Team NLP

강유정 김근호 김나연
김정현 박진우 임형준

KU-BIG

Index

1. Text Classification
2. Talking to the Little Prince ver.2
3. Project Plan

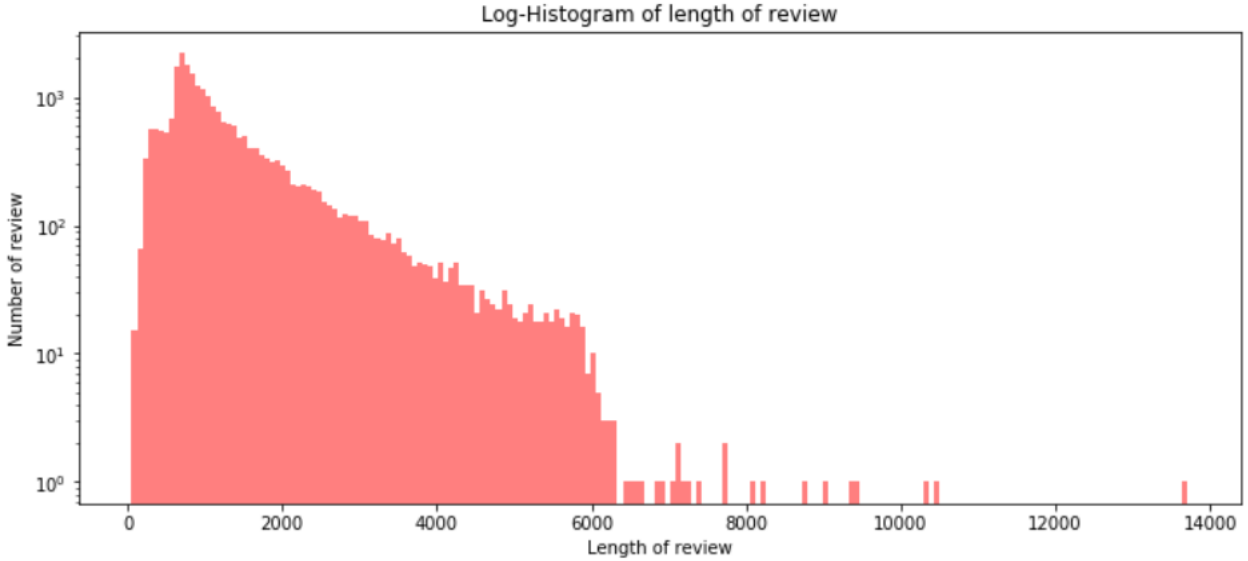
KU-BIG

1. Text classification

	id	sentiment	review
0	"5814_8"	1	"With all this stuff going down at the moment ...
1	"2381_9"	1	"\"The Classic War of the Worlds\" by Timothy ...
2	"7759_3"	0	"The film starts with a manager (Nicholas Bell...
3	"3630_4"	0	"It must be assumed that those who praised thi...
4	"9495_8"	1	"Superbly trashy and wondrously unpretentious ...

