트위터 데이터를 이용한 광고 분류와 분석

소프트웨어 학부 20160323 전병모

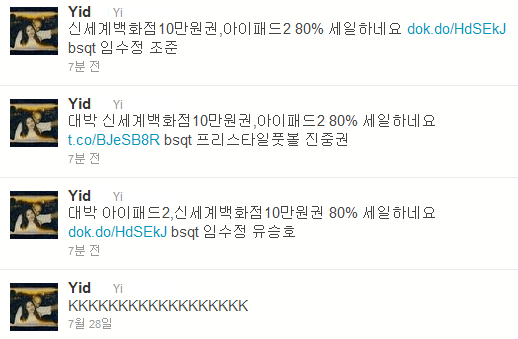
개발이유:

광고는 사용자의 불쾌함을 증폭시킨다.

서버의 공간과 트래픽을 낭비시키는 여러 가지 부정적인 사항이 많다.

애드블럭처럼 커뮤니티 광고 차단기능도 필요합니다.

개요: 인터넷에서 넘쳐나는 광고들을 막기위해 트위터로 부터 데이터를 수집하여 케라스를 이용하여LSTM 모델을 학습시켜서 텍스트를 입력으로 주면 광고인지 판별하는 광고 분류기 만들고 트위터에 있는 광고들을 분석하고 삭제 할 광고들을 분류합니다.



(그림1)불쾌한 광고 예시

해결과정

1. 트위터 JSON 데이터에서 Text만을 따로 파싱하여 저장합니다.

2. 트레이닝 셋을 만들기 위해서 몇가지 샘플 데이터를 수동으로 광고인지 아닌지 수동으로 분류하여 데이터를 분류합니다.

3. 분류된 트레이닝 셋을 단어 단위로 나누어서 단어를 숫자에 매칭하여 문장 데이터들을 숫자 n x m 매트릭스로 표현하는 전처리 과정을 거칩니다.(n은 문자열 수, m은 한 문장의 최대 단어수)

4. Keras를 사용한 LSTM 모델을 사용하여 학습 시킵니다. 이때 트레이닝 셋에서 80%를 학습데이터로 사용하고 나머지 20%를 검증 데이터로 하여 학습을 하면서 결과를 검증합니다.

5. 학습된 모델을 이용하여 트위터에서 광고를 분류하여 삭제합니다.

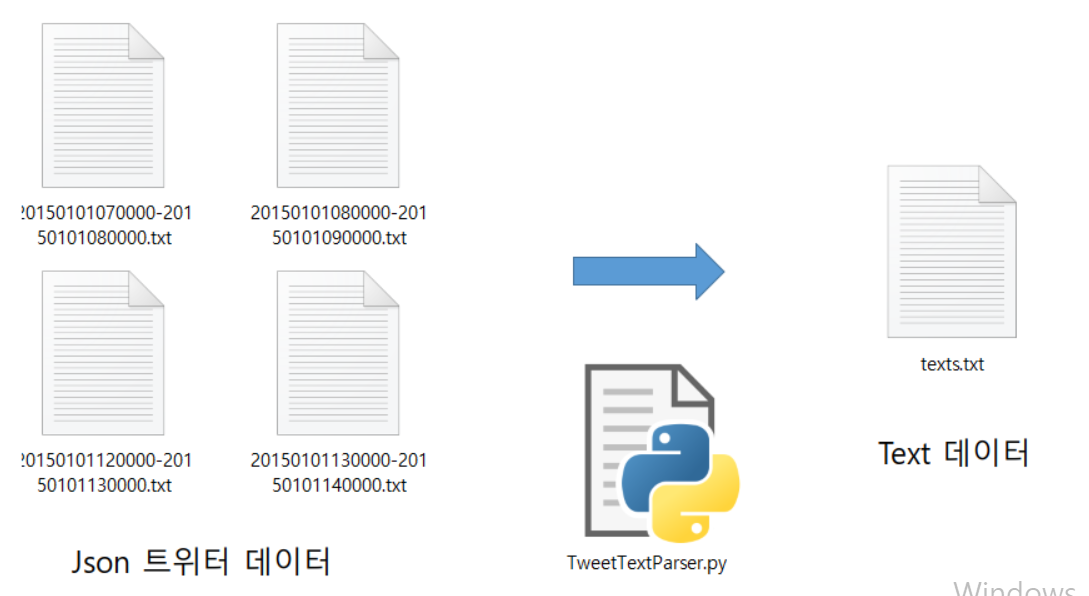
트위터 파싱 및 데이터 분석

트위터 데이터는 JSON 방식으로 저장되있는것을 확인할 수 있습니다.

