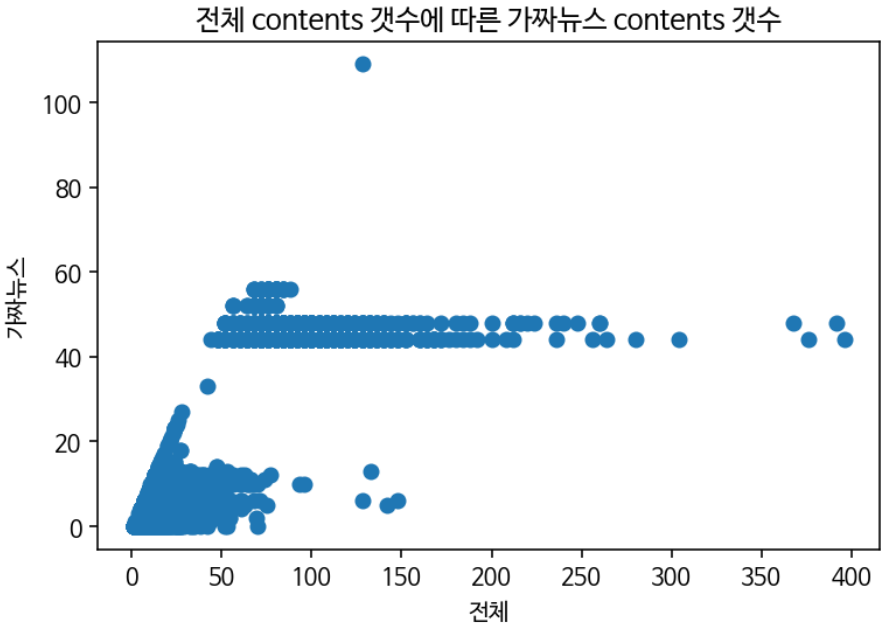
EDA 결과, 날짜는 별 상관이 없는 것으로 보인다.

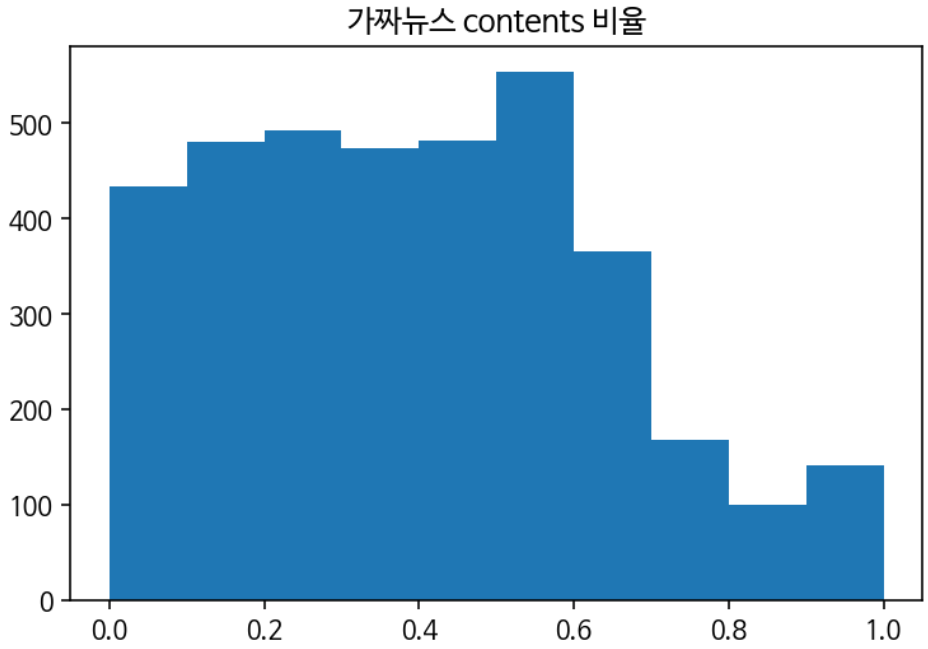
[코드공유실습] 뉴스 contents 갯수에 따른 가짜뉴스 비율



뉴스 길이가 증가할수록 진짜 뉴스는 줄어들고 가짜뉴스가 증가하는 것처럼 보이는데 이는 데이터 제공에서의 문제이지 않을까 추측.

