СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

КУРСОВ ПРОЕКТ ПО СИСТЕМИ, ОСНОВАНИ НА ЗНАНИЯ

Тема:

Разпознаване на авторство

Студент:

Симеон Емилов Христов, Група 3, ФН: 71845

София, януари 2021 г.

1. Формулировка на задачата

Да се реализира класификатор, който разпознава текстове на Иван Вазов от текстове на Йордан Йовков.

Класификаторът е написан на Python. Съставен е от 4 различни файла, разработени в Google Colaboratory и съответно с разширение “.ipynb”. За правилно тестване е необходимо предварително качване на текстовите материали, изброени в т. 3.

2. Използвани алгоритми

Използва се вид машинно самообучение с учител - самообучение чрез запомняне. Реализиран е методът на логистичната регресия. За характеристики на входните данни се създава таблица от вида “честота на думата - обратна честота на документа”. За оценяване работата на логистичната регресия се използват разнообразни методи, включ. матрица за разбиране на объркването.

Най-често посещаваните източници са документациите на Python, и библиотеките Pandas, Sklearn, Numpy, Matplotlib и Pickle.

3. Описание на програмната реализация

Проектът се състой от 4 файла, всеки представляващ отделна стъпка в процеса на машинното самообучение чрез запомняне:

1. 1_data_cleaning.ipynb (цел: създаване на corpus и tfidf vectorizer):
    a. извличане на текстовете, използвани за обучаващи примери.
        i. текстовете са: “Под игото” (Иван Вазов), “Епопея на забравените” (Иван Вазов), “Чифликът край границата” (Йордан Йовков), “Приключенията на Гороломов” (Йордан Йовков), “Старопланински легенди” (Йордан Йовков), “Последна радост” (Йордан Йовков), “Вечери в Антимовския храм” (Йордан Йовков).
    b. формиране на “тяло” - таблица с две колони, представляващи съответно текст и авторът му.
    c. “изчистване” на отделните текстове:
        i. заместване на главните букви със съответните им малки.
        ii. премахване на пунктуация.
        iii. премахване на думи, които не се състоят от български букви и последователности от символи, в които има такива, които не са букви.

        iv. премахване на междуметия и често използвани словосъчетания.
        v. заместване на думи с коренната им такава.
    d. създаване на биграми.
    e. създаване на таблица “честота на думата - обратна честота на документа”.
2. 2_exploratory_data_analysis.ipynb (цел: извличане на закономерности от данните):
    a. генериране на таблица с обща информация за всички колони, които съдържат числа.
    b. намиране отговор на въпросите: “Коя е “най-тежката” дума? Какво е нейното тегло? Кой автор я е изплозвал? Коя дума има най-висока средна тежест? Колко е тя? Кой автор я е изплозвал?”
    c. създаване на хистограми, показващи взаимоотношенията между различните думи.
3. 3_apply_techniques.ipynb (цел: чрез използване на логистична регресия да се създаде можел):
    a. проверка за най-подходящ модел измежду LogisticRegression, KNeighborsClassifier, GaussianNB, MultinomialNB чрез изплозване на k итеративни разделяния.
    b. чрез използване на модел, реализиращ логистична регресия да се намерят думите, отличаващи най-много двамата автори.
4. 4_testing_ground.ipynb (цел: да се провери и оцени работата на модела чрез използване на 150 случайно подбрани части от текстове):
    a. формиране на случайна извадка от всички текстове.
    b. за всеки текст от извадката да се вземе случаен интервал от символи.
    c. прилагане на модела върху всеки откъс.
    d. оценка на метода:
        i. намиране на процент коректни класификации;
        ii. генериране на матрица за разбиране на объркването;
        iii. генериране на метрики, произхождащи от матрицата за разбиране на объркването;
        iv. построяване на хистограма, показваща честотите на вероятностите за класификации на Иван Вазов;
        v. построяване на графика, показваща зависимостта между реалните правилни класификации и нереалните правилни класификации.

