

UNIVERSITATEA "ALEXANDRU IOAN CUZA" DIN IAȘI
FACULTATEA DE INFORMATICĂ



LUCRARE DE LICENȚĂ

**Clasificarea, clusterizarea si cartografierea datelor de pe Twitter
folosind Apache Spark si Folium**

propusă de

Rareș Bradea

Sesiunea: *Iulie, 2018*

Coordonator științific

Lect.dr. Cristian Frăsinaru

UNIVERSITATEA "ALEXANDRU IOAN CUZA" DIN IAȘI
FACULTATEA DE INFORMATICĂ

Clasificarea, clusterizarea si cartografierea datelor de pe Twitter folosind Apache Spark

Rareș Bradea

Sesiunea: *Iulie, 2018*

Coordonator științific
Lect.dr. Cristian Frăsinaru

DECLARAȚIE PRIVIND ORIGINALITATE ȘI RESPECTAREA DREPTURILOR DE AUTOR

Prin prezenta declar că Lucrarea de licență cu titlul „*Clasificarea, clusterizarea și cartografierea datelor de pe Twitter folosind Apache Spark si Folium*” este scrisă de mine și nu a mai fost prezentată niciodată la o altă facultate sau instituție de învățământ superior din țară sau din străinătate. De asemenea, declar că toate sursele utilizate, inclusiv cele preluate de pe Internet, sunt indicate în lucrare, cu respectarea regulilor de evitare a plagiatului:

- toate fragmentele de text reproduse exact, chiar și în traducere proprie din altă limbă, sunt scrise între ghilimele și dețin referința precisă a sursei;
- reformularea în cuvinte proprii a textelor scrise de către alți autori deține referința precisă;
- codul sursă, imaginile etc. preluate din proiecte *open-source* sau alte surse sunt utilizate cu respectarea drepturilor de autor și dețin referințe precise;
- rezumarea ideilor altor autori precizează referința precisă la textul original.

lași, *data*

Absolvent *Rareș Bradea*

(semnătura în original)

Cuprins

1 Colectarea datelor in timp real	12
2 Clasificare.....	16
2.1 Descrierea problemei.....	16
2.2 Incercari ulterioare.....	17
2.3 Solutia.....	17
3 Clusterizare.....	22

Introducere

Motivatie si gradul de noutate

In ultimul timp, retelele sociale online s-au dezvoltat enorm, plecand de la firavele inceputuri lipsite de multe functionalitati si importante si ajungand in ziua de azi sa reprezinte structuri sociale si chiar politico-economice de uriase proportii. O retea sociala online poate fi definita in contextul recentului fenomen numit Web 2.0 (care se refera la noua generatie de website-uri unde accentul se pune pe usurinta folosirii si crearea de continut de catre utilizatori) [4] ca fiind un site web unde utilizatorii fac parte dintr-o structura sociala si reprezinta „actori” sociali ce sunt conectati prin mai multe legaturi de tip „one-to-one”, creand astfel graful retelei sociale. Printre cele mai populare retele sociale online se numara Facebook, Instagram si Twitter, avand 2234, 813 si respectiv 330 de milioane de utilizatori, potrivit unui recensamant din Aprilie 2018 [5].

Prin popularitatea lor, aceste retele sociale devin foarte usor tinta utilizatorilor malitiosi ce au ca scop propagarea interactiunilor si fenomenelor de tip spam si phishing. Acestea au ca obiectiv convingerea utilizatorilor firesti, prin cai daunatoare, deranjante si de obicei greu de descoperit, sa isi expuna datele personale sau sa piarda bani in favoarea celor care recurg la aceste tactici.

Pe langa aceste practici malitioase, pe Twitter exista si foarte multe postari irelevante pentru multi din utilizatori. O data postat un tweet, acesta ajunge pe feed-ul oricarui utilizator ce urmareste persoana ce posteaza. Acest lucru face ca uneori sa existe o neconcordanza intre doleanțele unui utilizator si ce citește acesta pe propriul feed.

