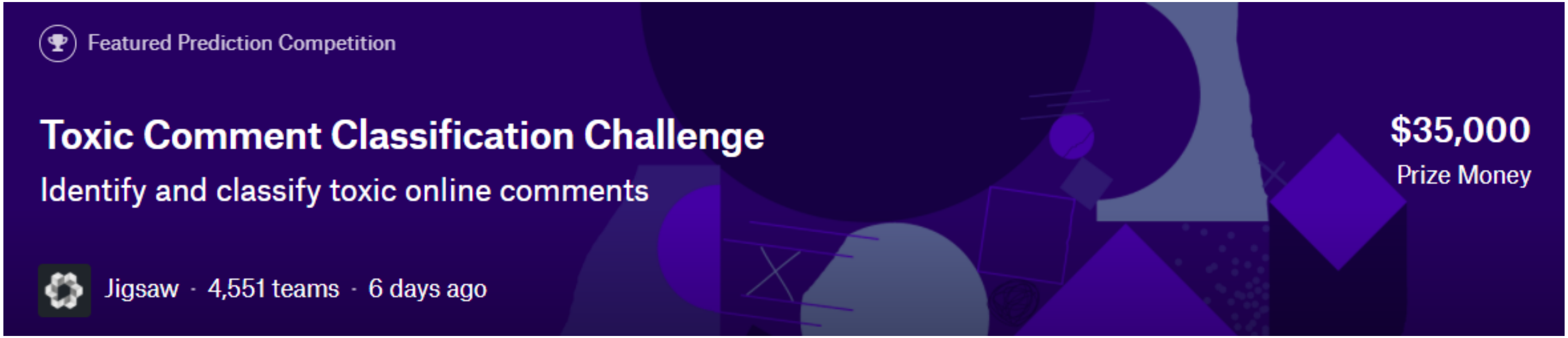
Toxic Comment Classification

김지현님 조 중간발표

(김지현, 이상열, 윤상필, 박희준, 오영택)


Featured Prediction Competition

Toxic Comment Classification Challenge

Identify and classify toxic online comments

$35,000

Prize Money

Jigsaw · 4,551 teams · 6 days ago

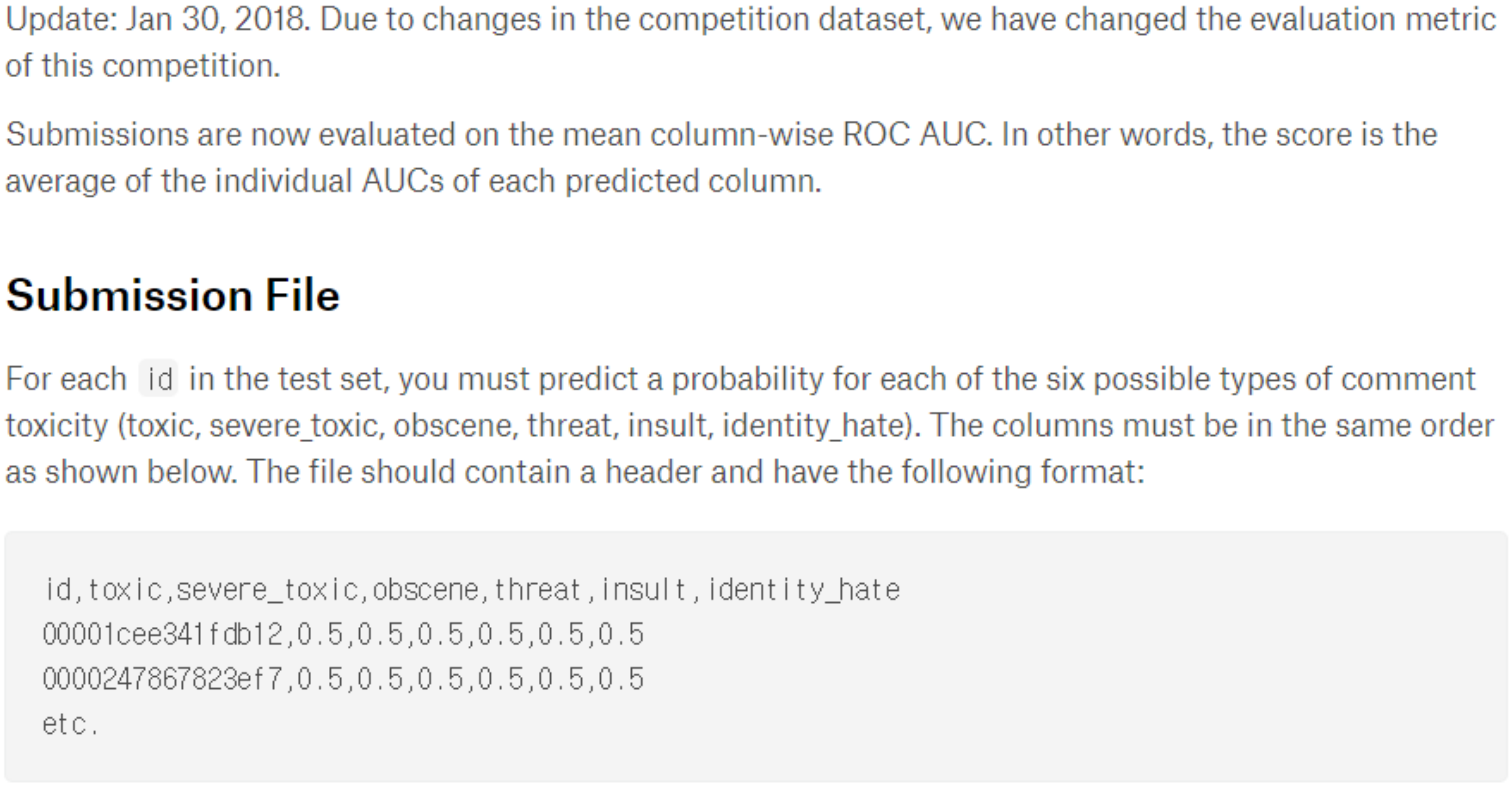
온라인에서 학대와 괴롭힘의 위협은 많은 사람들이 자신을 표현하지 않고 다른 의견을 찾는 것을 포기하는 것을 의미합니다. 플랫폼은 대화를 효과적으로 촉진하기 위해 노력하여 많은 커뮤니티가 사용자 의견을 제한하거나 완전히 종료시킵니다.

대화 AI팀은 온라인 대화를 향상시키는 도구를 개발하고 있습니다. 한 가지 초점 영역은 독성적인 설명 (예 : 무례하거나 무례하거나 누군가 토론을 떠날 가능성이있는 의견)과 같은 부정적인 온라인 행동을 연구하는 것입니다.

Perspective API를 통해 공개된 다양한 모델을 독성을 포함하여 만들었습니다. 그러나 현재의 모델은 여전히 오류를 남기며 사용자가 찾는 독성 유형 (예 : 일부 플랫폼은 욕설이 있을 수 있지만 다른 유형의 독성 콘텐츠는 사용할 수 없음)을 찾는 것을 허용하지 않습니다.

경쟁에서, 당신은 Perspective의 현재 모델보다 위협, 외설, 모욕, 신원 기반 혐오와 같은 다른 유형의 독성을 탐지 할 수있는 다중 헤드 모델을 구축해야 합니다. Wikipedia의 토론 페이지 편집 내용의 데이터 세트를 사용하게 됩니다.
(대회의 데이터 세트에는 모욕적이거나 저속하거나 불쾌감을주는 텍스트가 포함되어 있습니다.)

Evaluation



Update: Jan 30, 2018. Due to changes in the competition dataset, we have changed the evaluation metric of this competition.

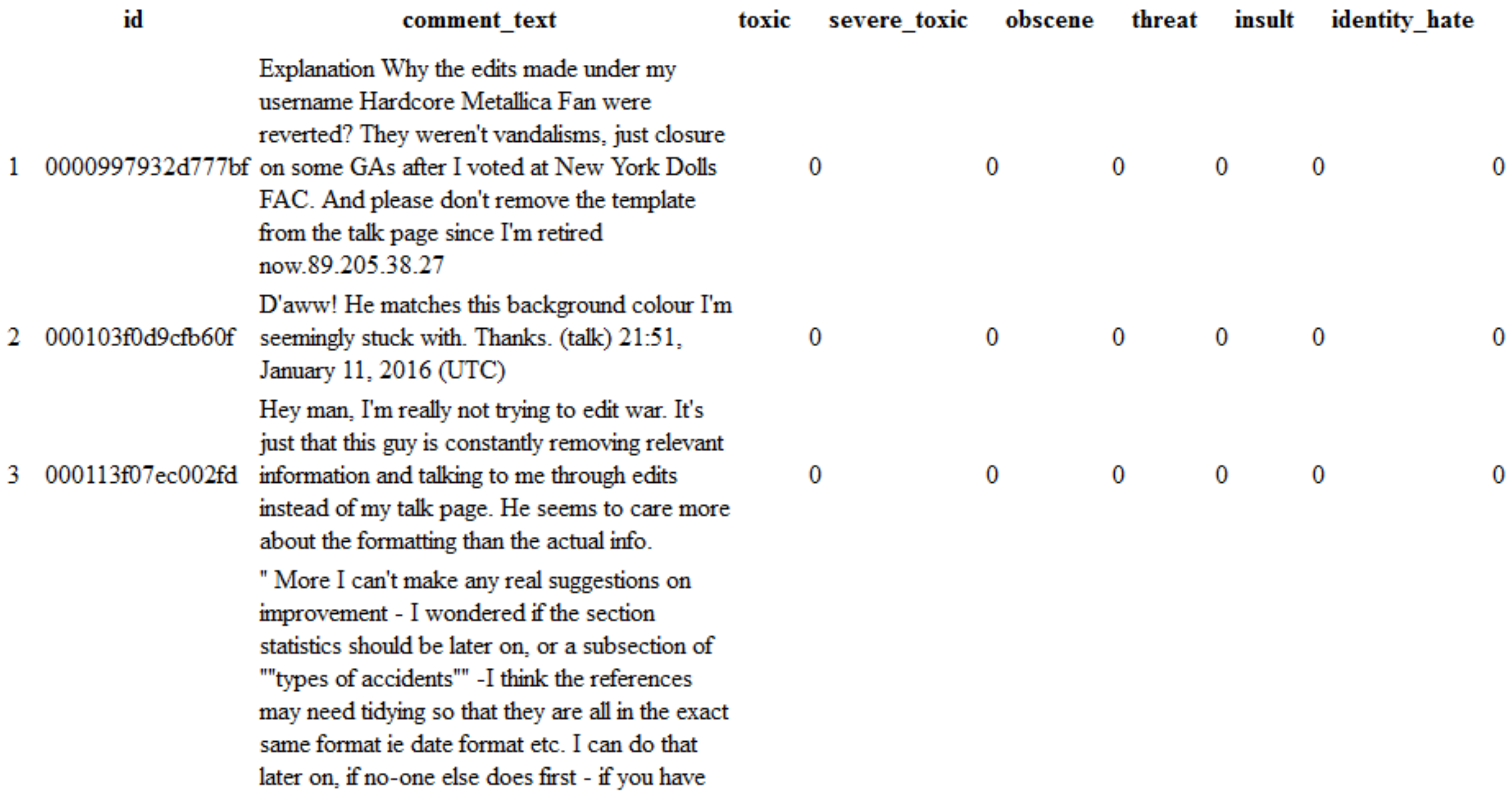
Submissions are now evaluated on the mean column-wise ROC AUC. In other words, the score is the average of the individual AUCs of each predicted column.