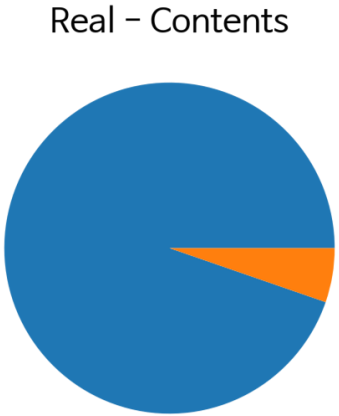
뉴스 contents 갯수에 따른 가짜뉴스 비율

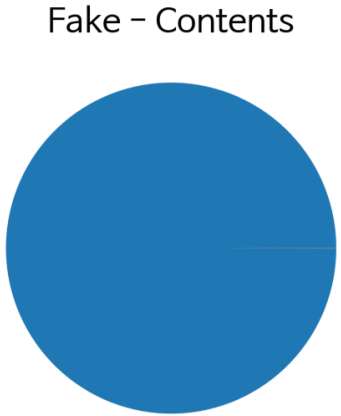


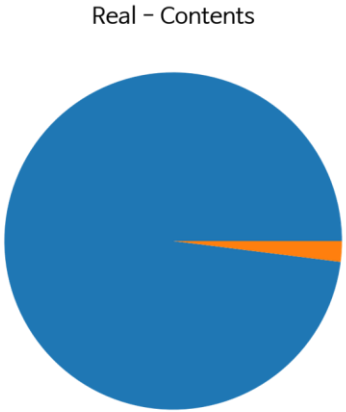


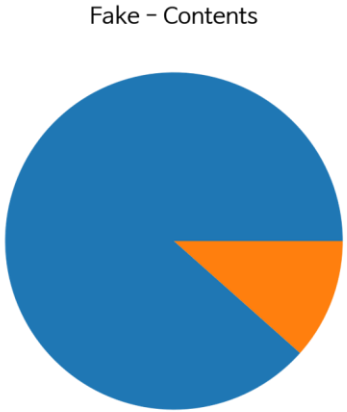
보통 가짜뉴스의 비율이 0~0.5 정도 차지

[코드공유실습] 시작과 끝 문자



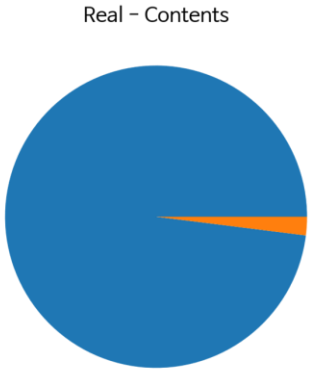


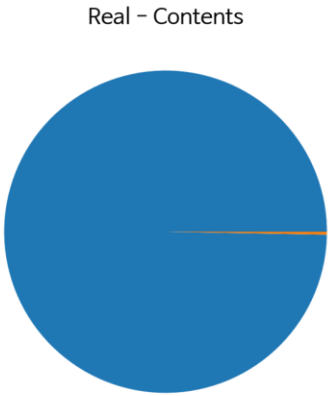


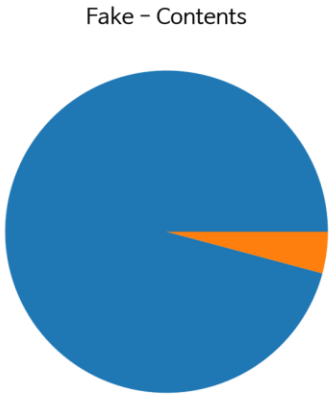


- [또는 (로 시작하는 경우 : 진짜뉴스에서 상대적으로 높은 비율을 보임. 가짜 뉴스의 contents는 거의 없음.
-] 또는) 로 끝나는 경우 : 가짜 뉴스의 contents에서 압도적으로 높은 비율을 보임

[코드공유실습] 시작과 끝 문자

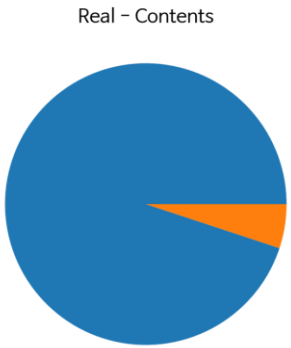


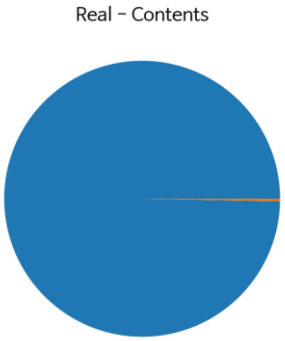


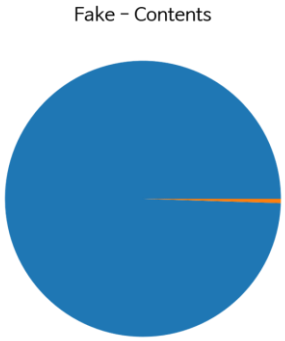


- ' 또는 "로 시작하는 경우 : 가짜 뉴스의 contents에서 압도적으로 높은 비율을 보임
- ' 또는 "로 끝나는 경우 : 가짜 뉴스에서의 비율이 좀 더 높음. 진짜뉴스는 거의 없음