Obiectivele lucrării

Unul dintre cele doua principale obiective ale lucrării este acela de a studia in profunzime frameworkul de cluster computing numit 'Apache Spark'. Dintre componentele sale, cele mai relevante pentru implementarea unei aplicatii folosind frameworkul acesta au fost studiate mai in profunzime si descrise in capitolele relevante fiecarui modul din aplicatie. Spark Streaming este descris in primul capitol si Spark MLlib in capitolele 2 si 3.

Al doilea obiectiv este implementarea unei aplicatii ce foloseste tehnologii existente in frameworkul Apache Spark. Acest software are scopul de a filtra multe din aceste mesaje folosind tehnici de clasificare din invatare automata, explicate in al doilea capitol. De asemenea, se dorește crearea unor grupuri de tweeturi ce abordeaza subiecte asemanatoare si sunt de inalta calitate, pentru a veni in ajutorul persoanelor ce sunt interesate de a 'lua pulsul' societatii in care se afla, precum jurnalistii, fara sa fie nevoiti sa sorteze printr-o multime de mesaje ce se pot dovedi a fi irelevante,

astfel avand parte de o experienta sigura si utila.

Aceste grupuri de tweeturi vor fi afisate pe o harta si vor putea fi cautati termeni ce se doresc a fi gasiti in grupuri special create dupa acei termeni, astfel creandu-se anumite topicuri ce sunt relevante pentru inputul unui utilizator. Cartografierea acestor tweeturi este utila deoarece creeaza o perspectiva noua si ofera o imagine de ansamblu ce poate descrie foarte usor de unde a pornit un anumit fenomen social, fie el o stare, un zvon, o idee pentru o miscare sociala sau chiar un dezastru natural sau uman.

Metodologia folosita

Pentru a atinge aceste obiective, adica de a clasifica binar un tweet, fie intr-o categorie de continut de calitate inalta, fie o categorie de continut de calitate scazuta, si de a grupa (in termeni de invatare automata, de a clusteriza) o colectie de tweeturi in mai multe subcolectii ce sunt asemanatoare intre ele, iar apoi cartografia acestor tweeturi intr-un format usor de interpretat am apelat la biblioteca 'folium' din Python si frameworkul de cluster computing 'Apache Spark'. [Anexa1]

Lansat in anul 2014, avand ca autor initial pe Matei Zaharia, in cadrul proiectului AMPLab al universitatii Berkley, din California, acesta are la baza o abstractizare a datelor numit resilient distributed dataset (RDD). Si din denumire se poate infera ca acesta descrie un multiset (un set ce poate contine mai multe instante ale aceluiasi element) de date distribuite pe un cluster de computere.

In cadrul acestui framework exista mai multe componente ce ofera diverse functionalitati:

- pentru a utiliza comenzi SQL peste o abstractizare a datelor numita DataFrame, ne este pus la dispozitie Spark SQL

- componenta Spark Streaming se ocupa cu analiza stream-urilor de date. Datele sunt aduse in memorie in mini-batch-uri (grupuri mai mici de date) si se pot efectua transformari RDD pe acestea

- componenta Spark MLlib ofera implementari ale unor algoritmi de invatare automata, capabili de calcul computational distribuit

- componenta GraphX este un framework de procesare a grafurilor

Din toate acestea, cele utilizate in aplicatie sunt Spark SQL, Streaming si MLlib, pentru citirea datelor de pe Twitter, aplicarea unor algoritmi de invatare automata peste ele si pentru manipularea de DataFrames.