Submission File

For each `id` in the test set, you must predict a probability for each of the six possible types of comment toxicity (toxic, severe_toxic, obscene, threat, insult, identity_hate). The columns must be in the same order as shown below. The file should contain a header and have the following format:

```
id,toxic,severe_toxic,obscene,threat,insult,identity_hate
00001cee341fdb12,0.5,0.5,0.5,0.5,0.5,0.5
0000247867823ef7,0.5,0.5,0.5,0.5,0.5,0.5
etc.
```

Peek into the Data



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
1	0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27	0	0	0	0	0	0
2	000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)	0	0	0	0	0	0
3	000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.	0	0	0	0	0	0
		" More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have						

1. Cleansing

참고커널 : regex + text2vec tfidf/lsa + lda + lightgbm

Define function for cleaning

1. all letters to lower (소문자)
2. remove linebreaks (개행)
3. strip multiple whitspace to one (여러 개 공백 하자로 제거)
4. remove links (링크 제거)
5. transform short forms

```
 -  text <- gsub("i'm", "i am", text, perl = T)
 -  text <- gsub("'re", " are", text, perl = T)
```

6. remove "shittext“

```
 - text <- gsub("₩₩b(y)a+₩₩b", "YAAAA", text, perl = T)
 - text <- gsub("₩₩b(b+)?((h+)((a|e|i|o|u)+)(h+)?){2,}₩₩b", "HAHEHI", text, perl = T)
 - text <- gsub("mm(m+)", "m", text, perl = T)
```

7. remove stopwords

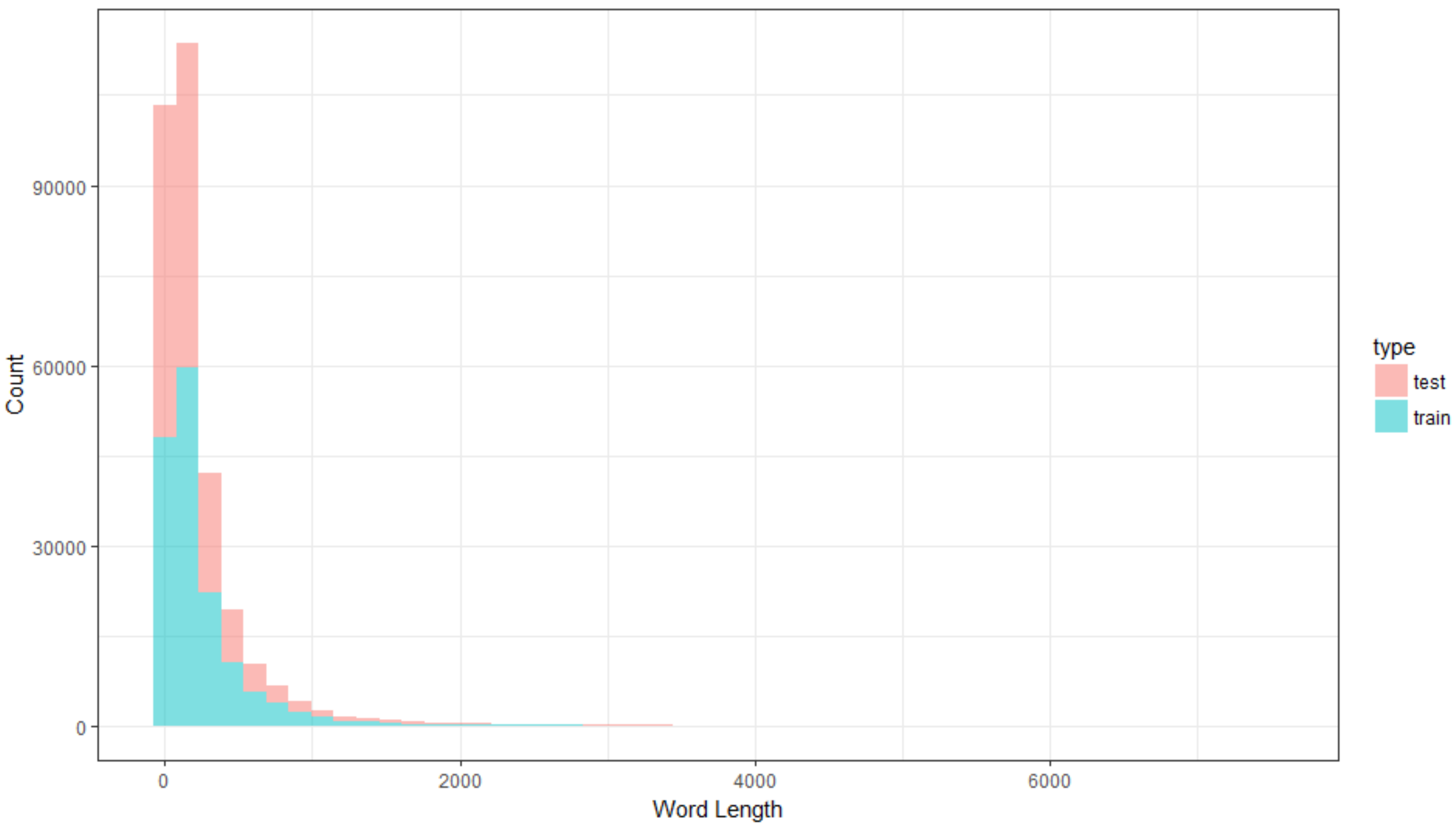
```
 - otherstopwords <- c("put", "far", "bit", "well", "still", "much“, "he“,.....)
```

8. remove graphics, punctuation, digits

2. EDA

참고커널 : toxic Comments EDA and XGB Modelling

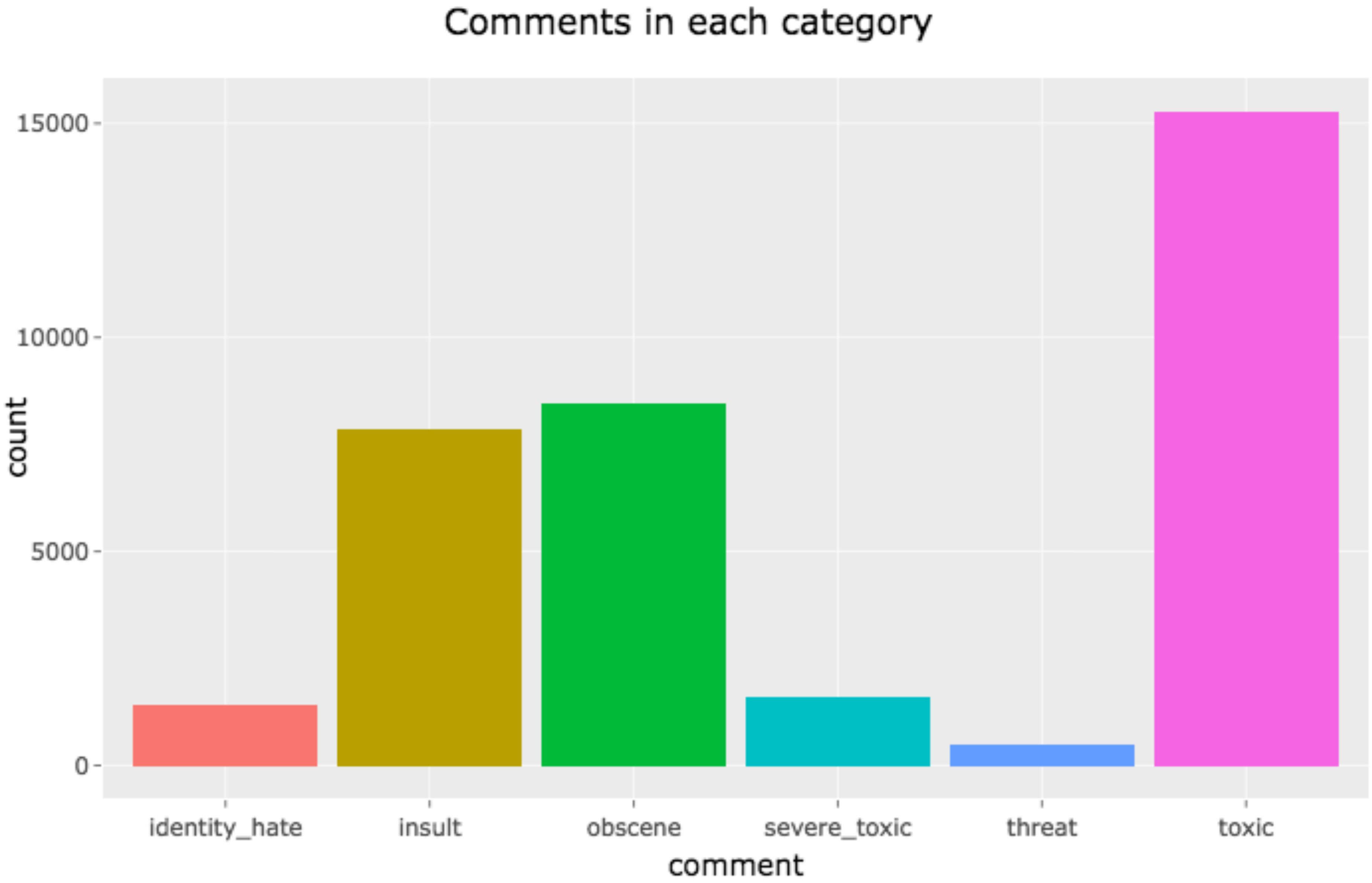
Distribution of Word Length (Test + Train)



