20130711 김기태

SNS 빅데이터 분석 기말 과제

목차

1. 문제 정의 및 예상 결과
2. 문제해결을 위한 방법 및 모델링
3. 실제 자료분석
4. 결론
5. **문제 정의 및 예상 결과**

저희 과목의 제목이자, 기말과제 자체를 나타내는 sns빅데이터 분석. 저는 이 제목 자체에 의문을 가지게 되었습니다. 과연 sns에서 넘쳐나는 그 수많은 글들 중 일부가 조금 바뀐다고, 그것이 여론을 대변할 수 있을지 확인을 하고 싶어졌습니다.

실제로 sns는 사람들이 특정한 목적을 가지고 사용하는 경우도 많습니다.

일부는 특정 아이돌을 지지하고 팬들과 소통하기 위해서, 또한 그 글들을 rt를 하면서 글을 쓰고 쓰는, 일명 팬질을 위해 sns계정을 만들어 활동을 하는 경우도 있고, 일기장처럼 쓰기 위해서 사용하는 사람도 있을 뿐만 아니라, 자신의 회사를 홍보하기 위해, 자신의 생각을 주장하기 위해, 등등 여러 목적을 가지고 사람들은 sns를 사용한다고 볼 수 있습니다.

우리는 여론을 파악할 때, 특정 단어군의 사용 빈도가 확연히 늘어나면 그러한 여론이 생겨나고 있다고 판단할 수 있습니다. 하지만, 앞에서 설명한 sns들은 항상 비슷한, 혹은 같은 단어들을 꾸준히 사용해 높은 사용 빈도를 나타낼 것이고, 이것은 현재 여론을 나타낸다고 말하기엔 좋지 않을것않을 것 이라고 판단됩니다.

따라서 저는 일상적으로 사용하는 단어들의 빈도를 파악해둘수 있다면, 이러한 단어들이 현재 여론에 따라 생기는것인지, 아니면 현재 일어나는 사회문제에 영향을 받아 생겨난 단어들인지 파악 할 수 있을것이라고 생각했습니다.

따라서 저는 이번 과제로 받은 데이터들을 활용해 특정 달, 일에 한해 항상 자주 쓰이는 단어들, 여론과는 상관없이 꾸준히 쓰이는 단어들을 분류해 내고, 그러한 단어들의 경향성이나 특징들을 골라낼 수 있도록 하는 것을 목표로 삼았습니다.

**예상되는 단어들의 특징 및 경향성:**

1. 다른 년도 같은 월,일의 경우 비슷한 단어가 항상 사용될 가능성이 높다.

EX) 1월 1일의 새해복 많이받으세요, 8월 15일 광복절, 12월 25일 메리 크리스마스, 등등

1. 평소에도 일반적으로 쓰는 단어일수 있다.

EX) 트위터, 미국, 안녕하세요, 등등

이때 이러한 단어들은 평소에도 어느 정도 쓰이다가 특정 뉴스에 의해 횟수가 폭발적으로 증가할 수 있다. 따라서 이러한 자료는 기준점 정도가 되지 않을까 싶다.

1. 특정 유명인사의 인지도에 따라 꾸준히 언급될 가능성도 고려해봐야 한다.

EX) EXO, 방탄소년단, 오빠, 악동뮤지션, 한국당 총재

역시 2번과 같이 폭발적인 증가가 있을 수 있다.

1. 유행어, 혹은 인터넷 특유의 유행에 따라 여론과 상관없이 자주 쓰이는 단어도 있다.

EX) 앙기모띠, UBD, Yee~, (대충 과제물에서 예시 생각해내는 트윗)

1. **문제해결을 위한 방법 및 모델링**

이때 우리는 문제해결을 위해 방법을 생각해야 되는데, 단순히 특정 단어가 많이 쓰이고 말고를 알아보는 일에 대해서는 결국 전체 데이터에서 빈도를 분석해 월별, 일별로 구분해서 비교를 해보는 수밖에 없다고 판단이 들었습니다. 따라서 전체 데이터에서 텍스트들을 형태소 분석, 단어들의 개수를 세어 각 기준별로 단어들의 수를 확인 할 수 있도록 하는 것이 문제해결을 위한 첫 걸음이자 모델링의 기본이 되었습니다.

따라서 우선 받은 800기가에 달하는 큰 데이터들이 저장되어 있는 형태를 살펴보니, 매일 만개를 기준으로 짤라서 텍스트에 json형식으로 저장해둔 것을 확인했습니다. 너무 큰 데이터를 매번 부르는 것은 너무 비효율적이기 때문에 각각의 텍스트를 불러 해당 텍스트 내에서의 모든 트윗의 text를 형태소 분석하여 date값을 주어준 후 새로운 json파일에 저장하였습니다.