Mai in profunzime, Apache Spark este un framework open source destinat procesarii unor volume de date la scara mare. Vizeaza aplicatiile construite pe sisteme distribuite si APIul expune functionalitati pentru limbajele Java, Scala, Python si R. Folosind Spark Application Frameworks, Spark, scris in Scala, simplifică accesul la algoritmi de machine learning și analiză predictivă. Spark Core, componenta de bază a frameworkului, se bazează pe o abstractizare a datelor numita “resilient

distributed dataset” (RDD), ce reprezintă o mulțime, o colecție imutabilă de elemente distribuita pe un cluster de sisteme computaționale peste care se poate opera în paralel. Caracteristicile acestui tip de date sunt, după cum sugerează numele:

- Rezilient, există posibilitatea de a recomputa partiții cu probleme
- Distribuit, datele se află pe mai multe noduri într-un cluster
- Este un dataset cu valori primite sau valori de valori (tuple sau obiecte)

Spark, rulat în mod nelocal, are nevoie de un manager de clustere și un sistem distribuit de stocare a datelor. Pe lângă soluția nativă a managerului, există suport pentru Hadoop Yarn și Apache Mesos. Pentru stocarea datelor se poate utiliza Hadoop Distributed File System (HDFS), MapR File System (MapR-FS), Cassandra, OpenStack Swift, Amazon S3, Kudu,

Spark mai poate fi descris prin capacitățile din modulul Spark Streaming. Acesta permite utilizatorului să lucreze cu cantități mari de date ce sunt servite în timp real, prin deschiderea unui stream și “ascultând” date precum statusuri de Twitter sau, de pildă, streamuri custom construite în Kafka, Flume, Kinesis sau chiar socketuri TCP. Aceste date pot fi procesate utilizând algoritmi complecși de machine learning sau procesarea grafurilor cu funcții high-level precum map, reduce, join și window. Datele procesate pot fi exportate în fișiere, baze de date sau live dashboards. Spark Streaming lucrează astfel, primește un input live de data streams și le împarte în elemente numite batches, sunt apoi procesate și astfel rezultă streamul final de batches.



Spark Streaming oferă o abstractizare numită discretized stream (DStream), ce reprezintă un stream continuu de date. Acestea pot fi create din surse precum Kafka, dar și prin aplicarea unor operațiuni high-level pe alte DStreams. Intern, un DStream este reprezentat de o serie continuă de RDDs. Fiecare RDD dintr-un DStream conține date dintr-un anumit interval de timp.



Orice operațiune aplicată pe un DStream este de fapt tradusă în operațiuni pe RDDs din spate. Aceste operațiuni sunt efectuate de Spark engine. Operațiunile pe DStream ascund multe din detalii și oferă un API high-level, pentru facilitarea utilizării.

Fiecărui DStream de input (în afara celor care fac streaming din fișiere) îi este asociat un Receiver, o componentă, un obiect ce primește datele de la o sursă și o stochează în memorie pentru procesare. Există două tipuri de receivers, în funcție de fiabilitatea acestora. Surse fiabile precum Kafka sau flume permit datelor transferate să fie recunoscute. Dacă sistemul ce primește datele de la aceste surse fiabile recunoaște corect datele primite, atunci există siguranța că nu vor exista pierderi de informație ca urmare oricăror tipuri de defecțiuni. Astfel, receivers pot fi de două tipuri:

- Reliable receiver; acesta trimite confirmarea către o sursă fiabilă când datele au fost primite și stocate.
- Unreliable receiver; acesta nu trimite niciun fel de confirmare către sursă. Se folosesc pentru surse care nu suportă acest sistem de acknowledgment (admitere și confirmare).

Procesarea DStreamurilor și, deci, a RDDurilor se face prin transformări. Acestea sunt niște funcții care modifică datele. De exemplu, funcția `map(myFunc)` returnează un nou DStream prin aplicarea funcției `myFunc` peste toate elementele din DStreamul inițial. Funcția `transform(myFunc)` permite apelarea unor funcții RDD-to-RDD, care nu sunt aplicabile direct pe DStreams, pe fiecare RDD dintr-un DStream. Un exemplu de astfel de funcție este joinul dintre batchurile dintr-un stream și alt dataset.