4. Примери, илюстриращи работата на програмната система

Началото на файл със 150 случайно генерирани входове (параграфите са разделени със символите “@@@@”):

```
@@@@
/content/drive/MyDrive/Colab Notebooks/project/data/jovkov_separated/Jordan Jovkov -  - . Krystopyt - 7926.txt
ба, случайни пътници и арестанти — цялото това множество непременно ще се отбие в етапния пункт. Там всеки ще намери подслон през нощта, ще почине, ще се огрее, ще бъде най-точно упътен, може би ще научи последните новини и, което е най-важното, ще има и хляб.
        В Добруджа посоката на пътищата не се определя от естествената проходимост на местността в една или няколко само посоки. На всяка крачка тук се пресичат и се разделят безбройни пътища, които бог знае къде водят. Еднообразието на безкрайното поле приспива и най-изострената наблюдателност, окото не намира по какво да се ориентира и не един пътник тук посред бял ден се е въртял с часове, като в магьосан кръг. А войната сякаш е увеличила още повече разстоянията, като е направила полетата безлюдни и пусти. Затова всеки етапен пункт се догонва и очаква като оазис сред пустиня. Тия, които пътуват тъдява, имат точно отмерено време, не твърде лека участ и почти всякога — някое категорично и строго задължение. А упътване, сведение за каквото и да било, подкрепа, храна — всичко това може да даде само етапният пункт.
        И нищо друго не говори тъй красноречиво и ободрително, както пътепоказателите, които са поставени по тия безбройни пътища. На високи върлини, забити в земята, стоят дъсчени табели с всякакви надписи: български, немски и турски. Някъде, непокътнат още, може да се прочете и руски надпис: „На етап“. Всички тия надписи са кратки, но са изразителни и говорят много: една стрелка, която никога няма да заблуди, после точното число на километрите, които остава да се изминат. Ония, които познават отблизо войната и са преживели не едно скитничество из непознати краища, знаят с какво упование и с какво облекчение се прочитат тия надписи. Някаква невидима ръка грижливо сочи пътя и сякаш някой с трогателно участие прошепва: „Уморени сте, измокрени и гладни. Чакате да научите нещо, което не знаете и ви измъчва. Ей по тоя път! Още малко. Близо е — и там всичко ще се уреди.“
        Пренощувах в етапното комендантство в Кубадин. На сутринта, в едно много късо време можах да видя картини, които остават незабравими. Отчетливо и ярко войната хвърляше в тях и блясъка, и сенките си, показваше сложния механизъм на исполинските си усилия и тия дълбоки промени, които никоя друга стихия не е в състояние, да направи.
        Цялото пъстро множество, което беше пренощувало тук, се събуди и се готвеше да продължи пътя ся. Минава унгарски хонвед. Той езди хубав червен кон и върху рамената му с отпуснати ръкави е наметната синя венгерка, обшита с черни кожи. Още сънният татарин, като надднича от вратата на хижата си, учуден следи тоя конник, който му се струва като някое фанта-стично видение. Никога няма да разбере той отде и как е дошел по тия места. Пред комендантството се нареждат пъстроцветни групи. Всеки, преди да тръгне, ще получи хляба си. Най-напред стоят група наши войници, които отиват към фронта. Струват ти се извънредно познати и хиляди пъти виждани тия флегматични някак лица, малко тежки движения, под които веднага проглежда тая хладна разсъдливост, упорита и спокойна самоувереност, която навсякъде ги отличава. Близо до тях се нареждат и десетина турски войници. Те вече са получили точно упътване де ще намерят корпуса си. Тяхната редица е безукоризнено правилна и всеки един високо и старателно произнася името си, преди да получи хляба си. Изплашени и смутени, тихо се изправят и трима турци, селяни от Добричко. Заловени са без документи. Никой, разбира се, не вярва, че са тръгнали по работа, както уверяват. На лицата им стои огорчението, че наивно са повярвали ония които преди тях са дохождали за плячка.
        Малко настрана е построена голяма група руски пленници. Това са повече млади хора, уморени и бледи. Едни са с фуражки, други — с високи бели калпаци. Пленени са в последните сражения, но между тях има и такива, които още недоумяват и не могат да разберат как се е случило да попаднат в неприятелски ръце. Ето най-възрастния между тях, слаб и прегърбен, изгубен в шинела си, извънредно голям за него, препасан несръчно почти през средата на гърдите. Познава се, че е нестроевак. Казва се Гусев Михаил Андреич, от Курска губерния. Печално и развълнувано той разказва историята на пленяването си. Бил ординарец при ротния си командир. Надвечер тръгнал да занесе вечеря на негово високоблагородие. Било тъмно, мъгла, ситен дъжд шибал по лицето му и премрежвал очите му. Гусев внимавал преди всичко да не разлее гостбата. Ето вижда светлина и сред нея — хора, разбира се — свои. Но той и сега не може да разбере къде е отишел, че наместо на ротния си командир подал вечерята на един български офицер. Лицето му е тъй измъчено, печално и виновно, като че не толкоз му тежи самото пленяване, колкото тая непростима грешка — да остави без вечеря негово високоблагородие.