저장한 데이터 형식 : {‘date’:’20140101’, ‘data’: ~~

이러한 json들을 한 텍스트당 한줄로 채워가며 완성한 words.json파일의 총 예상 용량은 2GB가 되었습니다.

이때 예시에서 사용하신 꼬꼬마 형태소 분석기 kkma()의 경우 좀 더 꼼꼼하고 분류 효과가 높긴 하지만, 속도가 너무 느리고 대략 5만번의 형태소분석 도중, 분석을 멈추고 cpu가 터져버리려고 하는 오류가 발생해 사용하지 않았고, 인터넷 검색을 통해 정확도는 조금 떨어지지만, 속도가 매우 빠른 komoran 분석기를 사용하였습니다. 이때 분석 자체를 실패하는 경우도 조금 많아졌지만 체감 속도가 10배 이상 빨라졌습니다.

실제로 데이터 분석의 경우, 정확도가 떨어지더라도 빠른 komoran분석기를 사용했음에도 불구하고, 하루 종일 intel i7-7700을 자랑하는 제 cpu가 multi process없이 30퍼를 혼자 사용했는데도 하루 종일 돌려서 대략 8개월 정도를 줄여내었고, multi process를 활용해서 줄이는 코드를 짜보려 하였지만, 제데로 분석이 되지 않고 저장 또한 이상하게 되는 모습이 보여 포기했습니다.

결국 데이터를 줄이는데 있어 모든 데이터를 줄이는데는 실패하였고, 2014년 1월 ~ 8월 + 2015년 1월 ~ 5월 까지만 분류했습니다.(중간에 실수로 끊기고 그래서 이것도 실제 분류시간이 꽤 됩니다.)

결국 실제로 분류해낸 데이터는 1.17GB로 2014년 1월~8월 + 2015년 1월~5월 까지의 일별 단어수입니다.

이렇게 분류해낸 json파일을 다시 불러와 list 및 json형식으로 구분하였고 이 데이터들을 다시 불러와 실제 빈도분석을 시행하였습니다.

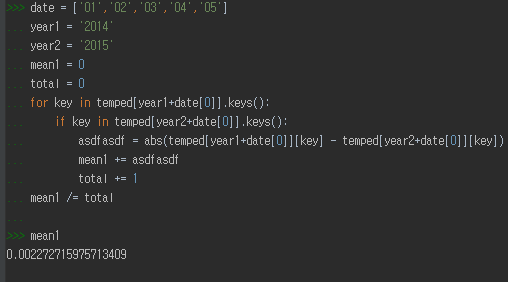
분류해낸 데이터들을 활용하여 문제해결을 위한 방법으로는 1p에 적은 예상에 따라

1. 같은 월, 같은 일일 때 년도의 차이가 단어 차이에 영향을 주지 않는 경우
2. 월, 일이 지났는데도 계속하여 동일한 비율을 차지하고 있는 경우

이러한 두가지 기준에 따라 단어를 분류해나가기로 하였습니다.

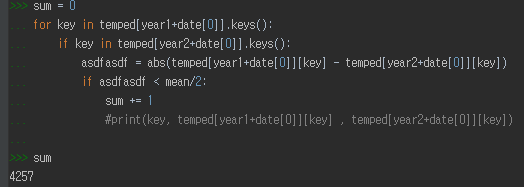
1. **실제 자료분석**

우선 첫번째 예상이었던 1월에 자주 쓰이는 단어에 대해 확인해 보기로 하였습니다.



여기서 temped는 연도별로 단어의 빈도json을 가지고 있는 json입니다.

빈도가 0.001%보다도 낮은 수들은 모두 합쳐서 빈도 분석을 했을 때 각각의 단어별로 따로 분석을 해야 하기 때문에 기존의 카이제곱테스트는 딱히 의미가 없고(월별로 통째로 비교하는 것은 아니기 때문에) 값이 다르다는 기준점으로 삼기 위해 평균값을 구해보았습니다.

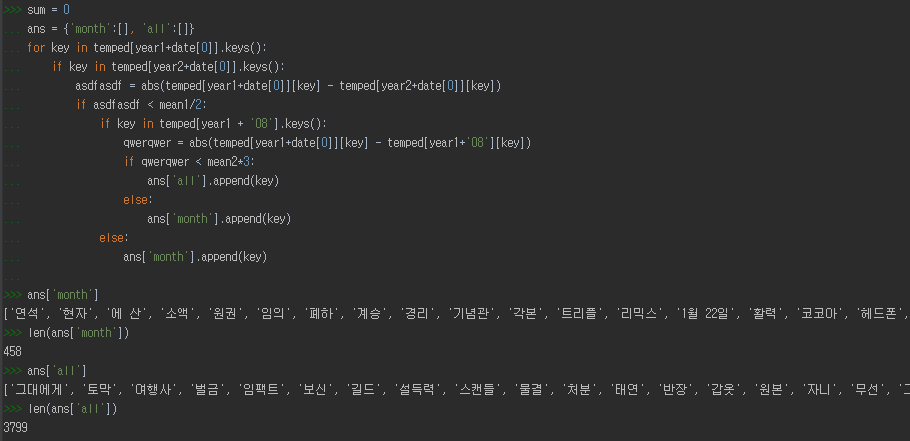


이때 개수가 너무 많아 mean/2를 기준으로 동일한 빈도를 가지고 있다고 판단되어지는 단어들의 수는 4257개로 나오며 주석 처리한 print함수를 이용해 실제로 출력을 해 보았을때