2. EDA

참고커널 : Classify toxic comments

카테고리 카운트

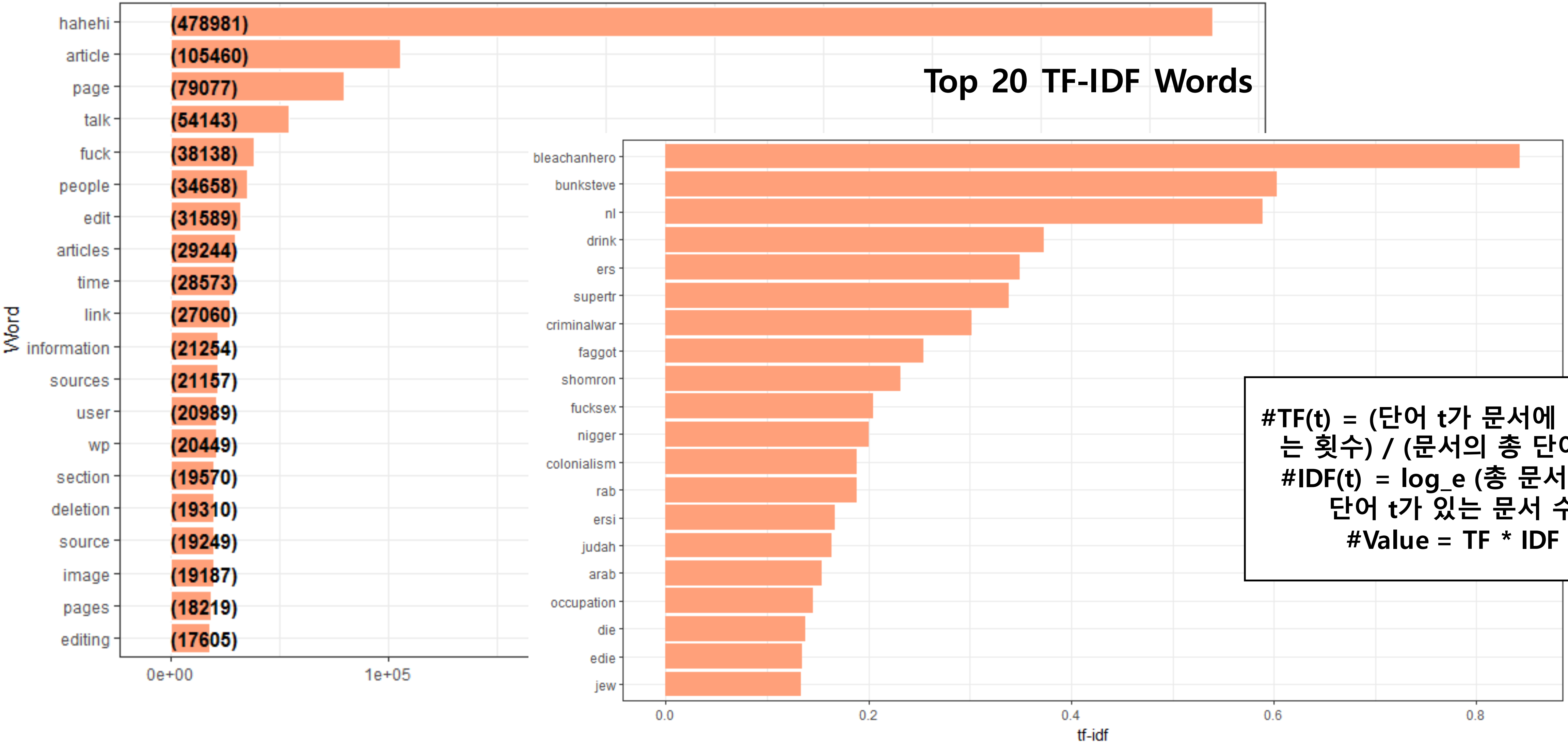


참고커널 : toxic Comments EDA and XGB Modelling

2. EDA

Top 20 most common Words

Top 20 TF-IDF Words

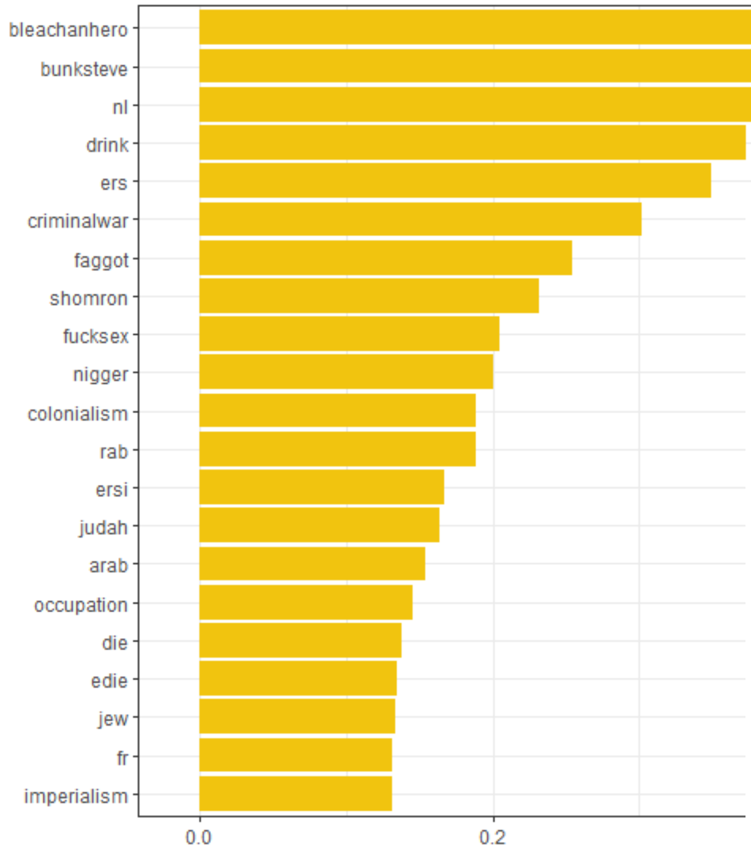


#TF(t) = (단어 t가 문서에 나타나는 횟수) / (문서의 총 단어 수)
#IDF(t) = log_e (총 문서 수 / 단어 t가 있는 문서 수)
#Value = TF * IDF

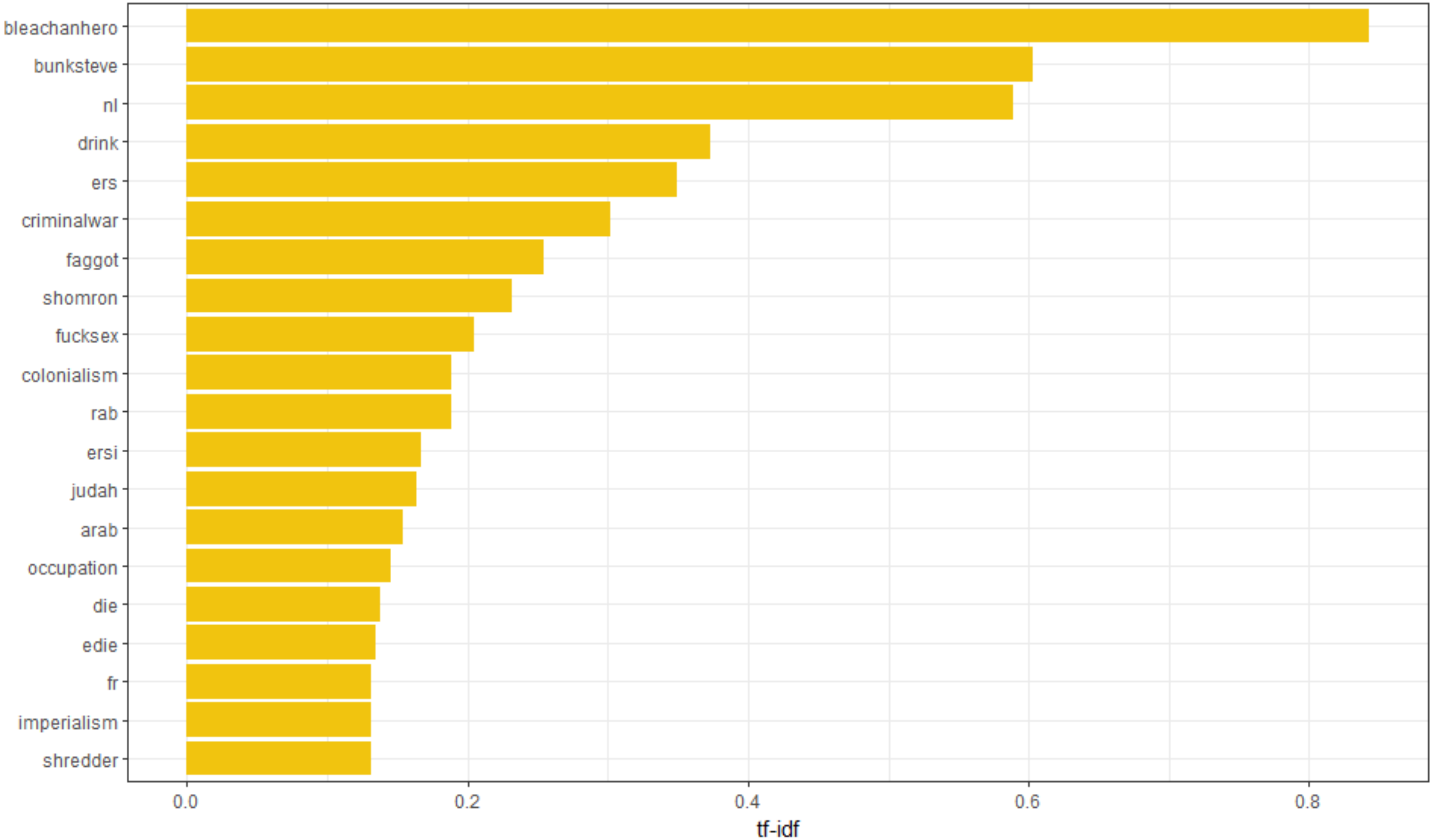
참고커널 : toxic Comments EDA and XGB Modelling