[코드공유실습] 시작과 끝 문자



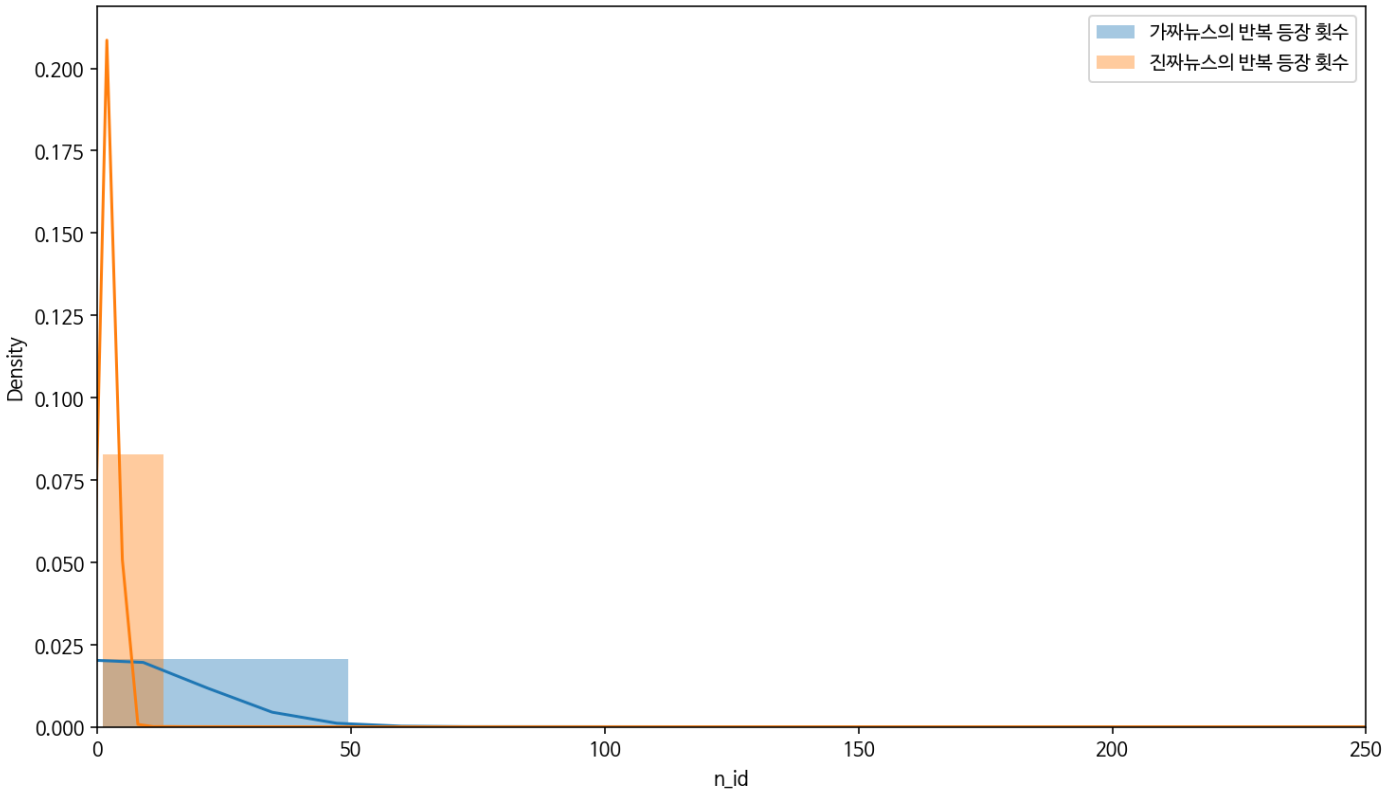




- 숫자 0~9로 시작하는 경우 : 가짜뉴스의 contents에서 압도적으로 높음
- 숫자 0~9로 끝나는 경우 : 모든 경우에서 별로 없음

결론 : 진짜뉴스의 경우 가짜뉴스보다 상대적으로 특수문자로 시작하거나 끝나는 경우가 적고, 그나마 있어도 숫자로 시작하거나 [로 시작하는 경우이다.

같은 content의 반복 등장 횟수



가짜뉴스가 반복 등장하는 경우가 훨씬 많다.