Există transformări ale căror apeluri sunt constrânse de timp. Prin aceste windowed



transformations, se pot aplica modificări pe datele peste care trece o “fereastră glisantă”, ca în exemplul de jos.

Pe DStreams pot fi utilizate, de asemenea, DataFrames și operațiuni SQL. Fiecare RDD este convertit într-un DataFrame, înregistrat ca un tabel temporar peste care se pot face interogări SQL.

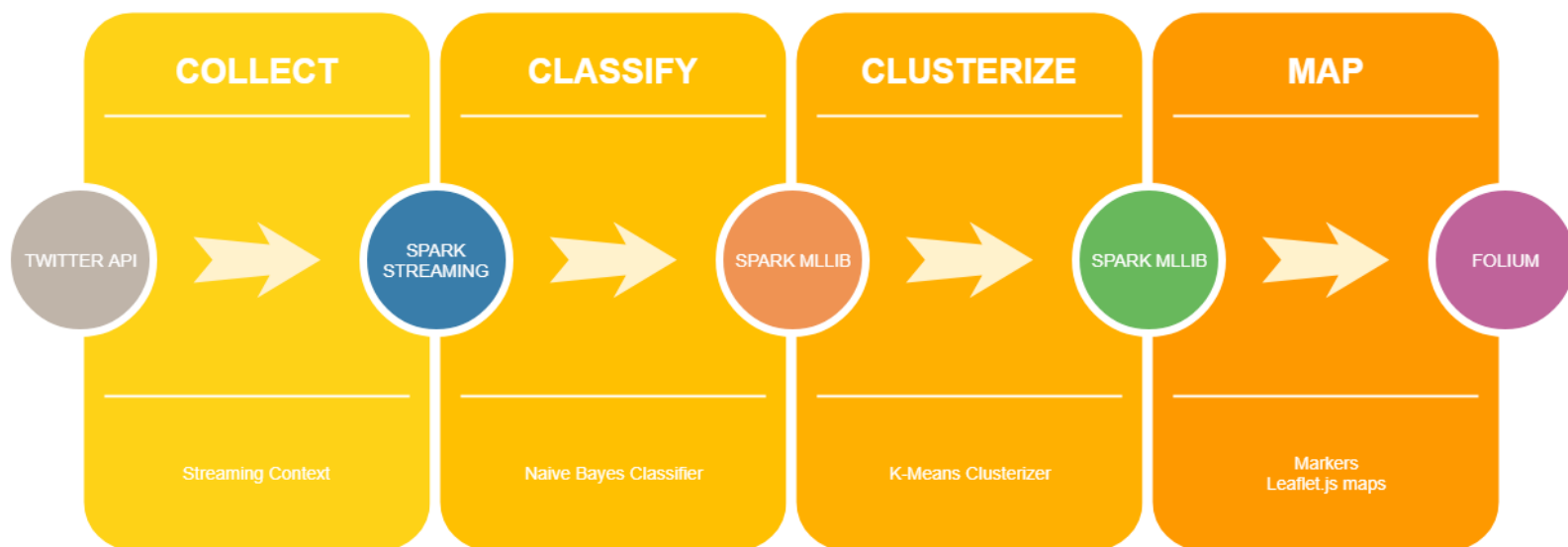
Algoritmii de streaming machine learning din MLib pot învăța din streamurile de date și, în același timp, să aplice aceste cunoștințe pe același stream de date. Menționez algoritmi capabili de aceste lucruri: Streaming Linear Regression, Streaming KMeans etc. Pentru alți algoritmi, se pot folosi date istorice pentru learning, mai apoi aplicându-se modelul pe streamuri de date.

În concluzie, Apache Spark este o tehnologie foarte puternică și foarte rapidă, ce permite cluster computing pe dataseturi foarte mari, utilizând machine learning, graph processing, streamuri de date și

alte metode.