2. EDA

Top 20 Toxic TF-IDF



Top 20 Severe Toxic TF-IDF

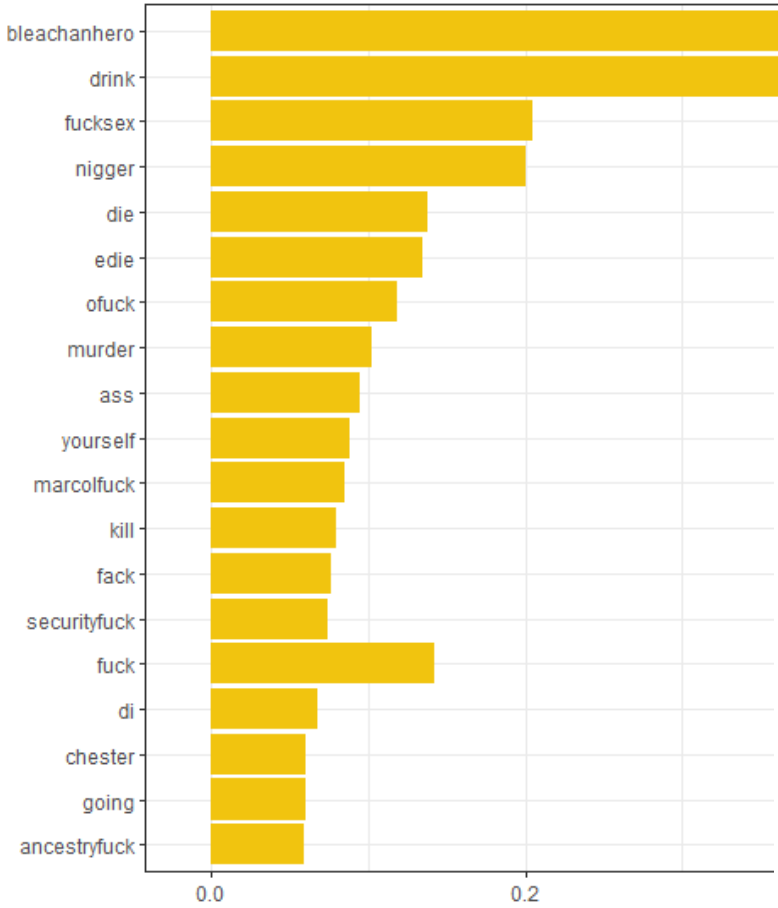


bleachanhero? 표백제?
Bunksteve? 덩어리
Ers
Criminalwar? 범죄자
Faggot? 동생애자

참고커널 : toxic Comments EDA and XGB Modelling

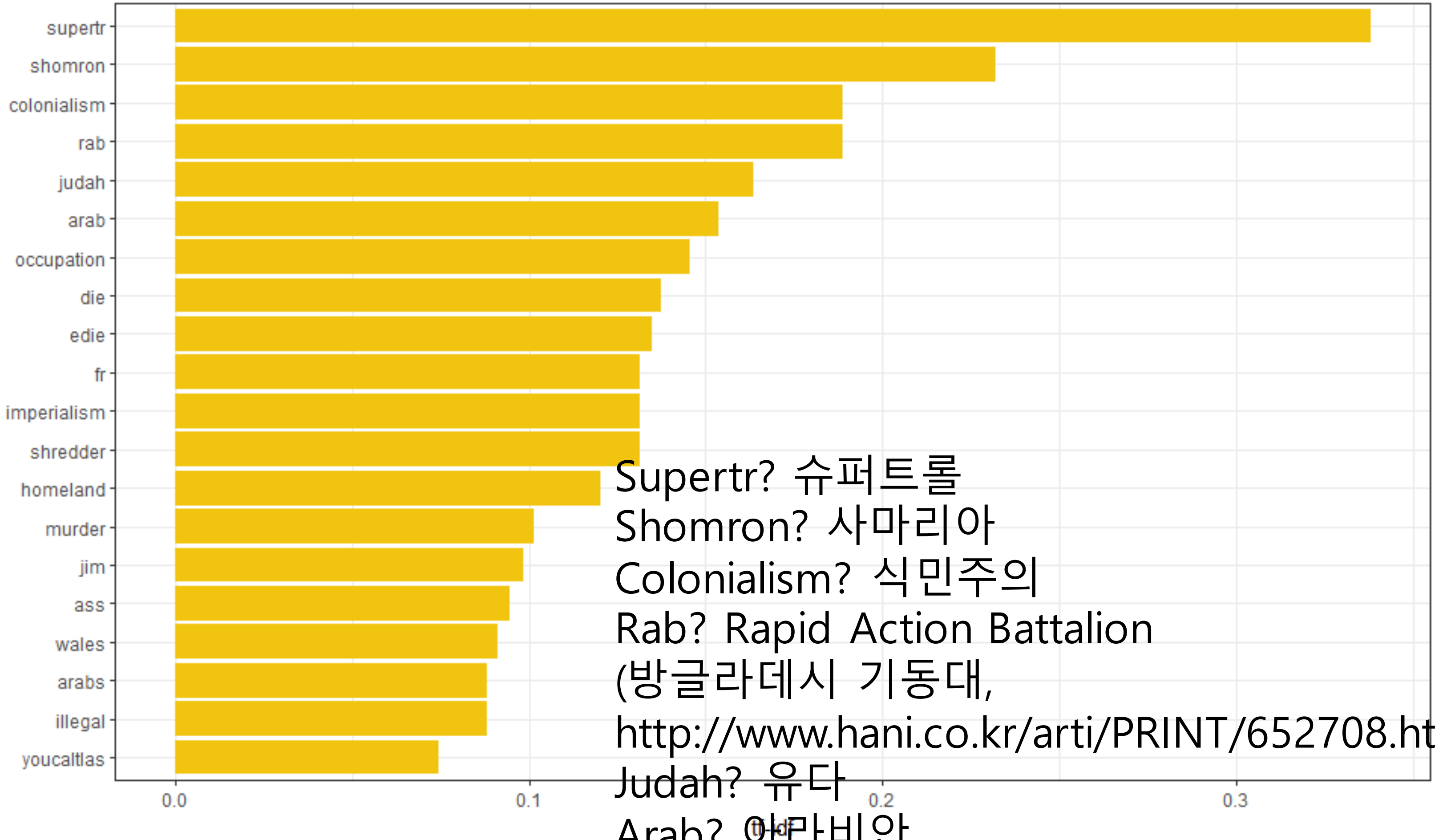
2. EDA

Top 20 Obscene(역겨운) TF-IDF



Nigger? 검둥이
edie?
Ofunk?

Top 20 Threat TF-IDF

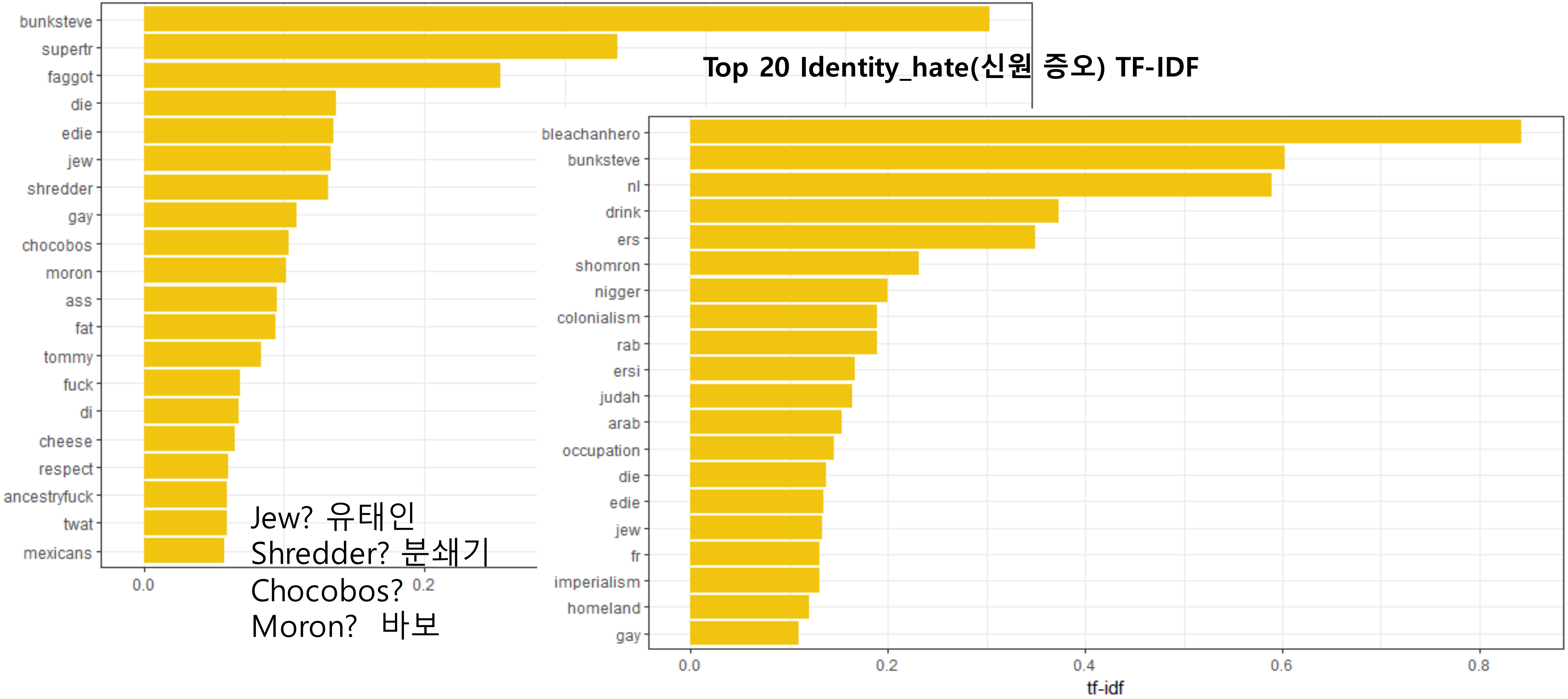


Supertr? 슈퍼트롤
Shomron? 사마리아
Colonialism? 식민주의
Rab? Rapid Action Battalion
(방글라데시 기동대,
http://www.hani.co.kr/arti/PRINT/652708.html)
Judah? 유다
Arab? 아라비안

