Лабораторная работа №3

Тема «Применение логистической регрессии для определения вредоносных URLs »

Дан файл data.csv, в котором содержатся около 400 000 URL как вредоносных так и безопасных. Записи в файле имеют следующий вид:

floridarentfinders.com/uploads/ws/css/www.paypal.com/cgi-bin/webscrcmd=_login-run/update.php,bad
paypollar.com.p12.hostingprod.com/wbsc.php,bad
01453.com/,good
015fb31.netsolhost.com/bosstweed.html,good

Каждая строка содержит URL и метку ‘good’ или ‘bad’.

Необходимо:

1. Извлечь признаки. Так как данные представлены в виде текста, то необходимо привести их к векторному виду. Для этого сначала нужно использовать токенезацию текста, т.е. разбить каждый url на токены (можно почитать здесь https://habr.com/company/ods/blog/325422/#teksty, простые примеры: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/). Так как url отличаются от обычного текста, то разбивать их на токены лучше по символам “/”, ‘-’,’.’ . Токен ‘com’ будет встречаться очень часто и не является индикатором «хорошего» или «плохого» url, поэтому его можно выбросить из словаря.

   После того как все url будут представлены токенами их необходимо преобразовать в вектора. Используйте для этого TdidfVectorizer() вместо Bags of Words.
2. После того как данные будут подготовлены, разбить их на 2 выборки для обучения и тестирования, используя функцию train_test_split().
3. На учебной выборке обучить модель логистической регрессии LogisticRegression()
4. На тестовой выборке проверить производительность построенной модели, используя функцию score()
5. Предсказать класс url:

   Приведенные ниже ссылки не открывать, они могут быть вредоносными!!!!!

   wikipedia.com
   google.com/search=facebook/
   pakistanifacebookforever.com/getpassword.php/
   www.radsport-voggel.de/wp-admin/includes/log.exe

ahrenhei.without-transfer.ru/nethost.exe
www.itidea.it/centroesteticosothys/img/_notes/gum.exe