나이프, 코너, 저도, 생계, 바른, 명심, 하차, 현황, 맞춤, 합리, 조카, 업데이트 등등의 일반적으로 자주 나올법한 단어들이 눈에 띄는 모습을 보였습니다.

이때 여기서 추가로 1월에 자주 쓰이는 단어가 다른 월에도 자주 쓰이는지, 아니면 1월에만 자주 쓰이는지를 구분하기 위해 1월가 가장 먼 데이터를 가지고 있는 2014년 8월의 데이터도 비교해보기로 하였습니다.

그 결과

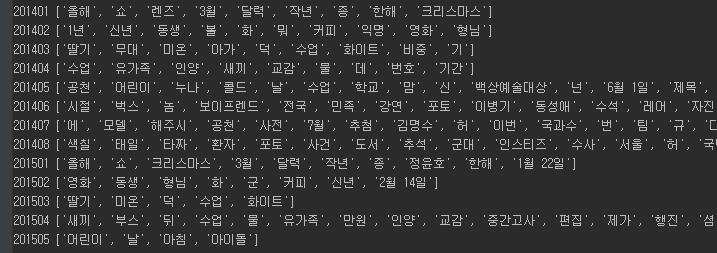


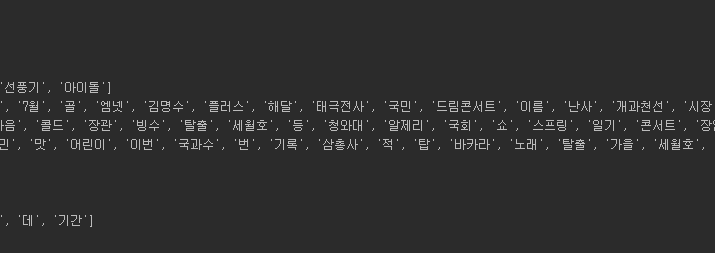
보이는 것과 같이 현자, 헤드폰 등등 1월에 국한될 것 이라고 판단되지 않은 데이터들도 조금씩 포함되어 기준점을 \*3 하여 늘렸지만 완벽히 구분하지는 못하는 모습을 보여주고 있습니다.

하지만 확실히 ‘1월 22일’ 이라는 단어라던가, 여름인 8월에는 마시기 힘든 겨울음료 코코아, 연초에 나올 수 있을만한 단어인 폐하, 계승, 기념관(신년 이벤트가 있을 수 있다.)등이 구분되는 것이 보이고 전체 달을 기준으로 비교해본 것이 아니라 8월 하나만 가지고 비교를 한 결과만으로 특징적으로 보이는 단어들이 보여지니 전체 데이터를 가지고 분류를 하면 좀 더 좋은 데이터들이 분류가 될 것임을 예상할 수 있습니다.

따라서 비록 데이터 자체는 2014년 1월 ~ 8월 + 2015년 1월 ~ 5월 까지 분류하였지만 모든 년도에 대해 분류를 한다면 각 월별로 특징적인 데이터, 전체적으로 자주나오는 데이터를 분류해 낼 수 있을 것 이라고 생각이 되었다. 마찬가지로 코코아 같은 단어를 보아 계절별 분류 또한 필요하지 않을까 생각이 들었다. (데이터 분류가 완벽하지 않아 시행하지는 못하였습니다.)

따라서 앞에서 사용한 방식으로 모든 데이터를 사용하는 코드를 짜보기로 하였습니다





이렇게 분류를 하였을 때 결과입니다.

각각 달에 분류되는 값을 확인해 보았을 때 해당 월에 어울리는 특징적인 단어, 예를 들면

1월에 크리스마스, 올해, 한해, 2월에 1년, 신년, 영화, 형님 3월에 화이트, 수업, 딸기 5월에 어린이 등등 각각 월에 어울리는 특징적인 단어가 눈에 띄게 보이는 모습을 보아 분명히 해당 분류방식이 잘못 되지 않았다고 판단할 만한 근거가 되었다 생각이 됩니다.

