






Week6

 Created By	 DongGu Kim
 Last Edited	@Apr 24, 2020 5:05 PM
 Property	
 Tags	

1. Preprocessing Text

가. 특수문자 갯수를 나타내는 Feature 추가

나. 문자메시지 길이를 나타내는 Feature 추가

다. 원본 문자메시지 변형

- 1) 알파벳 소문자로 통일시키기
- 2) 불용어, 일반/스미싱 문자 공통적으로 많이 쓰인 단어 제거
- 3) 특수문자 제거

참고)

- DataFrame 바꿔주는 기본 함수
- Global Variables (특수문자, 불용어)

2. 워드 임베딩

3. Trainset - Validationset - Testset 구분하기

1. Preprocessing Text

가. 특수문자 갯수를 나타내는 Feature 추가

```
%%time
train['spe_num'] = 0
for idx, text in enumerate(train['text']):
    num = 0
    for spe in spe_char_list:
        temp = text.count(spe)
        num += temp
    train['spe_num'][idx] = num
```

나. 문자메시지 길이를 나타내는 Feature 추가

```
%%time
train['length'] = train['text'].apply(lambda x: len(x))
```

다. 원본 문자메시지 변형

1) 알파벳 소문자로 통일시키기

```
%%time
def alp_to_lower(mms):
    """ (str) -> str
    Return the mms after changing Upper to Lower
    >> alp_to_lower("Kim Dong Gu")
    kim dong gu
    """
    mms = mms.lower()
    return mms
train = apply_replacement(train, alp_to_lower)
```

2) 불용어, 일반/스미싱 문자 공통적으로 많이 쓰인 단어 제거

```
%%time
def del_stop_words(mms):
    """(str) -> str
    Return the mms after deleting stop words
    >>> del_stop_words('사랑했습니다')
    사랑
    """
    for word in stop_words:
        mms = mms.replace(word, '')
    return mms

train = apply_replacement(train, del_stop_words)
```

3) 특수문자 제거

```
%%time
def del_spe_char(mms):
    """(str) -> str
    Return the mms after deleting special charaters
    >> del_spe_char("$5000달러 환전")
    5000달러 환전
    """
    mms = re.sub(spe_char, '', mms)
    return mms

train = apply_replacement(train, del_spe_char)
```

참고)

- 특수문자
- Global Variables

```
#spe_char, spe_char_list: 특수문자
#stop_words: 불용어

spe_char = "[-+=#/\?:^$@*~&%·!~\`'|\\(\`)\`[\`]\`<\`>`\`'...>\`$]"
spe_char_list = list(spe_char)
stop_words = ['x', 'xx', 'xxx', '으로', '습니다', '까지', '합니다', '에서', '입니다', '셔서', '세요']
```

2. 워드 임베딩 (Word Embedding)

- 자연어를 컴퓨터가 이해하고, 효율적으로 처리하게 하기 위해서는 컴퓨터가 이해할 수 있도록 자연어를 적절히 변환할 필요가 있습니다.
- 참고: transform, fit_transform 차이

Difference between fit and fit_transform in scikit_learn models?

Data Science Stack Exchange is a question and answer site for Data science professionals, Machine Learning specialists, and those interested in learning more about the field. It only takes a minute to sign up. Sign up

<https://datascience.stackexchange.com/questions/12321/difference-between-fit-and-fit-transform-in-scikit-learn-models>



```
# DTM 문서 단어 행렬
from sklearn.feature_extraction.text import CountVectorizer
vector = CountVectorizer()

# TF-IDF
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
tfvector = TfidfTransformer(smooth_idf=False)
```

```
vectorizer=TfidfVectorizer(ngram_range=(1, 3),
                           min_df=2,
                           max_features=10000,
                           sublinear_tf=True,
                           lowercase=False,
                           use_idf=True)
```

```
%%time
vec_train = vectorizer.fit_transform(train['text'])
vec_df = pd.DataFrame(vec_train, columns = ['vec_text'], index = train.index)
```

3. Trainset - Validationset - Testset 구분하기

```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
```

```
import lightgbm
import joblib
```

```
s_train = pd.concat([vec_df, train[['spe_num', 'length']]], axis=1)
s_label = train['smishing']
```

```
# # train - test 분류
%%time
X_train, X_test, y_train, y_test = train_test_split(s_train,
                                                    s_label,
                                                    test_size = 0.2,
                                                    shuffle = True,
                                                    random_state = 3077)

# # train - val 분류
X_train, X_val, y_train, y_val = train_test_split(s_train,
                                                    s_label,
                                                    test_size = 0.3,
                                                    shuffle = True,
                                                    random_state = 3077)
```