Descrierea sumara a solutiei

Aplicatia se poate imparti in patru module ce lucreaza impreuna pentru a ajunge la rezultatul final, reprezentat de o colectie de tweeturi, impreuna cu locatia lor, ce au fost clasificate ca fiind de calitate inalta, clusterizate in grupuri relevante, dupa asemanarea dintre ele si apoi afisate pe o harta ca puncte ce contin textul tweetului, locatia acestuia si grupul semantic carui apartine.



Primul modul este compus din doua scripturi Python. Primul din ele apeleaza API-ul Twitter pentru a primi acces la un stream live de tweeturi. Acesta transmite prin TCP/IP tweeturi catre o instanta de Spark Streaming ce se afla pe un al doilea script. Acesta din urma preia tweeturi si le clasifica ca fiind de slaba sau inalta calitate, folosind un model de clasificare bazat pe algoritmul Bayes naiv.

Al doilea modul consta in scriptul Python ce a antrenat si testat diversi algoritmi de clasificare, implementati in Spark Mllib. Concluzia optima dupa multiple incercari a fost utilizarea algoritmul Bayes naive pe un dataset cu aproximativ 1200 de instante de tweeturi de slaba calitate si aproximativ 10000 de instante de tweeturi de inalta calitate. Acesta ofera o acuratete de 93.5% la testare, una destul de apropiata de celalte variante, cuprinse intre 91% si 92%.

Al treilea modul este alcatuit din functii ce pot clusteriza datele salvate in primele module si pot

produce rezultate ce constau în asocierea fiecărui tweet cu un cluster.

Al patrulea modul se referă la randarea tweeturilor clusterizate în modulul precedent, afisându-le pe harta globului pământesc și oferind o imagine de ansamblu asupra modului de propagare și naștere a unor subiecte de interes major pe rețeaua de socializare Twitter.

1 Colectarea datelor in timp real

Orice aplicație ce include în implementarea ei și un modul de învățare automată are imperioasă nevoie sau cel puțin beneficiază foarte mult de pe urma unui dataset cât mai extins, dar și curat. Twitter, prin popularitatea sa imensă atinge cu excelență punctul referitor la cantitatea datelor, dar se îndepărtează de un ideal al datelor relevante și curate. Dat fiind faptul că oricine poate să își facă un cont unde poate să exprime idei în limita a 280 de caractere, nu este o surpriză faptul că relevanța multor tweeturi este minimă pentru multe persoane. Partea de clasificare a aplicației se va ocupa de etichetarea acelor tweeturi ce conțin enunțuri fără sens, cuvinte deosebit de vulgare, încercări de comercializare sau spam. Înainte de a ajunge acolo, trebuie să clădim un dataset cât mai mare pentru a micșora impactul postărilor cu relevanță scăzută.

Acestea fiind spuse, Twitter este un ecosistem foarte complex și activ, iar pentru a facilita munca dezvoltatorilor de software interesați de datele ce rezidă în interiorul aplicației lor, s-a creat un API ce expune accesul la tweeturi în timp real. Aplicația mea urmărește să clasifice, clusterizeze și apoi să afișeze pe hartă tweeturi care au o vechime scurtă, astfel că este foarte utilă utilizarea API-ului de streaming oferit. Alături de acesta, voi folosi și o instanță de Apache Spark cu un context de Streaming ce permite procesarea datelor stream-uite live.

O altă variantă ar fi fost folosirea unor dataseturi deja existente, însă relevanța lor referitoare la vârstă și noutate ar fi fost discutabilă. De altfel, multe dataseturi urmăresc un singur subiect, cum ar fi tweeturi ce discută situația imigrării în Canada. Acest lucru se dovedește a fi util pentru partea de testare a clusterizării, dar scopul aplicației este să fie semi-realtime și să dispună de o varietate a subiectelor vastă și necontrolată a priori.