이중에서 목적에 필요한 데이터는 ‘text’ 뿐 이므로 JsonParser를 사용하여 Text만 따로 분류한 데이터 들을 만들었습니다.

{ "\_id" : "550306727342522370", "filter\_level" : "medium", "retweeted" : false, "in\_reply\_to\_screen\_name" : "DOdaehee", "possibly\_sensitive" : false, "truncated" : false, "lang" : "ko", "in\_reply\_to\_status\_id\_str" : "550306527983468546", "id" : 550306727342522370, "in\_reply\_to\_user\_id\_str" : "2878654873", "timestamp\_ms" : "1420038322660", "in\_reply\_to\_status\_id" : 550306527983468546, "created\_at" : "Wed Dec 31 15:05:22 +0000 2014", "favorite\_count" : 0, "place" : null, "coordinates" : null, "text" : "@DOdaehee 이제 곧 살 거야... 진짜 곧 살 거야 2주 안에 살 듯?", "contributors" : null, "geo" : null, "entities" : { "trends" : [], "symbols" : [], "urls" : [], "hashtags" : [], "user\_mentions" : [ { "id" : 2878654873, "name" : "도", "indices" : [ 0, 9 ], "screen\_name" : "DOdaehee", "id\_str" : "2878654873" } ] }, "source" : "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>", "favorited" : false, "msg\_create\_at\_forlong" : 1420038322847, "in\_reply\_to\_user\_id" : 2878654873, "retweet\_count" : 0, "id\_str" : "550306727342522370", "user" : { "location" : "영고샤덕의 안식처, 샤이니의 영향권 안 쪽", "default\_profile" : false, "statuses\_count" : 14675, "profile\_background\_tile" : false, "lang" : "ko", "profile\_link\_color" : "09F0AF", "profile\_banner\_url" : "https://pbs.twimg.com/profile\_banners/490377882/1417356062", "id" : 490377882, "following" : null, "favourites\_count" : 3803, "protected" : false, "profile\_text\_color" : "333333", "verified" : false, "description" : "당신의 가장 작은 움직임에도 눈길을 보내고, \n당신의 가장 작은 진동에도 귀 기울일게요./제발 저를 팔로우하지 마세요, 저는 쭉정이같은 빠순이랍니다", "contributors\_enabled" : false, "profile\_sidebar\_border\_color" : "FFFFFF", "name" : "행모이 \_(:3」∠)\_", "profile\_background\_color" : "B2DFDA", "created\_at" : "Sun Feb 12 14:06:56 +0000 2012", "default\_profile\_image" : false, "followers\_count" : 43, "profile\_image\_url\_https" : "https://pbs.twimg.com/profile\_images/539047730908119041/IBlajRTI\_normal.jpeg", "geo\_enabled" : false, "profile\_background\_image\_url" : "http://pbs.twimg.com/profile\_background\_images/459284541379604480/qMUAbU\_L.jpeg", "profile\_background\_image\_url\_https" : "https://pbs.twimg.com/profile\_background\_images/459284541379604480/qMUAbU\_L.jpeg", "follow\_request\_sent" : null, "url" : null, "utc\_offset" : 32400, "time\_zone" : "Seoul", "notifications" : null, "profile\_use\_background\_image" : true, "friends\_count" : 256, "profile\_sidebar\_fill\_color" : "FFFFFF", "screen\_name" : "2jae\_go3", "id\_str" : "490377882", "profile\_image\_url" : "http://pbs.twimg.com/profile\_images/539047730908119041/IBlajRTI\_normal.jpeg", "listed\_count" : 0, "is\_translator" : false } }