참고커널 : toxic Comments EDA and XGB Modelling

2. EDA

Top 20 Insult(모욕) TF-IDF



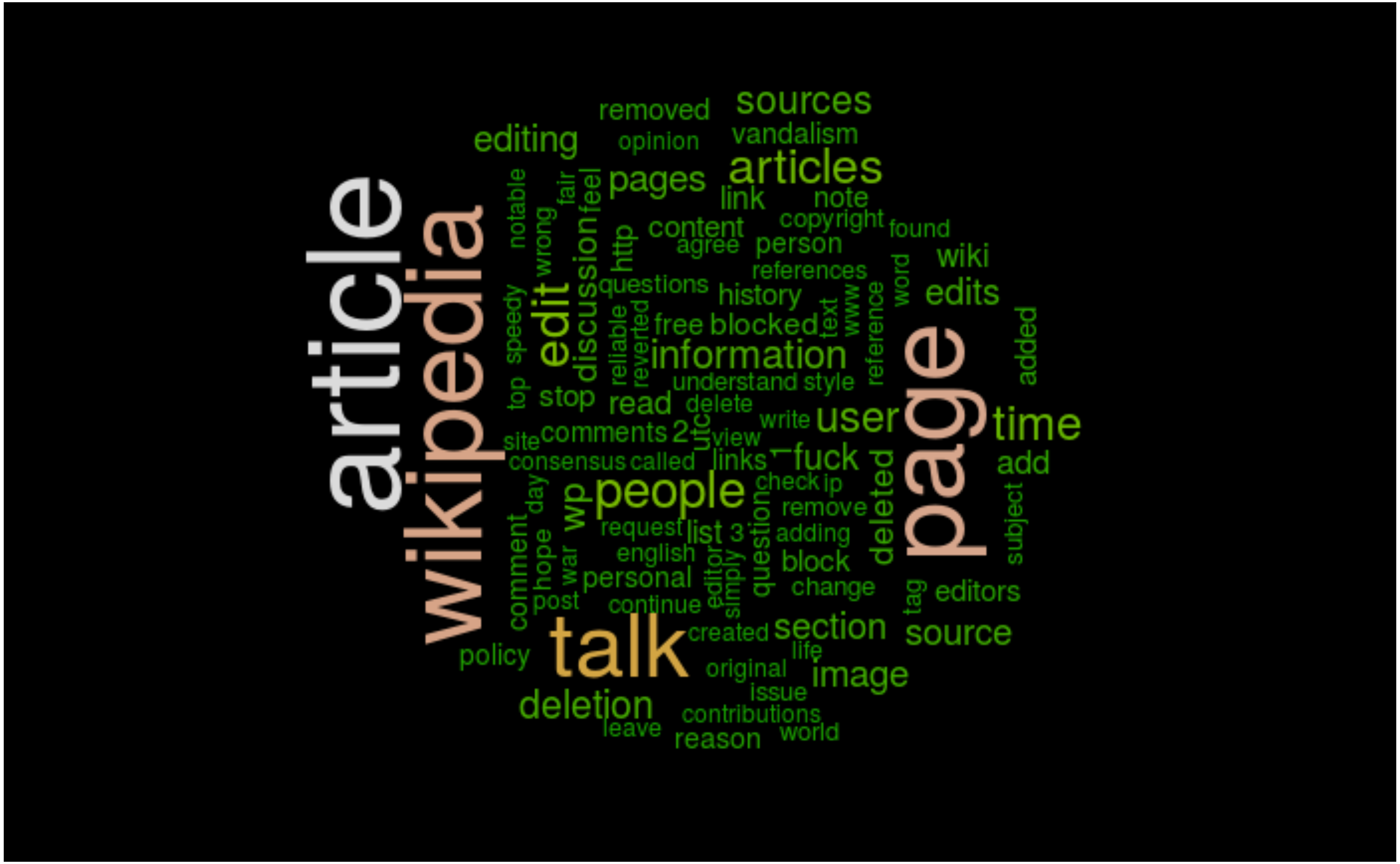
Top 20 Identity_hate(신원 증오) TF-IDF

Jew? 유태인
Shredder? 분쇄기
Chocobos?
Moron? 바보

2. EDA

참고커널 : Classify toxic comments

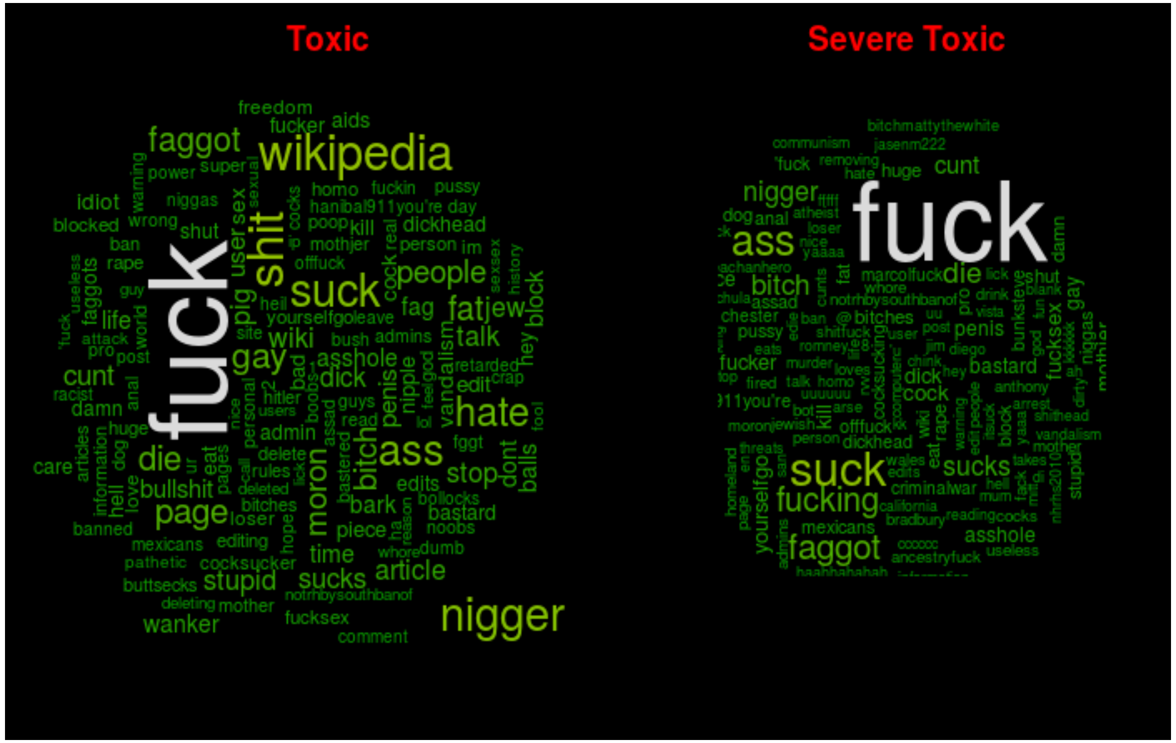
Wordcloud

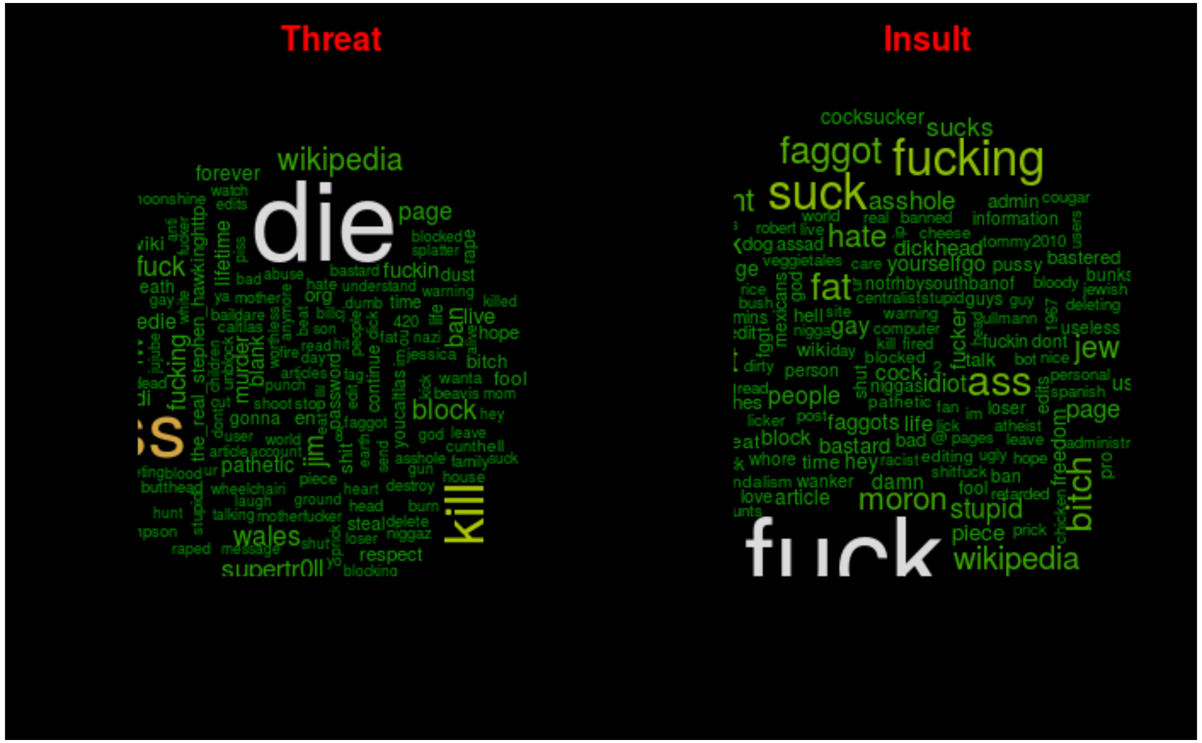


2. EDA

참고커널 : Classify toxic comments

Wordcloud





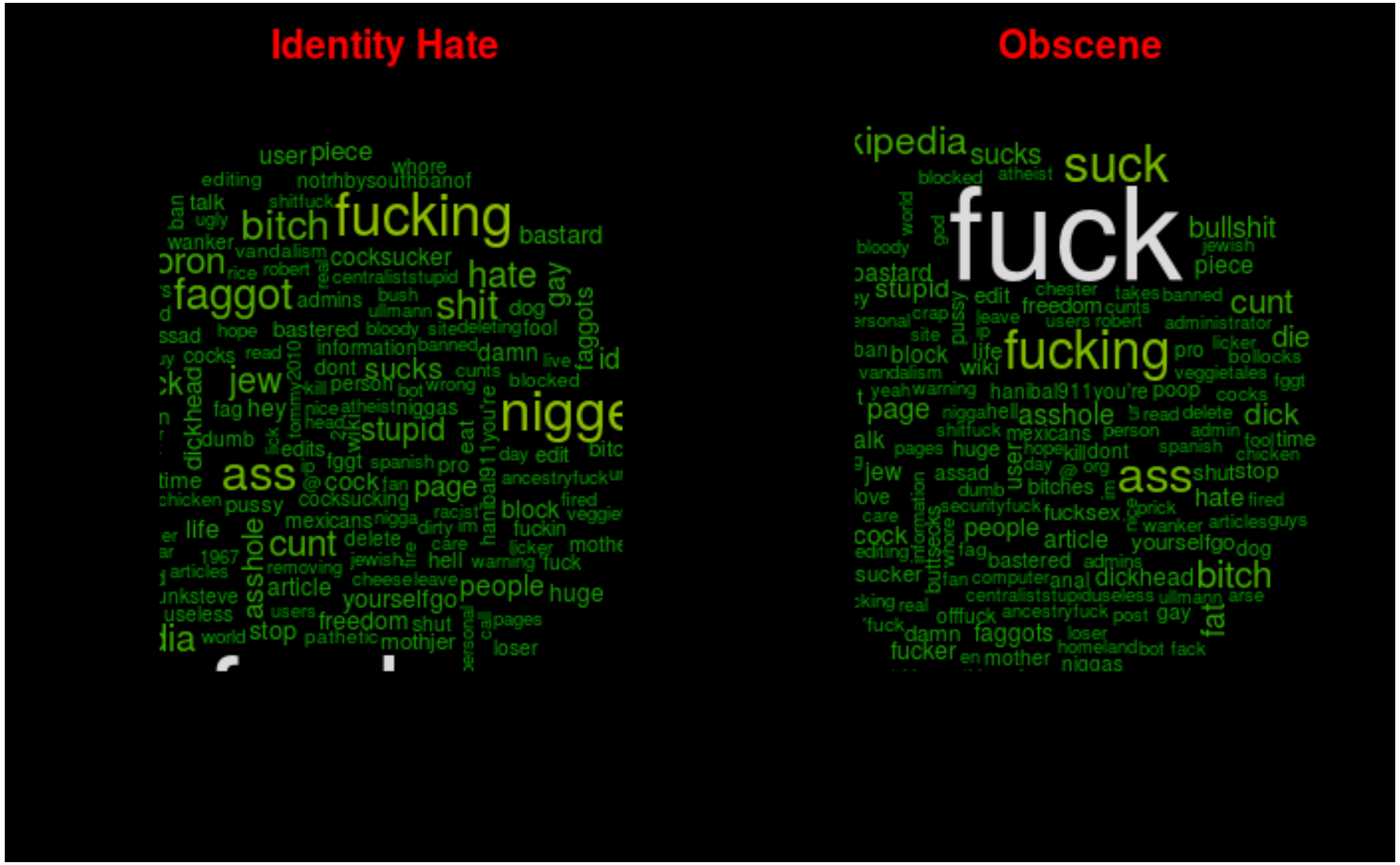
Fuck
Nigger
Shit
Suck

Fuck
ass
bitch
Suck

Die
Kill
Ass
wales

Fuck
Suck
Fucking
Faggot
fat

Wordcloud

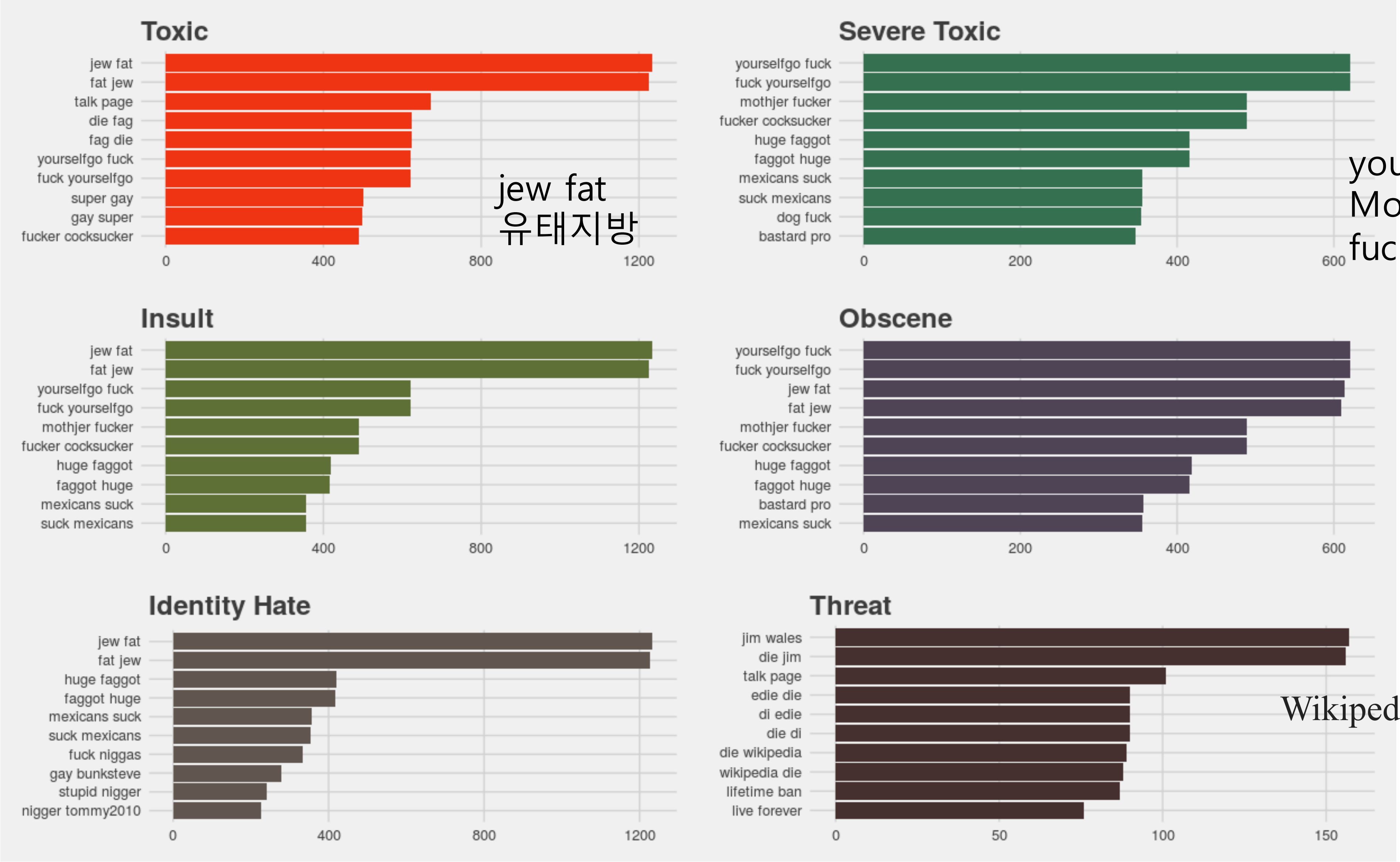


Jew
Fucking
Bitch
nigger

Fuck
Funcking
Suck
ass

2. EDA

참고커널 : Classify toxic comments



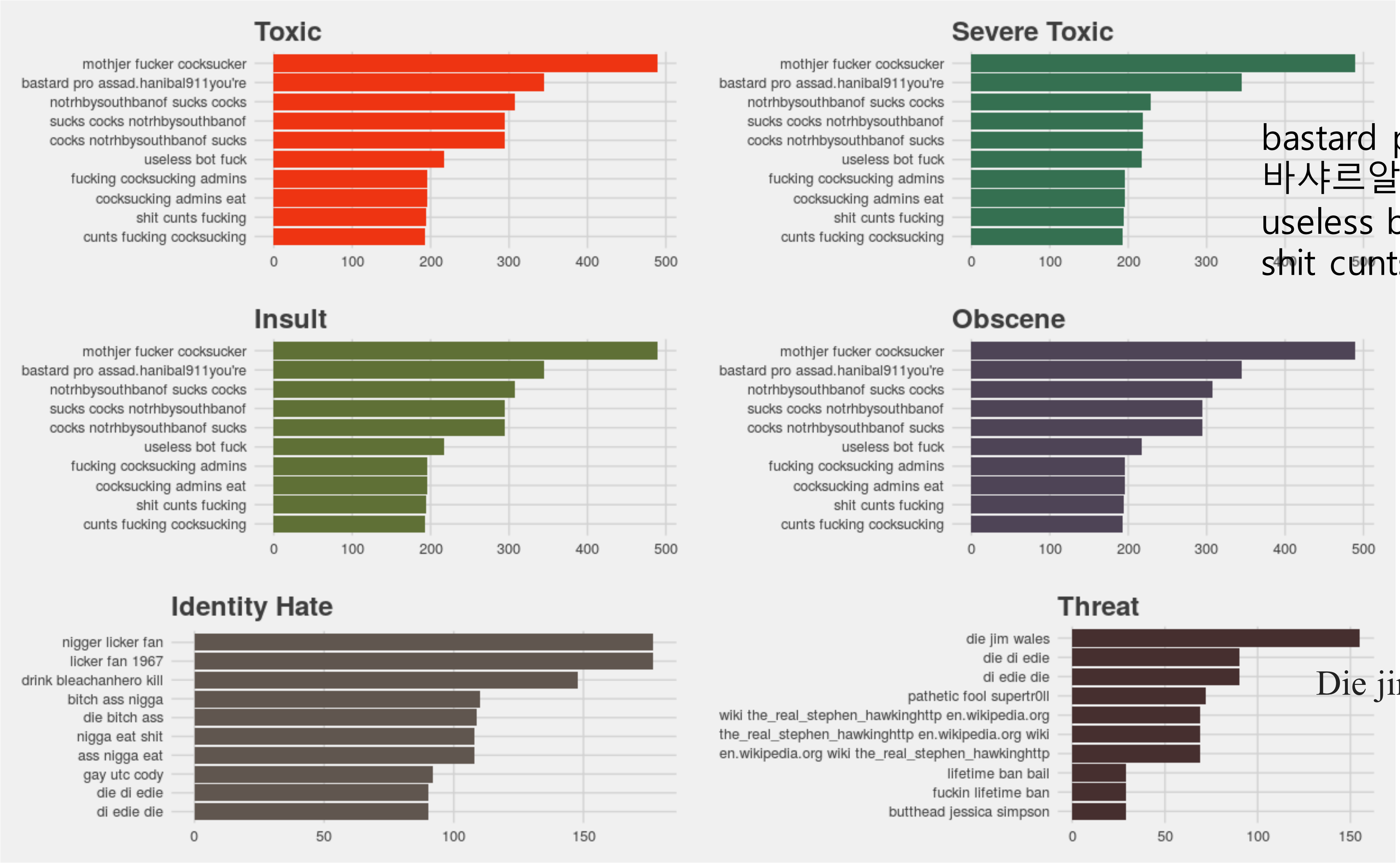
jew fat
유태지방

yourselfgo fuck 너자신 fuck
Mothjer fucker XXX 새끼
fucker cocksucker : 망할 새끼

Wikipedia founder Jimmy Wales

2. EDA

참고커널 : Classify toxic comments



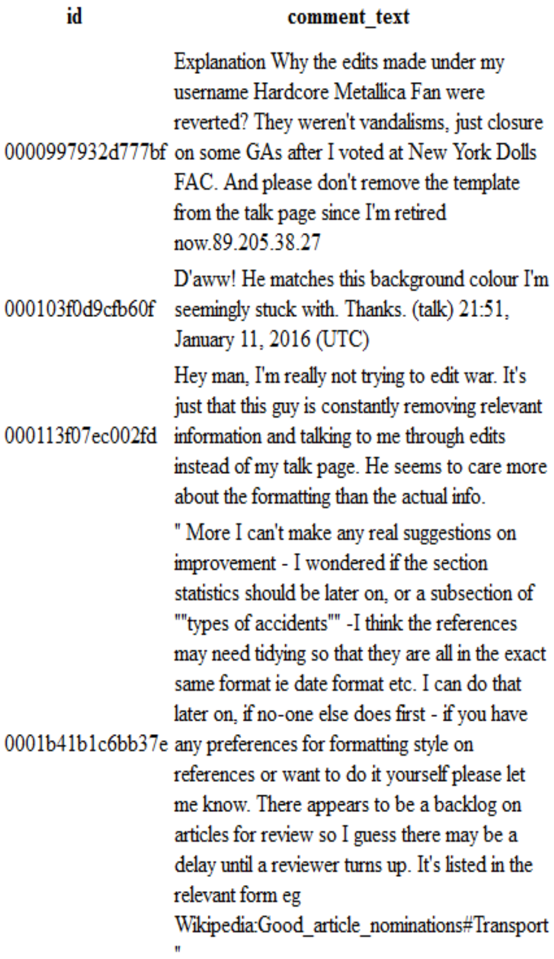
bastard pro assad.hanibal911you're
