






Week10

 Created By	 DongGu Kim
 Last Edited	@May 22, 2020 6:39 PM
 Property	
 Tags	

False records 분석

웹사이트 제작 (flask)

깃허브 진행 상황 업로드

To Do list

False records EDA

1. 예측이 잘못된 record 수: 222개(0.01% 정도)

- 전체 record : 약 30만 개
- false_positive: 197개

[illegible]

- false_negative: 25개



2. SVM Model 30개가 예측한 파일 30개 중 예측이 잘못된 primary key값인 ID를 합집합으로 합산함

→ LIME은 하나의 모델로 판단 근거를 보여주기 때문에 30개의 모델의 예측값이 다를 경우 문자메시지 해석이 애매함.

3. SVM Model의 false record를 각각 dictionary형태로 수합한 후, 30개 모델이 공통적으로 틀린(교집합) record를 우선적으로 분석함. (진행 중)

웹사이트 제작 (flask)

1. 가상환경 설치

```
python -m venv myvenv
```

2. 'templates/main.html'

- Input: 문자메시지 내용
- Output: 스미싱 문자일 확률, 판단에 대한 근거

```
<!doctype html>
<html>

<head>
  <title>Smishing? Smash!</title>
```

```

</head>
<form action="{{ url_for('main') }}" method="POST">
    <fieldset>
        <legend>Input values:</legend>
        Message:
        <input name="message" type="text" required>
        <br>
        <br>
        <input type="submit">
    </fieldset>
</form>

<br>
<div class="result" align="center">
    {% if result %}
        {% for variable, value in original_input.items() %}
            <b>{{ variable }}</b> : {{ value }}
        {% endfor %}
    <br>
    <br> Predicted number of bikes in use:
    <p style="font-size:50px">{{ result }}</p>
    {% endif %}
</div>

</html>

```

3. app.py (delpoy model)

1. **Input Data (string → sparse matrix 변환 문제)**

```

import flask
import pickle
import pandas as pd

from sklearn.svm import LinearSVC
from sklearn.calibration import CalibratedClassifierCV
import joblib

from flask import Flask, jsonify, request
import os

#####
##### INPUT DATA 전처리 #####
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer

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

```

```

model = joblib.load('model/svm_model.pkl')

app = flask.Flask(__name__, template_folder='templates')

@app.route('/', methods=['GET', 'POST'])
def main():
    if flask.request.method == 'GET':
        # Just render the initial form, to get input
        return(flask.render_template('main.html'))

    if flask.request.method == 'POST':
        # Extract the input
        prepared_train = pd.read_csv('model/prepared_train.csv', encoding='utf-8')
        del prepared_train['Unnamed: 0']
        vec_train = vectorizer.fit_transform(prepared_train['origin'][:1000])

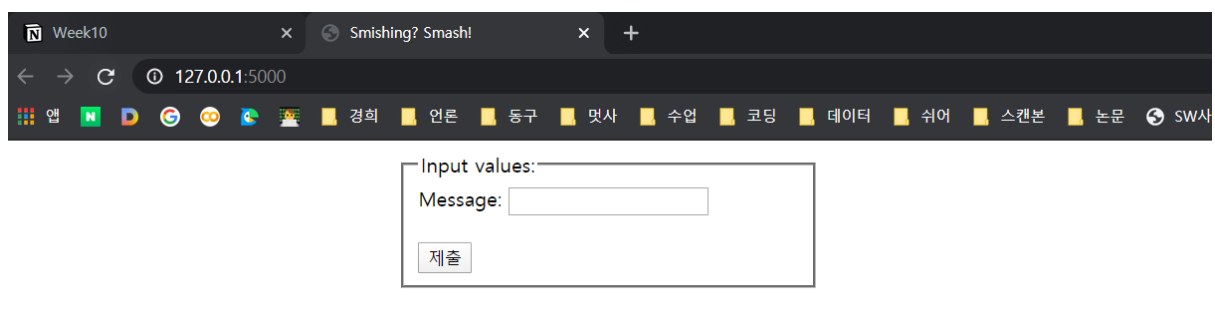
        message = flask.request.form['message']
        message_data = vectorizer.transform([message]).toarray()
        prediction = model.predict(message_data)

        return flask.render_template('main.html',
                                     original_input={'Message':message},
                                     result=prediction,
                                     )

if __name__ == '__main__':
    app.run()

```

4. flask run




깃허브 진행 상황 업로드

- 주소

ehdrn463/dataanalysis_capstone

20-1학기 데이터분석캡스톤디자인. Contribute to ehdrn463/dataanalysis_capstone development by creating an account on GitHub.

 https://github.com/ehdrn463/dataanalysis_capstone

20-1학기 데이터분석캡스톤디자인

Edit

[Manage topics](#)

28 commits

1 branch

0 packages

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

ehdrn463 Update README.md

Latest commit c169e38 15 hours ago

Weekly Notebook

Update README.md

15 hours ago

Weekly_Report

Add files via upload

15 hours ago

README.md

Update README.md

15 hours ago

README.md



경희대학교 2020-1학기 데이터분석캡스톤디자인

주제: 스미싱 문자를 판독해주는 챗봇

스미싱 문자 데이터 출처

데이콘 금융 문자 분석 경진대회 <https://dacon.io/competitions/official/235401/overview/>

데이콘, 스폰서의 협의로 인해 대회 종료 후 데이터 다운로드 불가능합니다.

따라서 repository에서도 제공받은 데이터를 공개하지 않을 것입니다.

7	하이퍼 파라미터 선택		1 주		완료
8	LSTM 모델 구축 및 검증		2 주		LGBM, SVM 성능이 우수함 -> 생략
9	하이퍼 파라미터 선택		2 주		LGBM, SVM 성능이 우수함 -> 생략
10	LightGBM과 LSTM 모델 중 정확도, 속도를 고려하여 더 우수한 모델 선택		2 주		SVM(속도, F-measure) 모두 우수
11	LIME 학습 및 모델 적용		3 주		완료
12	카카오톡·라인 플러스 친구, 웹사이트 중 이번 모델에 적합한 플랫폼 선정		4 주		카카오톡·라인 모두 고려 중
13	8)의 a에 의해 선정된 플랫폼 학습		4 주		진행 중
14	웹 서비스 개발			1, 2 주	진행 중
15	웹서비스 배포			2, 3 주	진행 예정

To Do list

- Input Data: string to sparse data
- 웹 서비스 배포
- 챗봇 구현
- SVM Model의 false record를 각각 dictionary형태로 수합한 후, 30개 모델이 공통적으로 틀린(교집합) record를 우선적으로 분석함. (진행 중)
 - 20개 모델이 틀린 record: 141개