        Пълна противоположност на злочестия Гусев е другаря му Остап Петренко — висок, здрав казак, кавалерист от някакъв си пограничен полк. Той гледа открито и спокойно, говори правилно и разумно. И той разказва как са го пленили, но наместо простодушната скръб на Гусева, в разказа му не липсва високомерна ирония към сляпата случайност и собствената си небрежност. Слязъл от коня да пристегне подпругите на седлото, а нашите тъй бързо и неочаквано го заобиколили, че не останало време за нищо. „Попалсь“ — заключава той конфузно и пресилено се усмихва. Това е хитра лисица, неочаквано попаднала в капана.



@@@@
/content/drive/MyDrive/Colab Notebooks/project/data/jovkov_separated/Jordan Jovkov -  - . Te pobediha - 7909.txt
а сърцето си и шепне:
        „Сполай ти, боже. Добър плод ни си дал…“
        И когато една вечер камбаните в глухото малко село тревожно разнасят зловещата вест за войната, той със същата загриженост, спокойствие и твърдост се готви за новия труд. Труд също тъй тежък, като една сеитба, като една жътва. И ето те двамата напущат родния кът, тъгуват за милувките и смеха на малките палави любимци, скитат се далече по чужди пътища, под чужди небосклони. Но навсякъде те показват своите беззаветни усилия, изпитана душа, железни мишци и сила. Победи гърмят една след друга. Светът се учудва, светът очарован гледа: и далече, всред посърналите пустини на Чаталджа, под дъждовното небе, в кални пътища, той съзира тях — двамата другари. Да, те победиха!
        Еднаква добросъвестност и усърдие и във войната, и в мирния труд. Аз си спомням как тоя честен, добър селянин, тъгуващ сред боевете за своите ниви, често вземаше шепа пръст, размазваше я между пръстите си и казваше: „Да, харна земя. Захар!“… И може би той, широкоплещияг работник, влюбен в земята и труда, изправен на брега на Мраморното или Бялото море, също като мужика на Горки, загледан в безкрайната ширина, с въздишка е казвал: „Ех, да беше поле това… че да го разореш!…“
        Сюблимен герой в боя и в мирния труд! И сега, в тия тежки, безрадостни дни в едно тъмно упование аз го виждам да се изправя пред мене, да расте като мощен, гранитен исполин, здрав, силен, с девствена чиста дута, с една вяра и една обич: земята и труда. А ниско при неговите нозе, малки, дребни хора горещо се препират за него, викат, махат ръце, вещаят неговите съдбини, чертаят пътя на щастието му. Но никой достатъчно не го разбира и никой достатъчно не го обича. А той — сега са зимни дни и се попийва — накрехва нехайно рунтав калпак, очите му дяволито се смеят, подмигва лукаво и казва: „Дявол да го вземе! Ще пусна една червена бюлетина, да видим какво ще излезе!“…

        * * *
```

Изходи (True - коректно класифициран, 0 - некоректно класифициран):

```
array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
        True, False,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True, False,
        True,  True,  True, False,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True, False,  True,
        True,  True,  True,  True, False,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True, False,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True])
```

5. Литература

- https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/ : дата на първо посещение (26.12.2020г.)

- Python 3 documentation: https://docs.python.org/3/tutorial/index.html дата на първо посещение (26.12.2020г.)
- Sklearn documentation: https://sklearn.org/documentation.html дата на първо посещение (26.12.2020г.)
- Pandas documentation: https://pandas.pydata.org/pandas-docs/stable/reference/index.html дата на първо посещение (26.12.2020г.)
- Numpy documentation: https://numpy.org/doc/1.20/reference/index.html дата на първо посещение (26.12.2020г.)
- Matplotlib documentation: https://matplotlib.org/contents.html дата на първо посещение (26.12.2020г.)
- Pickle documentation: http://docs.picklesdoc.com/en/latest/ дата на първо посещение (26.12.2020г.)
- Много полезни Youtube канали:
    - Ken Jee дата на първо посещение (26.12.2020г.);
    - Data Professor дата на първо посещение (26.12.2020г.).