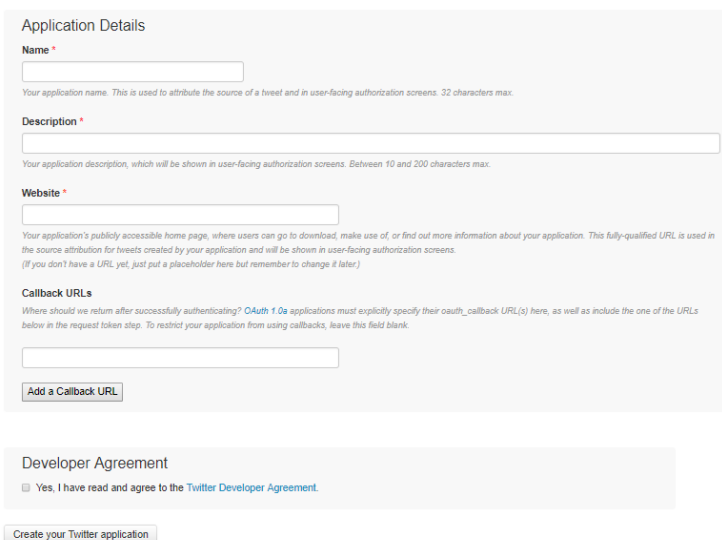
În cele ce urmează, voi descrie succint pașii parcurși de mine pentru a ajunge la un modul ce accesează API-ul de tweet streaming și trimite date către o instanță stream-ready de Spark ce le procesează.

În primul rând, pentru a avea acces la API-ul Twitter este nevoie un cont simplu de Twitter și mai apoi de înregistrarea unei aplicații ce dorește accesul la API, accesând pagina <https://apps.twitter.com/>.

Dupa completarea acestui formular, vom avea acces la un dashboard cu anumite informatii. De acolo vom prelua patru coduri importante si necesare autorizarii noastre la serviciul oferit.

Acestea sunt un access token, un access token secret, consumer key si consumer key secret. Le voi folosi pentru a face autorizarea printr-un request OAuth1 folosind biblioteca 'requests_oauthlib' din Python, astfel.

Create an application



Application Details

Name *

Description *

Website *

Callback URLs

Developer Agreement

☐ Yes, I have read and agree to the [Twitter Developer Agreement](#).

Create your Twitter application

```
ACCESS_TOKEN = 'sample'
ACCESS_TOKEN_SECRET = 'sample2'
CONSUMER_KEY = 'sample3'
CONSUMER_SECRET = 'sample4'
auth_worker = requests_oauthlib.OAuth1(
    CONSUMER_KEY,
    CONSUMER_SECRET,
    ACCESS_TOKEN,
    ACCESS_TOKEN_SECRET)
```

Mai departe, pornind un server care asteapta cereri pe localhost la un port oarecare (aici 9999), voi face accesul unui stream de date de la Twitter catre o instanta de Apache Spark.

```
def stream_tweets():
    query_url = 'https://stream.twitter.com/1.1/statuses/filter.json?language=en&locations=-136,15,-45,55&track=#'
    response = requests.get(query_url, auth=auth_worker, stream=True)
    print(query_url, response)
    return response
```

Funcția 'stream_tweets' returnează un răspuns http ce conține, pentru fiecare linie, un tweet complet respectând parametrii dați în query. Mai exact, cautăm tweeturi scrise în limba engleză, iar parametrul locations astfel setat ne garantează că locația mesajelor provine din aproximativ partea continentală a Statelor Unite ale Americii.