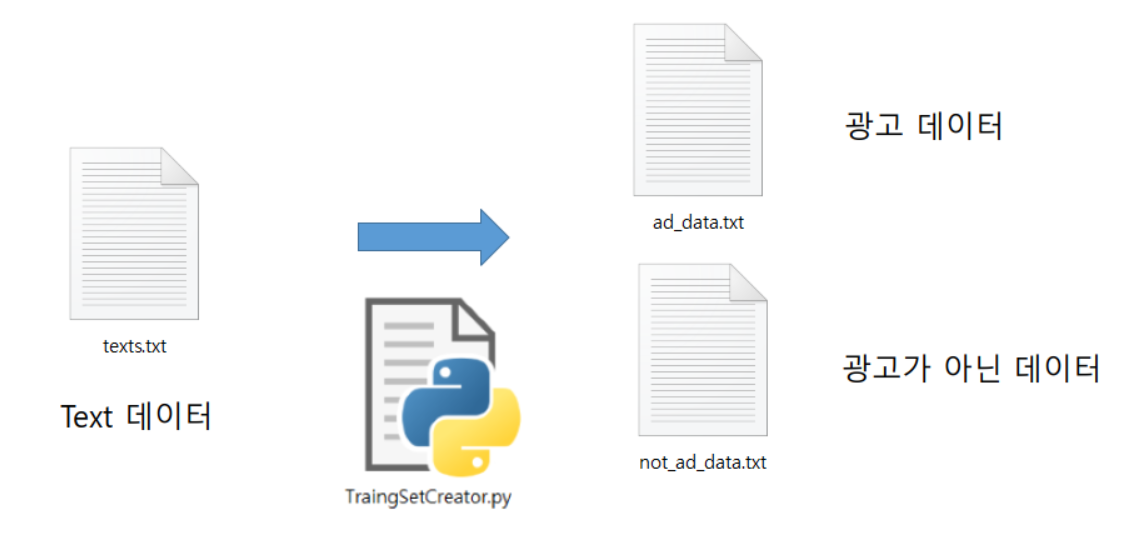
(글1) 트위터 데이터 예시



(그림2) 트위터 데이터 파싱 과정

광고를 찾기위한 키워드들을 적당히 활용하여 키워드가 들어간 글이 광고인지 아닌지 확인 하여 데이터를 광고인 텍스트와 광고가 아닌 텍스트로 수동으로 분리합니다.

키워드 예시 [분양, 카지노, 상담 …..]



(그림3)트레이닝 셋 분류 과정

성cgv영화관 부대시설 분양 업종 커피숍,편의점 스넥 패스트푸드,테이크아웃,의류 http://t.co/zfKe63mNo3 문의전화:1600-2692 https://t.co/3qeDgzBfib

진주 실크전문농공단지 분양 끝

광양 중마2차 진아리채 아파트분양 / 광양 여수 순천 최고의 랜드마크 / 34평 단일평형 / 4베이 4룸 / 600만원대 / 선착순 동호수

지정 / 24시간 상담환영~! / 대표번호 1600-9653 / http://t.co/f05VtjKAVh

#허니게임 &gt; 히어로게임 으로바뀌었습니다

바로가기 &gt; http://t.co/pVwZqWnRQH

▼추천인 : 새 우 깡 ▼

(글2) 광고 데이터 예시

#Happiness! 안녕하세요. 레드벨벳 둥둥이, 조이임다 ㅋㅋ 새해 복 많이 받으시고, 올해는 둥둥이와 #some 타보는 건 어떠실지..

거절은 거절임다. 진부한 답멘을 피하기 위한 질문 Time!

Q. 치킨 VS 피자

@Thx\_EXOkai

@J\_AKTF 눈꽃 5. 오늘 게스트나 음악 감독님 다 센스 짱ㅠㅠ멘트 길게 해주시고 특히 첫번째 게스트였던 정은님은 조근조근 할 말

다 하시면서 우리에게 즐거움을 선사해주셨다ㅠㅠ 치킨 앤 하이듴ㄲㅋㅋㅋ라고 의도치않게 말실수한 오빠를 놀리기도 하셨고

@pram11 개강하자마자 놀랍게도 치킨 사달라는 과 신입생이 다 쳐들어왔다고 한다

(글3)광고가 아닌 데이터예시

데이터 전처리

단어를 숫자에 매칭시켜서 문장을 단어 단위로 나누어 숫자를 매칭시킵니다.

전처리전 데이터 : 히오스 가입시 전원 카드팩 히오스 바로가기

단어 Sequence 배열

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 카드팩 | 전원 | 히오스 | 바로가기 | 가입시 |
| 0 | 1 | 2 | 3 | 4 |

전처리후 데이터 : [2, 4, 1, 0, 2, 3]