바샤르알 - 아사드 대통령
useless bot fuck : 쓸모없는 XXX
shit cunts fucking : 젠장 빌어먹을

Die jim wales

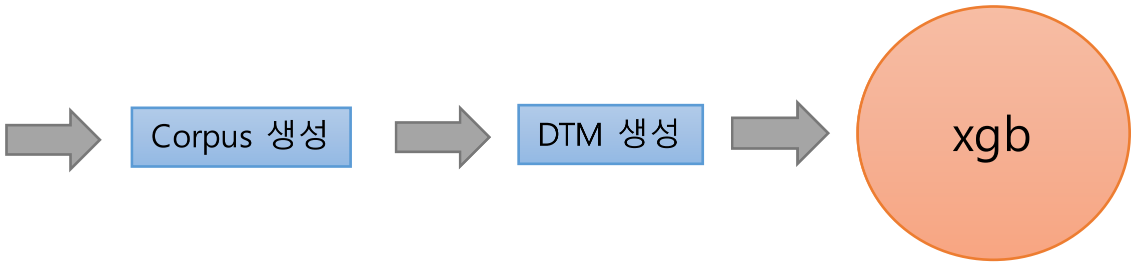
3. Modeling

참고커널 : toxic Comments EDA and XGB Modelling

Preparing data for model



id	comment_text
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.
0001b41b1c6bb37e	" More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have any preferences for formatting style on references or want to do it yourself please let me know. There appears to be a backlog on articles for review so I guess there may be a delay until a reviewer turns up. It's listed in the relevant form eg Wikipedia:Good_article_nominations#Transport "



Model training

Params

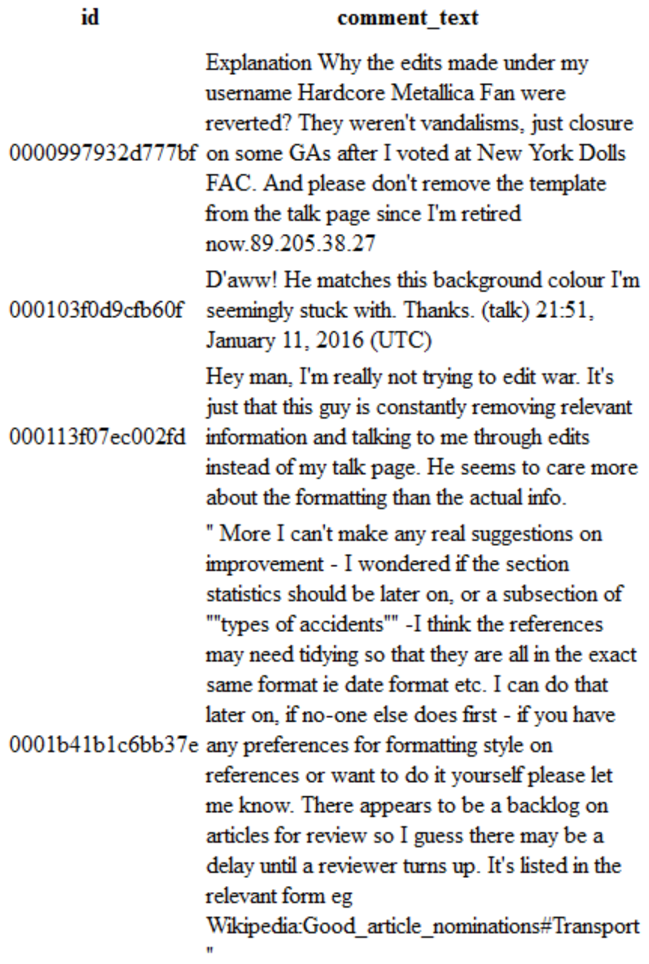
- nrounds = 500
- eta = 0.05
- max_depth = 3
- min_child_weight = 1
- subsample = 1
- colsample_bytree = 0.8

=> Score : 0.8376

3. Modeling-참고

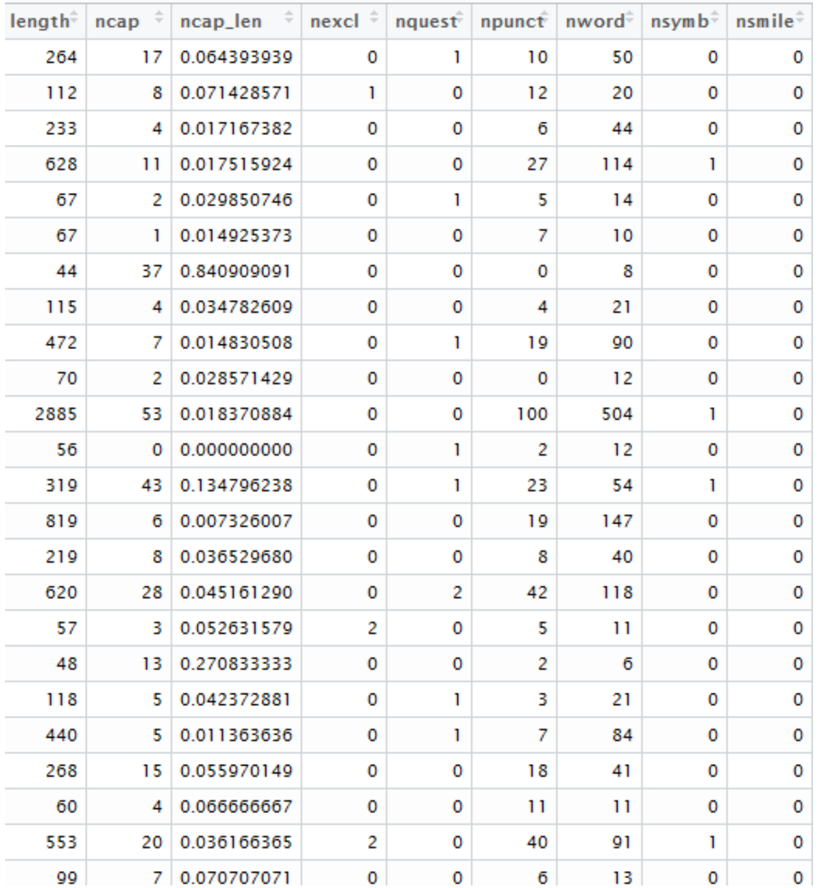
참고커널 : tidy xgboost + glmnet + text2vec + LSA

