트위터 데이터 분석을 통한 중요 사건 요약기

학과 : 컴퓨터학부

학번 : 20142350

이름 : 박선무



Fig. 1. 201401 ~ 201404 트위터 데이터 분석으로 그 기간동안 발생한 중요한 사건을 요약한 dataframe

문제 정의 및 모델링

트위터 데이터의 경우 Noise 가 많지만 데이터의 양도 방대하고 시국의 변화에 민감한 성질을 지녔다고 생각하였다. 또한 시계열 데이터 로써의 속성으로 해당 시기에 대한 정보도 존재하기 때문에 특정한 기간동안 주목할만한 이슈나 사건을 요약정리 할 수 있는 잠재성이 있다고 생각하였다. 따라서 해당 데이터를 분석해 사건을 요약정리하는데 사용한다.

트위터 데이터의 경우 200GB에 육박하는 상당히 거대한 데이터이고 따라서 주어진 컴퓨팅 환경에서 여러 데이터를 한번에 메모리에 실어서 활용할 수 없다(실제로 메모리가 많이 터졌습니다…). 따라서 우선 1시간 단위 데이터에서 keyword를 extraction한 후에 이데이터를 활용해서 문제에 접근을 할 생각이며 keyword extraction 후에는 상대적으로 데이터의 Noise가 적다고 생각하는 네이버 기사 데이터를 활용해 해당 사건과 관련된 기사데이터로 변환하고, 문서 요약에 자주 사용되는 TextRank 알고리즘을 적용하여 1줄로 요약해 데이터로 사용할 것이다.

USB자체의 용량제한으로 모든 파일을 한번에 압축 해제해서 사용할 수 없다. 따라서, 압축 해제, 데이터 전처리 후 keyword extraction까지 한후 pickle 라이브러리를 사용해 값을 저장한후 이미 사용한 데이터는 모두 삭제하는 방식으로 자동화 설계를 한다.

문제해결을 위한 도구 활용

1. 데이터 전처리

해당 프로젝트에서는 sns데이터에서 오직 text부분만 추출하여 이를 학습에 이용한다.

1-1 제일 먼저 pandas 의 read\_json을 이용해 dataframe에 data를 받아온뒤, text 부분만 추출한다.

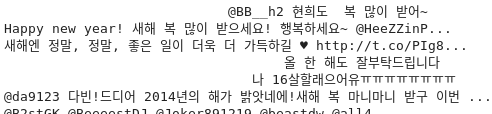


Fig. 2. data의 text 부분만 추출한 모습

1-2 keyword를 뽑아 내려하기 때문에 데이터의 noise를 제거하기위해 이모티콘, url, 리트윗등 을 제거하는 전처리 과정을 거친다.





Fig. 3. 데이터의 noise가 제거된 모습

1-3 keyword에서 핵심이 되는 부분은 단어에서 명사 부분이 생각하였고 Konlpy 라이브러리에서 속도와 성능면에도 잘 작동하는 mecab 라이브러리를 사용해 문장에서 명사를 뽑아낸다.



Fig. 4. sentence에서 명사만 추출된 모습

1-4 유의미한 단어만 학습에 이용하기 위해서 불용어를 지정하여 제거해준다.

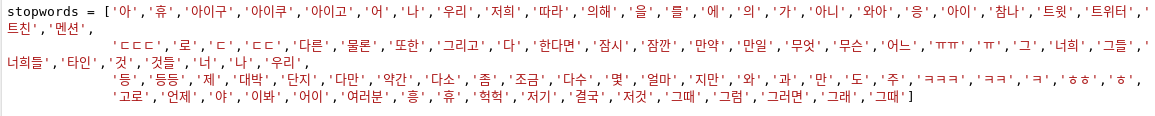


Fig. 5. 학습에 사용한 불용어들, Twitter의 특성을 제거하기 위해 ‘트친, ‘멘션’과 같은것도 추가한 모습이다.



Fig. 6. 최종적으로 학습에 이용될 데이터의 모습

2. 데이터 Keyword 화

2-1 keyword extracton

문서에 등장하는 단어의 빈도로 문서의 특징을 extraction할수 있을것이다. Sk-learn 패키지의 countvectorizer를 이용해 학습 할 수 있으며, min\_df와 max\_df 설정을 통해 총 문서에서 단어의 등장 빈도를 사용해 의미없을 수 있는 문장을 한번 더 걸러낸 후에, BOW형태의 Feature를 만들 수 있으며 이를 열방향으로 합할시 각 keyword의 sum이 도출된다. 한 countvectorizer 라이브러리에서 한 단어 뿐만아니라 2~3 어절의 단어까지 keyword로 사용가능하다. twitter데이터의 경우 noise가 많기 때문에 여러 어절의 단어도 keyword로 사용할 수 있도록 하였다.