하지만 세월호, 탈출, 국과수 같이 그 당시 있었던 가장 큰 사건인 세월호 사건도 특징으로 분류하는것을 보아 2015년 데이터가 없는 6~8월 데이터는 정확하게 분류해내지 못하는 모습을 보여주었습니다.

사실 자료분석을 하는데 있어 특징적인 모습이 나타나지 않는다면 일별로 추가로 분석을 해야 되나 고민을 하였지만 이미 이 정도로 충분히 고무적으로 보이는 성과가 보이고 레포트 7장 제한 또한 얼마 남지 않았기 때문에 이 정도로 하기로 하였습니다.

따라서 자료분석은 이 정도로 마무리 하기로 결정했습니다.

1. **결론**

우선적으로 이러한 자료분석을 통해 당연하다고 생각되지만 첫번째 문제인 년도 상관없이 매월 자주 쓰이는 단어가 있다는 점은 분명 존재한다고 할 수 있습니다.

비록 데이터 자체가 2014년 2015년밖에 없는데, 만약 년도가 훨씬 더 많아지게 된다면 시간은 오래걸리겠지만 완벽하게 이러한 특징적인 단어들을 구분해 낼 수 있다고 생각이 듭니다.

또한 이러한 연장선상으로, 매 시기 특징적인 단어를 분류해 내는데 있어서 응용 또한 충분히 가능하리라고 생각됩니다.

각 데이터를 분류하는데 있어서 기준점의 경우는 여전히 정확하지 않는 것은 사실입니다. 하지만 이 기준점을 적절히 조절해서 분류를 빡쎄게 할지 안 할지 정할수 있다는 점 덕분에 특별히 문제점은 없어 보입니다.

분류한 매월 특징적인 단어

1월: 올해, 쇼, 크리스마스, 3월, 달력, 작년, 한해, 종

2월: 영화, 동생, 형님, 화, 커피, 신년

3월: 딸기, 미온, 덕, 수업, 화이트

4월: 새끼, 수업, 번호, 기간, 교감

5월: 어린이, 날, 아이돌

어쩌다 보니 나온 거지만 당시의 여론, 상황을 보여주는 단어

2014년 4월: 유가족, 인양 물

2014년 5월: 공천, 백상예술대상

2014년 6월: 태극전사, 난사, 알제리, 시장, 세월호

등등

특징적인 단어를 살펴보면 확실히 그시기에 납득이 갈만한 단어들만으로 구성되어 있어 앞에서 예상한

1. 같은 월, 같은 일일 때 년도의 차이가 단어 차이에 영향을 주지 않는 경우
2. 월, 일이 지났는데도 계속하여 동일한 비율을 차지하고 있는 경우

의 기준에 따라 구분하였을 때 데이터가 많아질수록 원하는 특징적인 데이터를 분류 할 수 있을 것 이고 이에 따라 월별로 당시의 여론, 상황을 보여주는 단어를 알아낼 수 있을것입니다.

이는 처음 예상한 특징적인 단어들이고, 결과와 비교해보았을 때

1. 다른 년도 같은 월,일의 경우 비슷한 단어가 항상 사용될 가능성이 높다.

실제로 이것을 기준으로 분류하였고 이러한 형태로 단어가 분류된 것을 확인할 수 있었습니다.

1. 평소에도 일반적으로 쓰는 단어일수 있다.

역시 이러한 단어는 처음 데이터 빈도분석시 0.001%에 걸려 없어졌든지, 해서 고려되지 않은 것으로 보입니다. 어차피 절대값 평균을 기준으로 하기 때문에 분석시 고려되지 않을 것입니다.

1. 특정 유명인사의 인지도에 따라 꾸준히 언급될 가능성도 고려해봐야 한다.

2번과 마찬가지로 빈도분석시 걸러진것으로 보입니다.

1. 유행어, 혹은 인터넷 특유의 유행에 따라 여론과 상관없이 자주 쓰이는 단어도 있다.

아마 이러한 단어들은 형태소 분석시 error가 뜨거나 제대로 분석이 되지 않아 데이터가 사라지지 않았을까 생각이 듭니다.

이러한 가정들 역시 정답은 아니었지만, 이러한 가정들 자체가 가능성이 없는 것은 아니라서 시간이 좀 더 주어진다면 분석한 코드를 좀 더 가다듬어 이러한 단어들도 표시할 수 있게 만들수 있으면 좋을 것 같습니다.

우선적으로 분류해낸 데이터들로는 1번유형의 데이터만이 분류되는 모습이 보입니다. 데이터가 부족한게 아쉽습니다..

데이터를 좀더 많이 쓸 수 있었다면 좋았을 텐데 아쉽게도 컴퓨터 성능과 좋지 못한 코드로 제데로 분류를 못해서 너무 아쉽습니다. 이번 한 학기 수고 많으셨습니다.