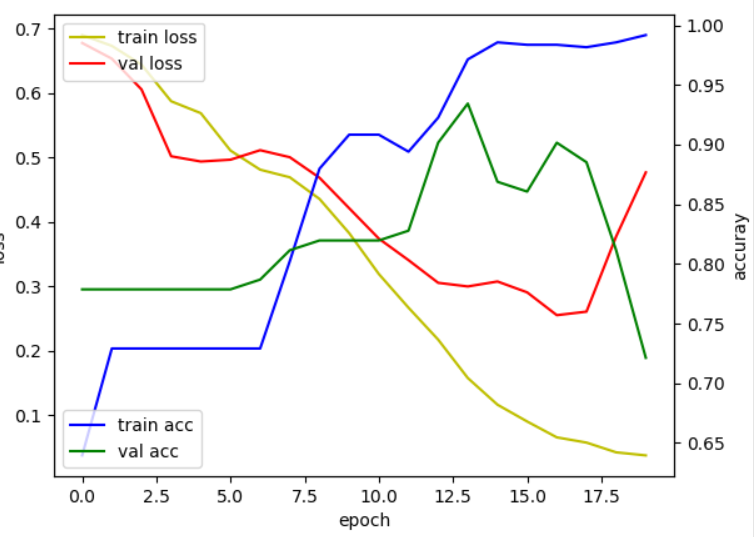
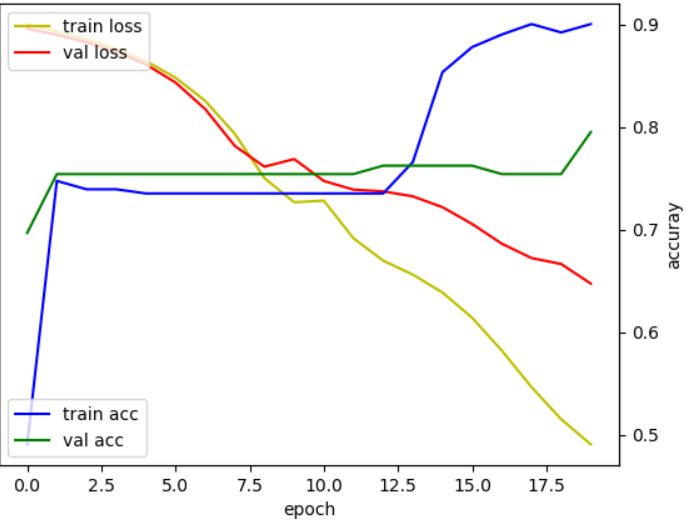
이런방법으로 n x m 매트릭스를 생성합니다. n = 데이터 셋 개수, m = 최대 문장의 단어수

이때 메모리와 시간을 절약하기 위해서 문자의 빈도수를 계산하여 빈도가 많은 순으로 표현 할 문자의 수를 10000개 정도로 제한합니다.

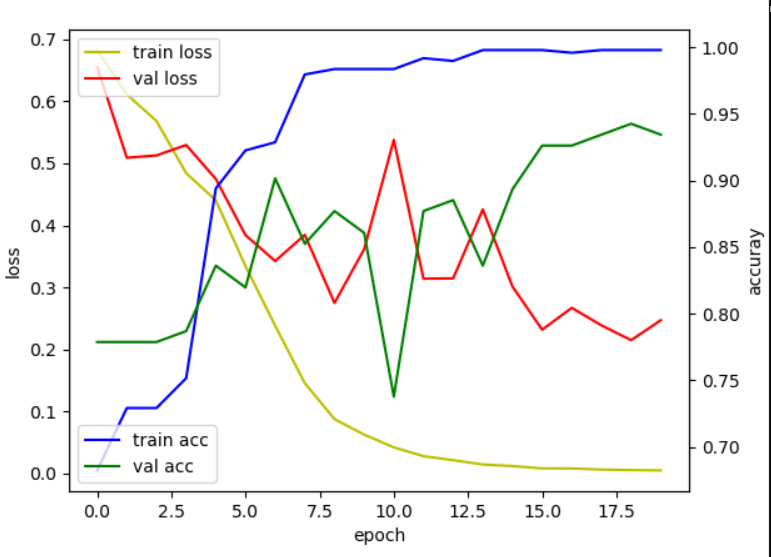
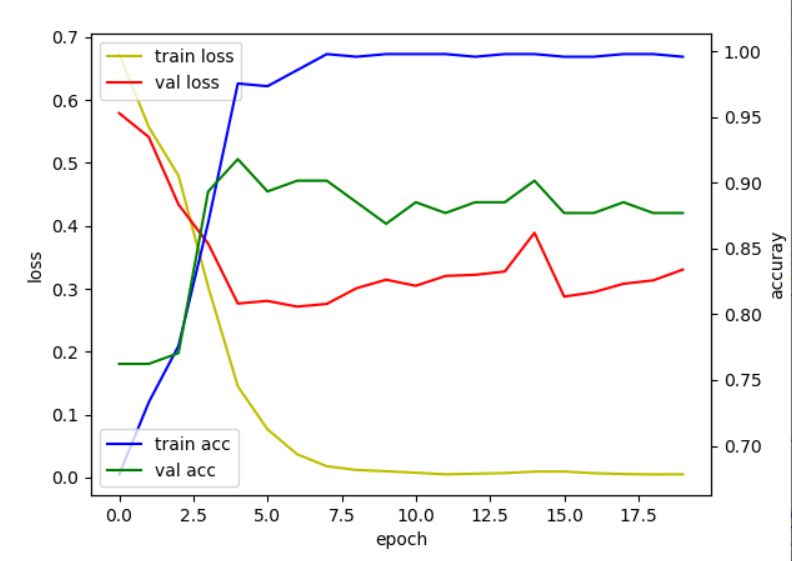
파이썬의 Keras 모듈을 사용하여 LSTM 모델을 구현하여 트레이닝 셋의 80% 정도를 학습데이터로 사용하고 나머지 20%를 검증데이터로 사용합니다.