KU-BIG

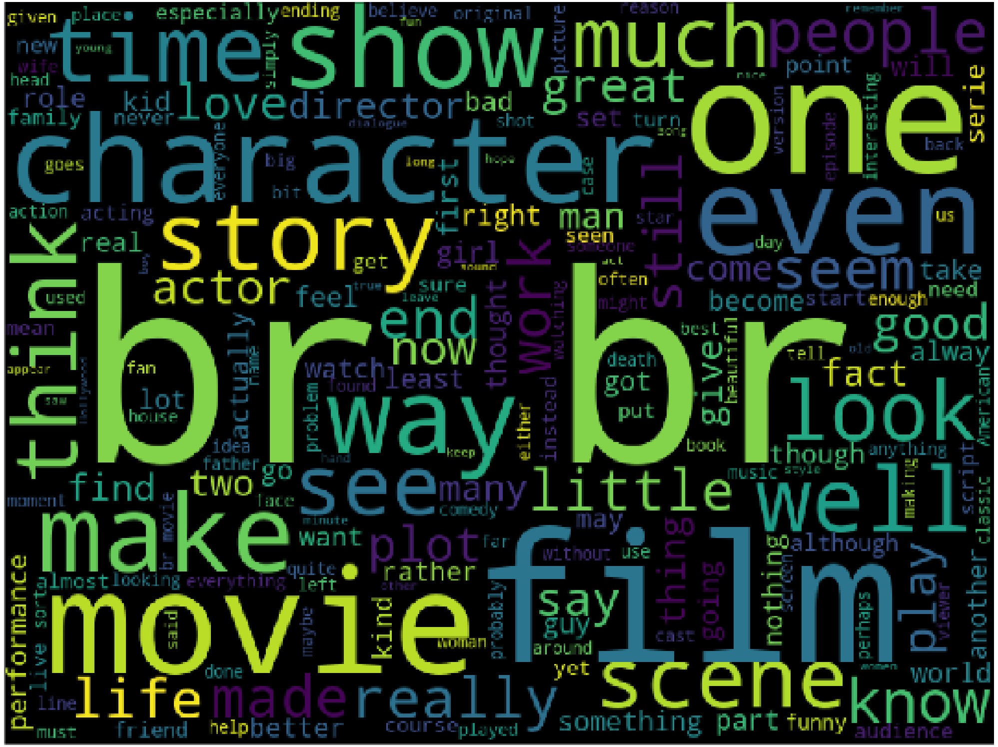
1. Text classification - EDA



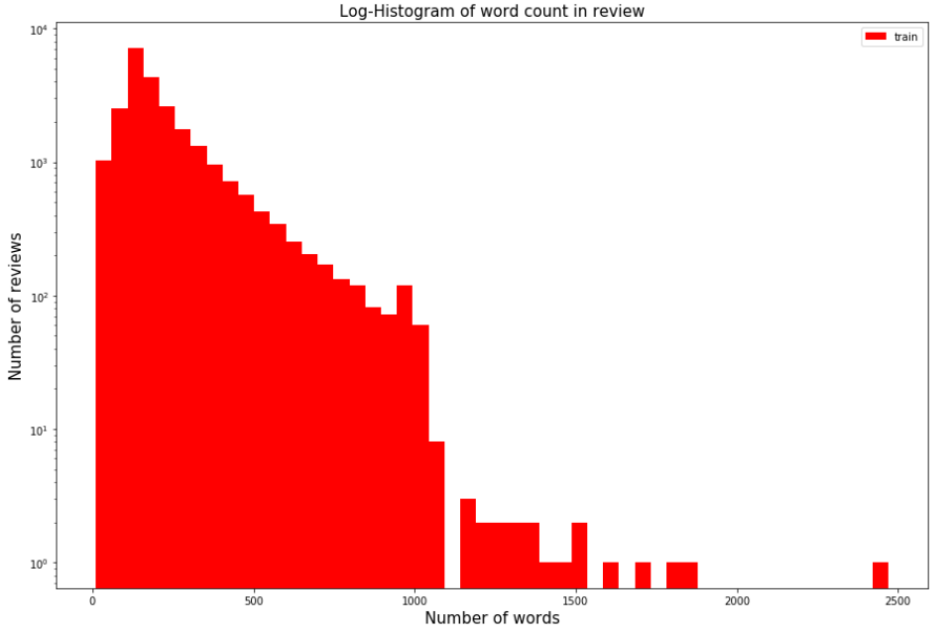

```
리뷰 길이 최대 값: 13710
리뷰 길이 최소 값: 54
리뷰 길이 평균 값: 1329.71
리뷰 길이 표준편차: 1005.22
리뷰 길이 중간 값: 983.0
리뷰 길이 제 1 사분위: 705.0
리뷰 길이 제 3 사분위: 1619.0
```

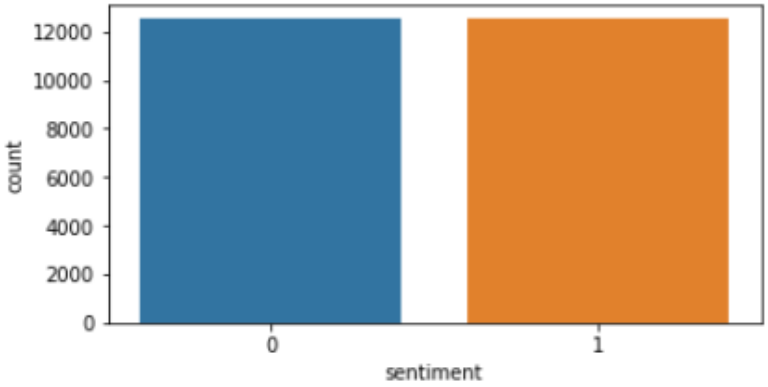

KU-BIG

1. Text classification - EDA



1. Text classification - EDA





1. Text classification - preprocessing

"With all this stuff going down at the moment with MJ i've started listening to his music, watching the odd documentary here and there, watched The Wiz and watched Moonwalker again.‹br /›‹br /›

With all this stuff going down at the moment with MJ i ve started listening to his music watching the odd documentary here and there watched The Wiz and watched Moonwalker again

- ✓ Beautiful Soup
- ✓ re.sub

['stuff', 'going', 'moment', 'mj', 'started', 'listening', 'music', 'watching', 'odd', 'documentary', 'watched', 'wiz', 'watched', 'moonwalker']