Preparing data for model



id	comment_text
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.
0001b41b1c6bb37e	" More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have any preferences for formatting style on references or want to do it yourself please let me know. There appears to be a backlog on articles for review so I guess there may be a delay until a reviewer turns up. It's listed in the relevant form eg Wikipedia:Good_article_nominations#Transport "

컬럼 생성	
Length	Text length
Ncap	A-Z length
Ncap_len	Ncap/Length
Nexcl	! 개수
Nquest	? 개수
Npunct	Punct 개수
Nword	단어수
Nsymb	특수문자 개수
nsmile	이모티콘 개수 :)

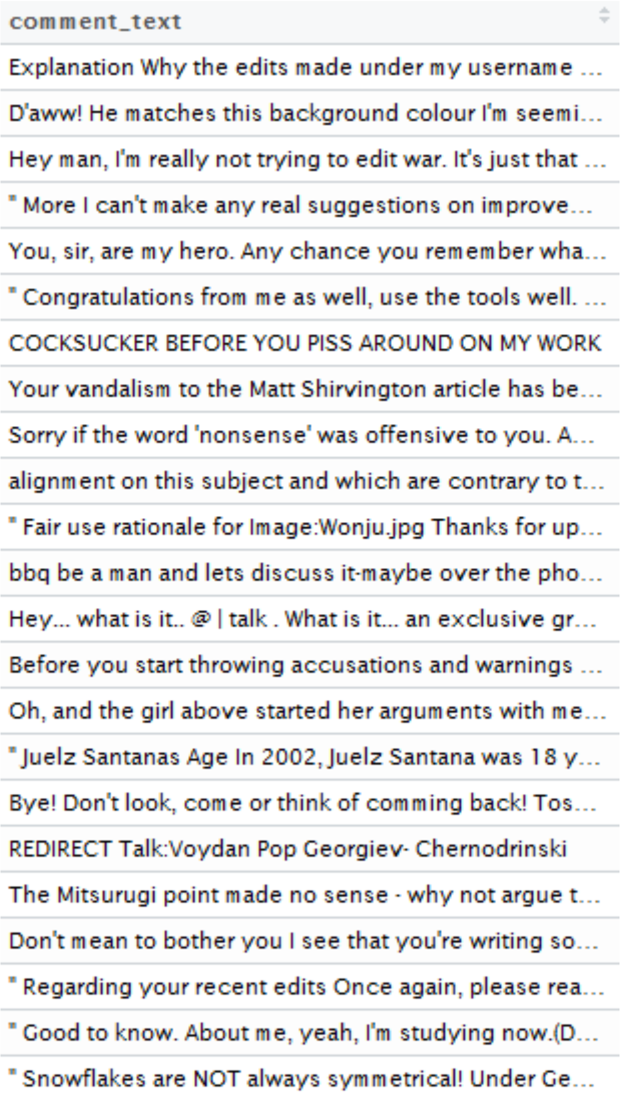


length	ncap	ncap_len	nexcl	nquest	npunct	nword	nsymb	nsmile
264	17	0.064393939	0	1	10	50	0	0
112	8	0.071428571	1	0	12	20	0	0
233	4	0.017167382	0	0	6	44	0	0
628	11	0.017515924	0	0	27	114	1	0
67	2	0.029850746	0	1	5	14	0	0
67	1	0.014925373	0	0	7	10	0	0
44	37	0.840909091	0	0	0	8	0	0
115	4	0.034782609	0	0	4	21	0	0
472	7	0.014830508	0	1	19	90	0	0
70	2	0.028571429	0	0	0	12	0	0
2885	53	0.018370884	0	0	100	504	1	0
56	0	0.000000000	0	1	2	12	0	0
319	43	0.134796238	0	1	23	54	1	0
819	6	0.007326007	0	0	19	147	0	0
219	8	0.036529680	0	0	8	40	0	0
620	28	0.045161290	0	2	42	118	0	0
57	3	0.052631579	2	0	5	11	0	0
48	13	0.270833333	0	0	2	6	0	0
118	5	0.042372881	0	1	3	21	0	0
440	5	0.011363636	0	1	7	84	0	0
268	15	0.055970149	0	0	18	41	0	0
60	4	0.066666667	0	0	11	11	0	0
553	20	0.036166365	2	0	40	91	1	0
99	7	0.070707071	0	0	6	13	0	0

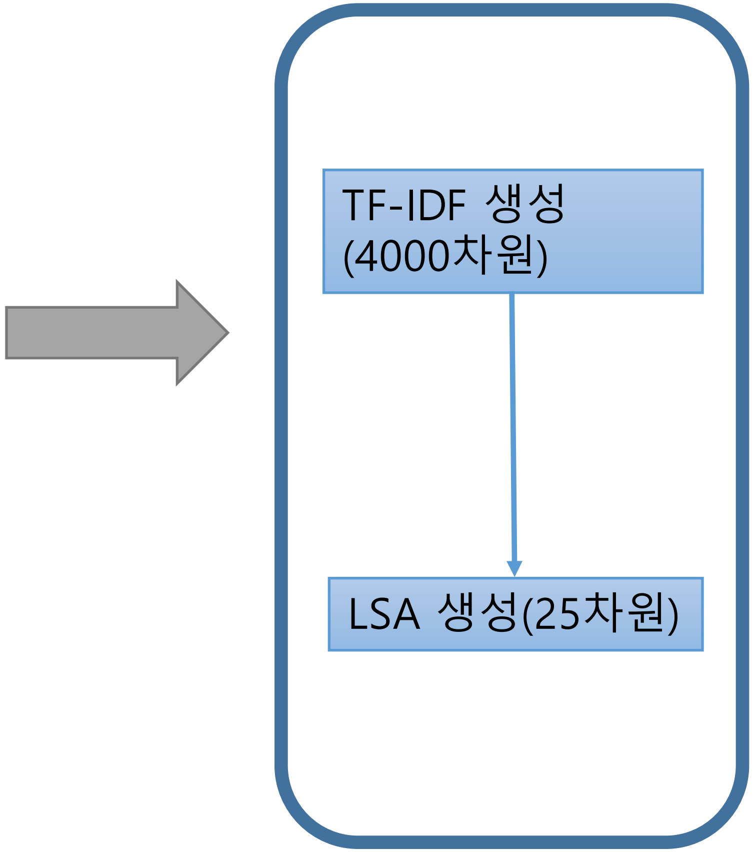
3. Modeling-참고

참고커널 : tidy xgboost + glmnet + text2vec + LSA

Preparing data for model



comment_text
Explanation Why the edits made under my username ...
D'aww! He matches this background colour I'm seemi...
Hey man, I'm really not trying to edit war. It's just that ...
" More I can't make any real suggestions on improve...
You, sir, are my hero. Any chance you remember wha...
" Congratulations from me as well, use the tools well. ...
COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK
Your vandalism to the Matt Shirvington article has be...
Sorry if the word 'nonsense' was offensive to you. A...
alignment on this subject and which are contrary to t...
" Fair use rationale for Image:Wonju.jpg Thanks for up...
bbq be a man and lets discuss it-maybe over the pho...
Hey... what is it.. @ | talk . What is it... an exclusive gr...
Before you start throwing accusations and warnings ...
Oh, and the girl above started her arguments with me...
" Juelz Santanas Age In 2002, Juelz Santana was 18 y...
Bye! Don't look, come or think of comming back! Tos...
REDIRECT Talk:Voydan Pop Georgiev- Chernodrinski
The Mitsurugi point made no sense - why not argue t...
Don't mean to bother you I see that you're writing so...
" Regarding your recent edits Once again, please rea...
" Good to know. About me, yeah, I'm studying now.(D...
" Snowflakes are NOT always symmetrical! Under Ge...



TF-IDF 생성 (4000차원)

LSA 생성(25차원)

#TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

#IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

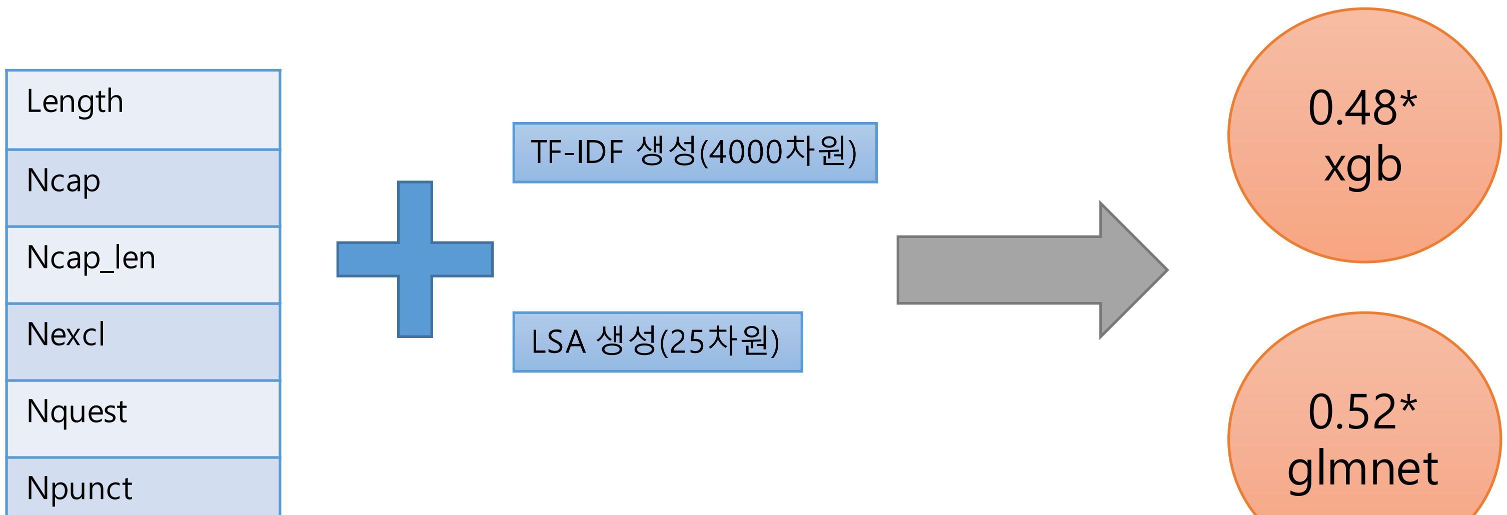
#Value = TF * IDF

3. Modeling-참고

참고커널 : tidy xgboost + glmnet + text2vec + LSA

Input

Model training



Length
Ncap
Ncap_len
Nexcl
Nquest
Npunct
Nword
Nsymb
nsmile

TF-IDF 생성(4000차원)

LSA 생성(25차원)