같은 content의 반복 등장 횟수 - 분포

```
repe_fake['n_id'].describe()
```

```
count     5794.000000
mean        8.104936
std       104.695503
min         1.000000
25%         1.000000
50%         1.000000
75%         1.000000
max      2420.000000
Name: n_id, dtype: float64
```

```
repe_real['n_id'].describe()
```

```
count    40367.000000
mean         1.778309
std          6.045905
min          1.000000
25%          1.000000
50%          1.000000
75%          1.000000
max        605.000000
Name: n_id, dtype: float64
```

가짜뉴스가 반복 등장하는 경우가 훨씬 많다.

가짜뉴스에서 자주 등장하는 단어

'종목': 9140, '가능': 7288, '공개': 5495, '한국': 5483, '추천': 5464, '주': 5061, '목표': 4765, '상한': 4740, '이상': 4468, '무료': 4168, '주식': 4091, '바이오': 3696, '금리': 3504, '테마': 3498, '최저': 3478, '방': 3461, '탁론': 3295, '카톡': 3102, '실적': 3079, '이용': 3029, '투자': 2997, '신용': 2967, '수익': 2962, '미수': 2939, '환': 2926, '혜주': 2912, '급등': 2873, '오늘': 2788, '대장': 2784, '당장': 2778, '매집': 2743, '젠': 2559, '줄': 2526, '로': 2520, '배': 2503, '효과': 2479, '수익률': 2475, '돈': 2474, '재료': 2474, '평가': 2463, '스': 2455, '업계': 2449, '지금': 2437, '시대': 2435, '라면': 2432, '연결': 2428, '인터넷': 2427, '레버리지': 2426, '소비자': 2423, '영웅': 2422, '박주': 2421, '똑똑해진': 2420, '소형차': 2420, '대중': 2420, '역대': 2340, '최종': 2302, '거래': 2296, '내일': 2293, '핵심': 2293, '클릭': 2277, '모집': 2261, '분': 2238, '끝': 2208, '다시': 2207, '책임': 2198, '여기': 2197, '주식시장': 2197, '도전': 2195, '사면': 2190, '정치': 2190, '역사': 2187, '니': 2185, '량': 2183, '가도': 2181, '인맥': 2180, '안집': 2180, '정부': 2179, '정책': 2080, '긴급': 2077, '코로나': 1861, '확인': 1529, '상반기': 1487, '관련': 1416, '수수료': 1375, '바로': 1339, '예상': 1176, '최대': 1147, '공략': 1082, '체험': 973, '바이러스': 934, '추가': 914, '부터': 911, '임박': 855, '것': 842, '직전': 842, '명': 838, '장주': 729, '유망': 721, '수': 698, '집중': 681}

‘종목’, ‘주’, ‘상한’, ‘주식’, ‘테마’ 등 주식 관련 단어, ‘추천’, ‘무료’ 등 광고성 단어

진짜뉴스에서 자주 등장하는 단어

'등': 11104, '것': 9595, '있다': 7371, '수': 6988, '이': 6105, '코로나': 5513, '재': 4343, '명': 4031, '및': 3679, '있는': 3488, '기자': 3447, '경제': 3173, '위': 3058, '말': 3046, '한국': 3032, '위해': 2920, '사업': 2791, '기업': 2781, '통해': 2588, '전': 2537, '지원': 2508, '관련': 2482, '지역': 2480, '를': 2475, '투자': 2397, '지난': 2334, '로': 2276, '대한': 2231, '중': 2216, '시장': 2201, '개': 2186, '최근': 2174, '거래': 2157, '금지': 2151, '정부': 2109, '기준': 2020, '이번': 1996, '현재': 1995, '대비': 1992, '배포': 1978, '무단': 1945, '날': 1936, '서울': 1865, '미국': 1861, '대해': 1829, '개발': 1795, '달': 1771, '헤럴드경제': 1752, '그': 1732, '계획': 1722, '의': 1680, '진행': 1657, '제공': 1655, '국내': 1636, '종목': 1636, '점': 1616, '이후': 1610, '금융': 1603, '기술': 1529, '올해': 1524, '고': 1523, '지난해': 1516, '예정': 1499, '내': 1472, '대표': 1472, '기록': 1470, '기사': 1459, '경우': 1445, '상황': 1440, '서비스': 1434, '상승': 1421, '중국': 1418, '정보': 1408, '확대': 1398, '뉴스': 1377, '증가': 1369, '산업': 1367, '확': 1365, '총': 1362, '글로벌': 1351, '대상': 1349, '진자': 1332, '더': 1329, '종합': 1324, '기관': 1311, '운영': 1307, '마스크': 1299, '때': 1295, '해당': 1288, '이상': 1279, '주가': 1275, '증권': 1269, '확산': 1262, '발생': 1256, '선': 1242, '관리': 1239, '분석': 1237, '관계자': 1220, '설명': 1219, '규모': 1218}

가짜뉴스와는 확연히 차이 난다.