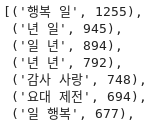


Fig. 7. 빈수가 높은 키워드 순으로 정렬한 모습이다.

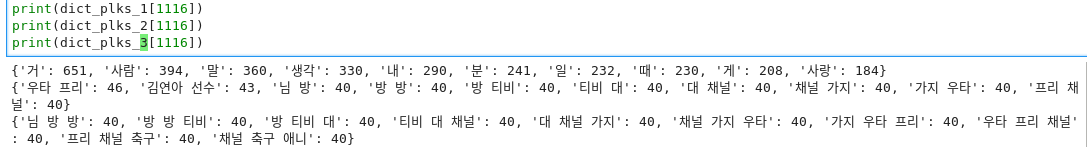


Fig. 8. 각 어절의 길이에 따른 키워드의 모습.

평균적으로 2어절 길이의 단어가 1,3 어절의 단어보다 조금더 의미있는 데이터를 담고 있음을 확인할 수 있다.

2-2 그룹화

현재 주어진 데이터는 1시간 단위로 주어져 있기 때문에 의미 있는 keyword 도출하기 힘들다 6시간 단위의 데이터로 그룹화하여 그 순간의 이슈를 포착할 수 있도록 한다.

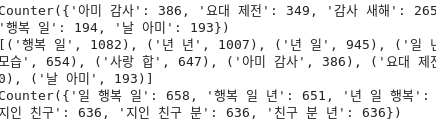


Fig. 9. 데이터를 합쳐 키워드의 수가 증가되고 고려된 시간의 범위도 넓어진 모습이다.

2-3 중요 keyword만 추출

주어진 keyword중 특정 수치 이상으로 사용된 keyword만 추출하여 해당 시간의 날짜와 함께 네이버 기사 검색에 사용할 수 있록 한다.

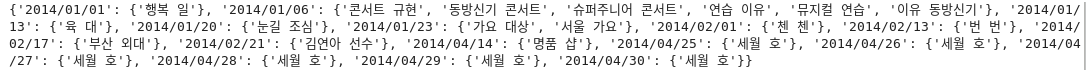


Fig. 10. 특정 수치 이상의 Keyword만 추출한 모습이다. 특정 Noise데이터만 제외하면 어느정도 해당시기의 이슈가 담겨있는 모습을 확인 할 수 있다.

3. 네이버 기사 크롤링

주어진 키워드와 날짜를 이용해 해당 날에 그 키워드로 나온 네이버 기사를 크롤링 한다.

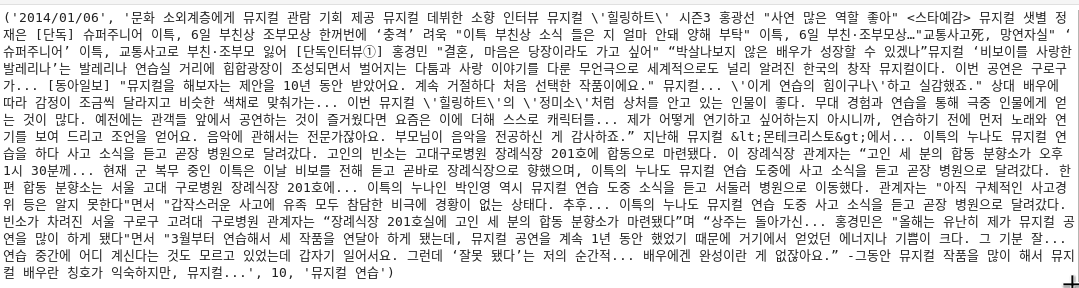


Fig. 11. 해당날짜에 해당키워드로 검색을 하였을때 나온 뉴스 데이터이다.

4. TextRank 알고리즘 적용으로 기사 요약

해당 뉴스와 관련된 이슈를 요약하기 위해 TextRank알고리즘을 사용한다. TextRank알고리즘은 문장이나 단어간의 인용비율을 통해 각 문장에 상대적 가중치를 적용하여 중요도가 높은 문장으로 요약할 수 있는 특징이 있다.



Fig. 12. 해당기사를 1줄로 요약한 버전이다.