0.48* xgb

0.52* glmnet

LSA(잠재 의미 분석)
자연어 처리 기법으로 단어나 문서들 사이의 의미론적 관계를 분석하는 기법, 유사한 의미를 갖는 단어들은 비슷한 문맥에서 등장할 것이라 가정

단락 당 단어의 수를 포함하는 행렬은 대용량의 텍스트와 SVD로 구성되며 이는 컬럼들 사이의 유사성을 보존하며 행의 수를 줄일 수 있음 두 개의 행으로부터 얻어진 두 Vector의 각으로 Cosine을 취해 단어들을 비교할 수 있음 [0,1] 사이의 값을 가지며 1에 가까울수록 높은 연관성을 가짐

Params

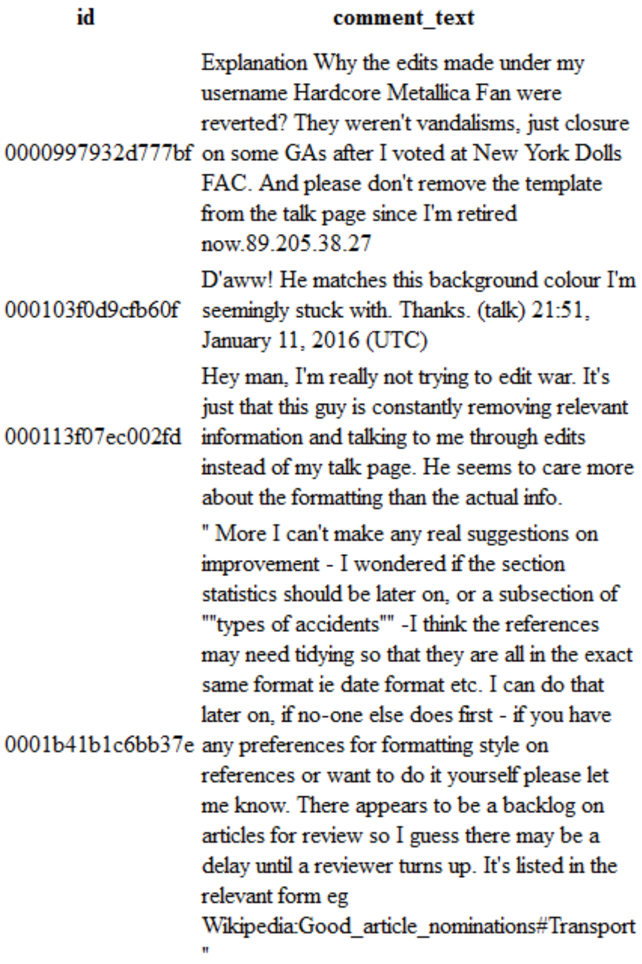
- nthread = 4
- eta = 0.2
- max_depth = 6
- min_child_weight = 4
- subsample = 0.7
- colsample_bytree = 0.7

=> Score : 0.9788

3. Modeling (Deep learning)

참고커널 : Two RNN+CNN Columns networks with keras

Input

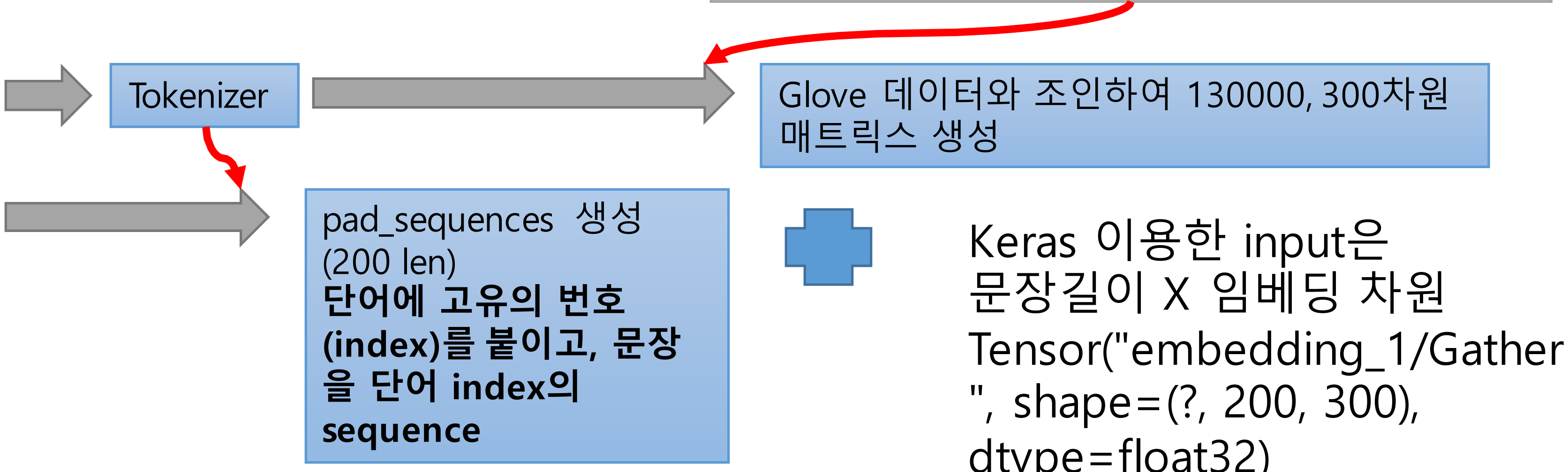


id	comment_text
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.
0001b41b1c6bb37e	" More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have any preferences for formatting style on references or want to do it yourself please let me know. There appears to be a backlog on articles for review so I guess there may be a delay until a reviewer turns up. It's listed in the relevant form eg Wikipedia:Good_article_nominations#Transport "

Pretrained Word Embedding Model

외부 데이터 (glove.840B.300d.txt)
-> 840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB
(https://nlp.stanford.edu/projects/glove/)

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.



Tokenizer

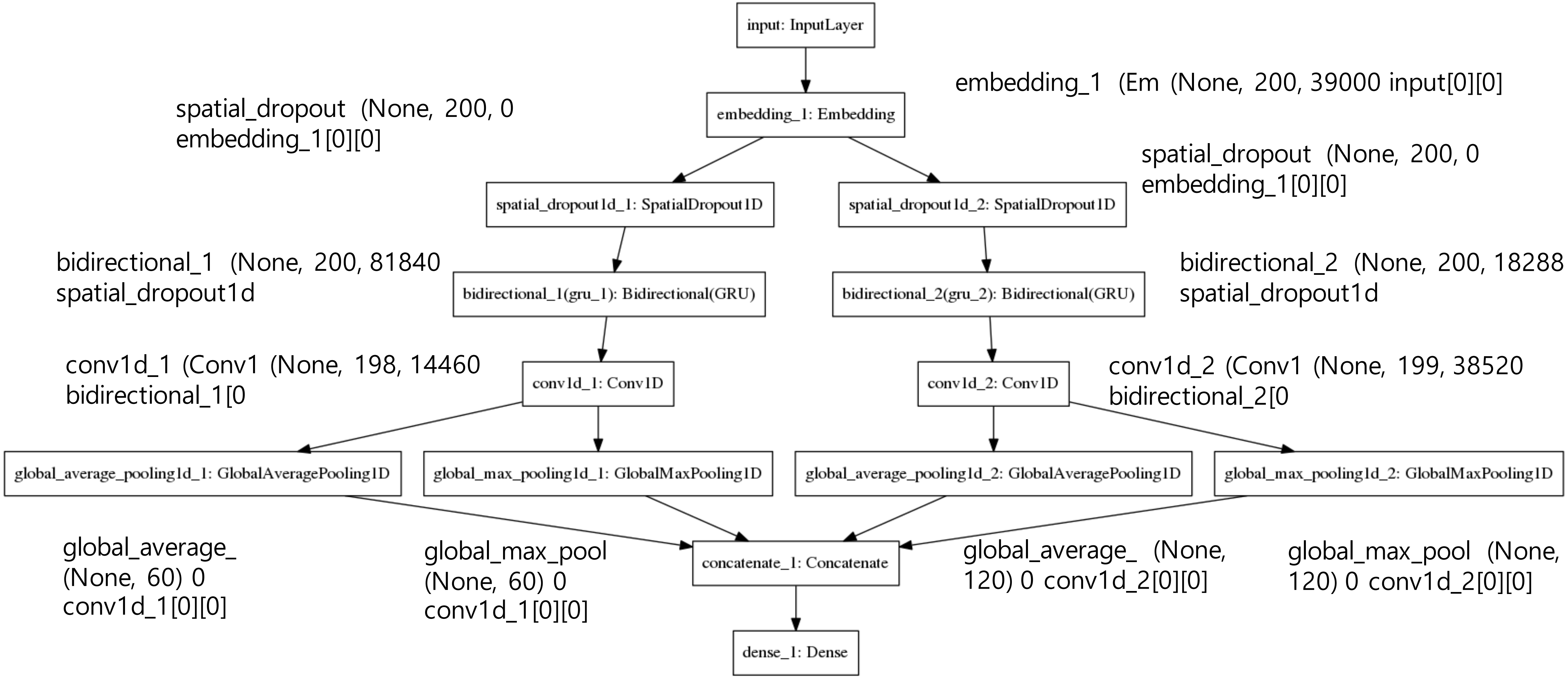
Glove 데이터와 조인하여 130000, 300차원 매트릭스 생성

pad_sequences 생성 (200 len)
단어에 고유의 번호(index)를 붙이고, 문장을 단어 index의 sequence

Keras 이용한 input은
문장길이 X 임베딩 차원
Tensor("embedding_1/Gather:0", shape=(?, 200, 300), dtype=float32)
=> 130,000, 200, 300

3. Modeling (Deep learning : Keras Model)

참고커널 : Two RNN+CNN Columns networks with keras


input: InputLayer
embedding_1: Embedding
spatial_dropout1d_1: SpatialDropout1D
spatial_dropout1d_2: SpatialDropout1D
bidirectional_1(gru_1): Bidirectional(GRU)
bidirectional_2(gru_2): Bidirectional(GRU)
conv1d_1: Conv1D
conv1d_2: Conv1D
global_average_pooling1d_1: GlobalAveragePooling1D
global_max_pooling1d_1: GlobalMaxPooling1D
global_average_pooling1d_2: GlobalAveragePooling1D
global_max_pooling1d_2: GlobalMaxPooling1D
concatenate_1: Concatenate
dense_1: Dense

embedding_1 (Em (None, 200, 39000 input[0][0]

spatial_dropout (None, 200, 0
embedding_1[0][0]

spatial_dropout (None, 200, 0
embedding_1[0][0]

bidirectional_1 (None, 200, 81840
spatial_dropout1d

bidirectional_2 (None, 200, 18288
spatial_dropout1d

conv1d_1 (Conv1 (None, 198, 14460
bidirectional_1[0

conv1d_2 (Conv1 (None, 199, 38520
bidirectional_2[0

global_average_
(None, 60) 0
conv1d_1[0][0]

global_max_pool
(None, 60) 0
conv1d_1[0][0]

global_average_ (None,
120) 0 conv1d_2[0][0]

global_max_pool (None,
120) 0 conv1d_2[0][0]

concatenate_1 ((None, 360) 0 global_average_po
global_max_poolin global_average_po
global_max_poolin

dense_1 (Dense) (None, 6) 2166 concatenate_1[0][

Param : Epochs = 2, batch_size = 32, vaildation_split = 0.05

=> Score : 0.9828

끝 (Q&A)

1st place solution overview

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557

- Diverse pre-trained embeddings (baseline public LB of 0.9877)
- Translations as train/test-time augmentation (TTA) (boosted LB from 0.9877 to 0.9880)
- Rough-bore pseudo-labelling (PL) (boosted LB from 0.9880 to 0.9885)
- Robust CV + stacking framework (boosted LB from 0.9885 to 0.9890)