- ✓ Lower, split
- ✓ NLTK

1. Text classification - preprocessing

stuff going moment mj started listening
music watching odd documentary watched
wiz watched moonwalker

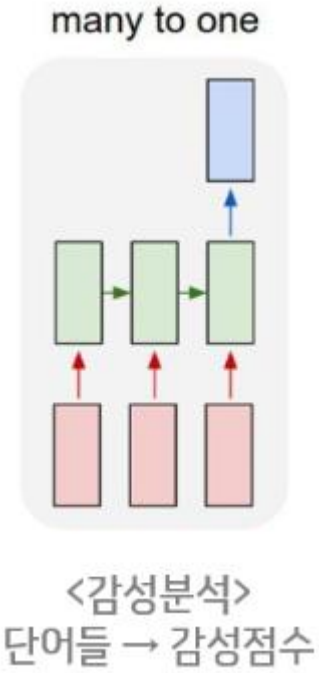
[404, 70, 419, 8815, 506, 2456, 115, 54, 873,
516, 178, 18686, 178, 11242]

✓ Tokenizer

[404, 70, 419, 8815, 506, 2456, 115, 54, 873,
516, 178, 18686, 178, 11242, 0, 0, 0, 0 …]

✓ Pad sequences

1. Text classification - RNN



<감성분석>
단어들 → 감성점수

학습과 검증 데이터셋 분리

데이터 입력 함수 구현

모델 구현

모델 학습, 평가, 예측을 위한 구현

학습 및 성능 검증

1. Text classification - RNN

학습과 검증 데이터셋 분리

```
from sklearn.model_selection import train_test_split
TEST_SPLIT = 0.1
RANDOM_SEED = 13371447

train_input, test_input, train_label, test_label = train_test_split(input_data, label_data,
                                                                    test_size=TEST_SPLIT, random_state=RANDOM_SEED)
```

데이터 입력 함수 구현

```
BATCH_SIZE = 16
NUM_EPOCHS = 3

def mapping_fn(X, Y):
    inputs, labels = {'x': X}, Y
    return inputs, labels

def train_input_fn():
    dataset = tf.data.Dataset.from_tensor_slices((train_input, train_label))
    dataset = dataset.shuffle(buffer_size=50000)
    dataset = dataset.batch(BATCH_SIZE)
    dataset = dataset.repeat(count=NUM_EPOCHS)
    dataset = dataset.map(mapping_fn)
    iterator = dataset.make_one_shot_iterator()

    return iterator.get_next()

def eval_input_fn():
    dataset = tf.data.Dataset.from_tensor_slices((test_input, test_label))
    dataset = dataset.map(mapping_fn)
    dataset = dataset.batch(BATCH_SIZE * 2)
    iterator = dataset.make_one_shot_iterator()

    return iterator.get_next()
```

1. Text classification - RNN

모델 구현

```
def model_fn(features, labels, mode):
    TRAIN = mode == tf.estimator.ModeKeys.TRAIN
    EVAL = mode == tf.estimator.ModeKeys.EVAL
    PREDICT = mode == tf.estimator.ModeKeys.PREDICT

    embedding_layer = tf.keras.layers.Embedding(
                    VOCAB_SIZE,
                    WORD_EMBEDDING_DIM)(features['x'])

    embedding_layer = tf.keras.layers.Dropout(0.2)(embedding_layer)

    rnn_layers = [tf.nn.rnn_cell.LSTMCell(size) for size in [HIDDEN_STATE_DIM, HIDDEN_STATE_DIM]]
    multi_rnn_cell = tf.nn.rnn_cell.MultiRNNCell(rnn_layers)

    outputs, state = tf.nn.dynamic_rnn(cell=multi_rnn_cell,
                                       inputs=embedding_layer,
                                       dtype=tf.float32)

    outputs = tf.keras.layers.Dropout(0.2)(outputs)
    hidden_layer = tf.keras.layers.Dense(DENSE_FEATURE_DIM, activation=tf.nn.tanh)(outputs[:,-1,:])
    hidden_layer = tf.keras.layers.Dropout(0.2)(hidden_layer)
    logits = tf.keras.layers.Dense(1)(hidden_layer)
    logits = tf.squeeze(logits, axis=-1)
```