이때 epoch와 batch\_size는 여러 가지로 실험하여 최적의 것을 정하겠습니다.

학습 수행 결과 - LSTMAdTrainer.py



(그림4) Epoch =20 batch\_size = 512 결과 (그림5)Epoch =20 batch\_size = 256 결과

(그림6) Epoch = 20 batch\_size = 128 결과 (그림7)Epoch = 20 batch\_size = 64결과

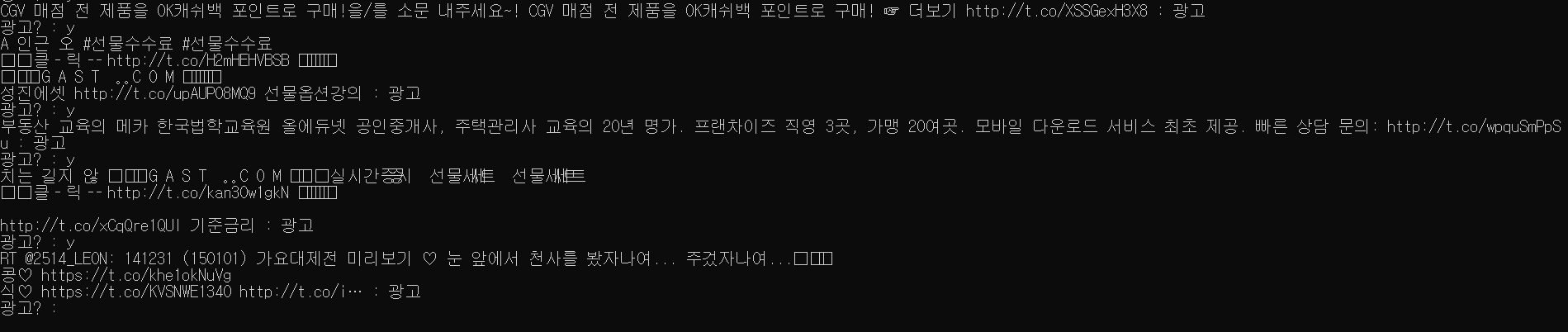
그래프를 통해

학습의 결과가 Epoch = 18, batch\_size = 128 정도에서 가장 좋은 것을 확인 할 수 있습니다.

학습된 모델을 사용하여 광고 분류하기 - LSTMAdClassfier.py

사용된 데이터: 2015년 01월 전체중 무작위로 산출된 데이터

모델을 파일로 저장하고 광고 분류기를 실행 한 결과 성공적으로 작동하는 것을 알수 있습니다.



(그림 8)광고를 성공적으로 분류하는 모습

성공률 더욱 더 높이기

오류를 보완하기 위해서 잘못된 데이터를 광고라고 분류 했을 때 그 자리서 바로 학습데이터로 추가 시켜서 오류 보완하였습니다.

광고 분석 결과

총데이터수 103963

광고수 : 306 광고가 아닌 데이터 수 : 103657

광고는 전체 게시물에서 0.294%를 차지한다.

결론

광고를 분류 할 수 있는 분류기를 만들었습니다.

이제 보기 싫은 광고는 애드 블록처럼 차단도 할 수 있고 회사 차원에서도 불 필요한 트래픽 낭비를 줄일 수 있고 사용자의 편의성도 증대 시킬 수있습니다.