이후 해당 날짜와 기사 요약본을 pandas dataframe화 하면 Fig. 1과 같은 dataframe이 완성된다.

USB환경에서 압축해제와 파일처리를 하였기때문에 그 과정에서 소요되는 시간이 너무 많아(한 tar파일 처리하는데 5시간 가량 소요) 201401 ~ 201404 데이터만을 사용하였다.

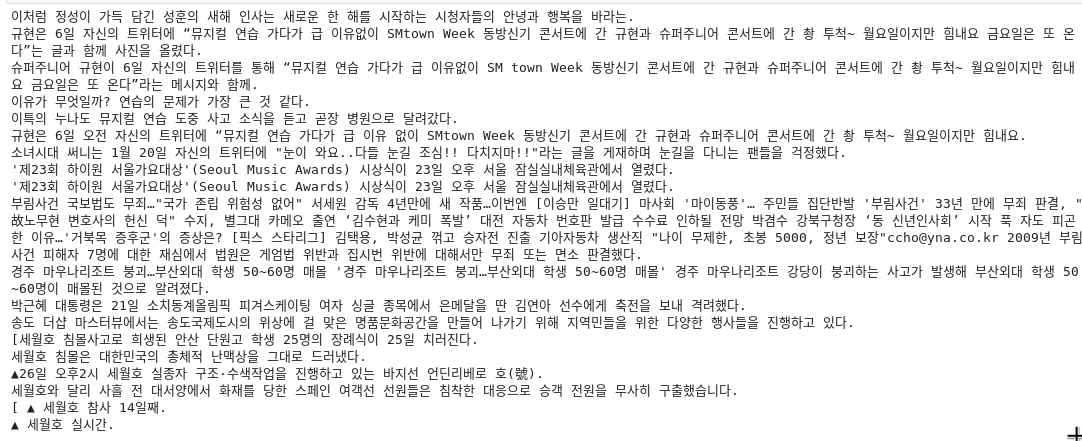


Fig. 13. 의미있는 키워드들로 요약된 모든 문장들

5. 자동화

200기가에 육박하는 데이터를 처리하기 위해 자동화 과정을 거친다.

0-1 tar파일 추출

glob 라이브러리를 사용해 tar이름을 가진 파일의 리스트 사용한다.

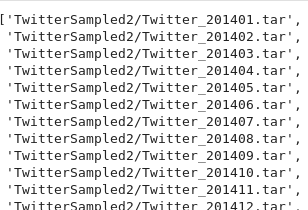


Fig. 14. 확장자가 tar인 파일만 추출한 모습

0-2 압축해제 & tar.gz파일 추출

tarfile라이브러리의 tarinfo기능을 활용해 내부 압축파일정보 가져오고 압축을 해제한다.

Tar.gz파일에도 마찬가지로 작동하며 사용이 종료될지 os.remove함수로 삭제를 하도록 하였다.

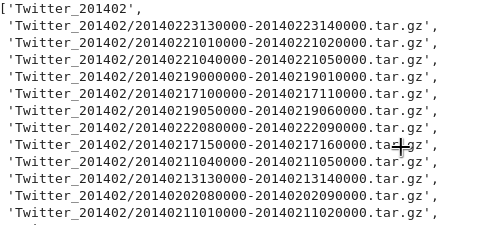


Fig. 15 해당하는 tar파일내부 파일의 정보

0-3 pickle라이브러리를 활용한 keyword 저장

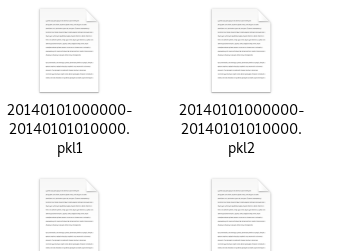


Fig. 16 pickle라이브러리를 활용해 저장된 keyword들

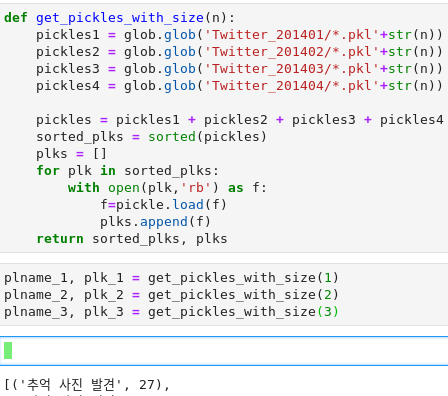


fig. 17. pickle 라이브러리를 활용해 load가 된 모습이다.