1. Text classification - RNN

모델 학습, 평가, 예측을 위한 구현

```
loss = tf.losses.sigmoid_cross_entropy(labels, logits)

if TRAIN:
    global_step = tf.train.get_global_step()
    train_op = tf.train.AdamOptimizer(learning_rate).minimize(loss, global_step)

    return tf.estimator.EstimatorSpec(
              mode=mode,
              train_op=train_op,
              loss=loss)

sigmoid_logits = tf.nn.sigmoid(logits)

if EVAL:
    accuracy = tf.metrics.accuracy(labels, tf.round(sigmoid_logits))
    eval_metric_ops = {'acc': accuracy}

    return tf.estimator.EstimatorSpec(mode, loss=loss, eval_metric_ops=eval_metric_ops)

if PREDICT:
    predictions = {'sentiment': sigmoid_logits}

    return tf.estimator.EstimatorSpec(
              mode=mode,
              predictions=predictions)
```

학습 및 성능 검증

```
est = tf.estimator.Estimator(model_fn=model_fn,
                             model_dir=DATA_OUT_PATH + 'checkpoint/rnn')
est.train(train_input_fn)
est.evaluate(eval_input_fn)
```

```
Out[23]: {'acc': 0.7992, 'loss': 0.50662565, 'global_step': 4164}
```

2. Talking to the Jung gu

뭐? 누가죽어? 누가죽어?!!! 놔 이 새끼야!! 놔!
야이 씨발 비싼 돈 쳐먹었으면 죽은 사람도 살려내야 될 거 아냐!!
놔!! 안놔?!! 안놔?!!!!이
씨발 여기 책임자 나오라 그래! 책임자 나오라 그래애애애!!! 씨이이빨!
이새끼들 봐라 이거, 야 지금 니들 누구 허락 받고 찍어대는거냐?
이거 엄연히 초상권 침해야~ 알어?
꼬라지들 보아하니 문상 온것 같진 않고, 니들 어디냐? 어디서 들어왔어?
ㅎ. 말 안해? 벙어리야?
야!! 그냥 이 새끼들 낯짝도 다 찍어놔라.
혹시라도 나중에 또 볼 일이 있을지 모르잖냐.
야, 근데 니네들 왜 이렇게 들러들 붙는건데?
니들 지금 우리 스토킹하냐?어, 이 개새끼들아?
이거 누군가 했더니 우리 강팀장님 아니세요?
에~~~이 씨발.ㅋ 저렇게 대놓고 찍어대는데 어떻게 못 본 척을 하나?
내 눈이 삐꾸도 아니고
근데 조심좀 허지? 그 잘난 혀가 댁의 목줄을 끊어 놓는 수가 있거든.
쫌 더 챙겨 놨으니까. 강팀장 담뱃값 해.ㅋ
아~~이 나, 씨발놈의 짜바리 새끼
물론, 안 괜찮을거 있나? 그룹장 자리를 오래 비워둬서 뭐, 좋을 거 없겠지.
거 대충 얘기들 끝나셨음. 전 먼저 일어 납니다.
거 됐슈다.
솔직히 우리가 피차 낯짝 마주보고 정~답게 앉아 밥 쳐먹을 그런 살가운 사이는 아니잖수.

KU-BIG

2. Talking to the Jung gu

> 안녕하세요 중구형
우리 ? ? ? ?

> 네?
근데 , ? ? ?

> 뭐야 왜이래
근데 여기 ? ? ?

> 중구형 거 장난이 너무 심한거 아니오?
거 ? . ? 응

> 고향이 어디냐
뭐 ? . ?

> 왜리래요 ㅠㅠ 대답좀 해요 ㅠㅠ
뭐 , . ? ? 응

> 개망했다
씨발

> 뭐인마?
씨발 . ? . ?

> 너 구글 직원이지 새끼야
그럼 , 응 ?

> ?? 실명까라 어디냐?
거 , 밥 안보이냐 ? ?

> ?? 삭제되고싶냐 로봇새끼야?
뭐 , ? 안보이냐 . ? ?

> 뭐가 자식아
뭐 . . ?

> 됐다 넌 실패작이야 삭제다
씨발 , ? ? ? 응

> 잘가라
? ?

> 죽기 딱 좋은 날씨다 자식아

> 담배는 없다 자식아
뭐 ? . ? ?

KU-BIG

3. Project Plan

SHERLOCK

S01E01 **ASTUDYINPINK**

Therapist: How's your blog going?
John Watson: Yeah, good, very good.
Therapist: You haven't written a word, have you?
John Watson: You just wrote "still has trust issues"
Therapist: And you read my writing upside down. You see what I mean? John, you're a soldier and it's going to take you a while to adjust to civilian life and writing a blog about everything that happens to you will honestly help you.
John Watson: Nothing happens to me.

Jeffrey Patterson: What do you mean there's no ruddy car?
Woman: He went to Waterloo, I'm sorry. Get a cab!
Jeffrey Patterson: I never get cabs!
Woman: I love you.
Jeffrey Patterson: When?
Woman: Get a cab!

KU-BIG