Am ales această locație deoarece Twitter este foarte popular și utilizat în respectiva țară, fapt ce

denota o posibila diversitate mai mare decat in alte locuri. Subiectele discutate sunt extrem de variate si ne pot oferi multe sanse de a obtine date interesante si din prisma geolocatiei, un atribut ce il are orice tweet al carui utilizator permite accesul serviciilor de locatie asupra contului.

```
def send_tweets_to_spark_with_location(http_response, connection):
    for line in http_response.iter_lines():
        try:
            full_tweet = json.loads(line)
            my_dict = dict([('text', full_tweet['text']), ('coordinates', full_tweet['coordinates'])])
            if my_dict['coordinates'] is not None:
                str_my_dict = json.dumps(my_dict)
                connection.send(bytearray(str(str_my_dict) + '\n', 'utf8'))
        except Exception as e:
            print(e)
```

Funcția 'send_tweets_to_spark_with_location' primește un răspuns http (în cazul acesta, cel returnat de funcția 'stream_tweets'), o conexiune și trimite prin aceasta un vector de bytes ce reprezintă o parte dintr-un tweet. Domeniul aplicației ne permite să folosim doar o parte din nenumăratele atribute ale unui tweet. Mai exact, avem nevoie doar de text și de coordonatele geografice reprezentant punctul de unde a fost trimis acel tweet.

Aceste funcții fac parte dintr-un script Python care la execuție face bind și listen unui socket pe localhost, port 9999. Când primește o cerere, acceptă și trimite prin conexiune, folosind funcțiile descrise mai sus, tweeturi către entitatea care face cerere.

A doua parte a procesului se referă la această entitate. Ea este un `DataStream` creat de un `StreamingContext` din biblioteca `pyspark.streaming`. `DataStream`ul este descris de un `socketTextStream` ce face cereri la serverul numit mai sus. Pe stream aplicăm o funcție pe datele live pentru a face split

```
ssc = StreamingContext(sc, 2)
dataStream = ssc.socketTextStream("localhost", 9999)
tweets = dataStream.flatMap(lambda line: line.split("\n"))
tweets.foreachRDD(print_with_location_rdd_with_prediction)
ssc.start()
ssc.awaitTermination()
```

între tweeturi. Apoi pentru fiecare tweet afișăm și salvăm într-un fișier, pentru fiecare tweet, textul, locația precum și predicția făcută de clasificatorul de spam.

Am folosit partea de Streaming din Apache spark pentru a avea acces la utilizarea în timp real a unor funcții peste aceste date. Aici folosesc funcția 'foreachRDD' care, după cum reiese și din denumire, apelează o anumită funcție pe fiecare RDD din datastream.

Funcția 'print_with_location_rdd_with_prediction' afișează și salvează tweeturile astfel: folosind pe întregul tweet funcția 'loads' din biblioteca json, putem încărca un dicționar dintr-un string.

Pentru a putea clasifica textul unui tweet ca fiind ori spam, ori non-spam, avem nevoie sa impartim stringul in mai multe substring-uri.

Acest proces se numeste tokenizare, iar pentru acest task se poate folosi foarte utila biblioteca creata pentru exact acest scop. Obiectul de tip TweetTokenizer din biblioteca nltk.tokenize imparte un string in tokenuri avand in vedere si structura unui tweet. Parametrii 'strip_handles' si 'reduce_len' folositi in constructorul acestei clase indeamna tokenizerul sa reduca dimensiunea tweetului daca este posibil. Acest lucru are loc daca se repeta foarte multe litere intr-un cuvant. Acest lucru este destul de des folosit, deoarece utilizatorii ar putea accentua anumite cuvinte prin repetarea unor litere.

De asemenea, referintele la alti utilizatori nu sunt incluse. Spre exemplu, textul '@remy: This is waaaaayyyy too much for you!!!!!!' este tokenizat si returnat ca intr-o lista astfel: [':', 'This', 'is', 'waaayyy', 'too', 'much', 'for', 'you', '!', '!', '!'].
[1]

Mai departe, predictia nu poate fi facuta pe niste simple cuvinte, astfel ca apelam la un procedeu numit feature hashing. Pentru fiecare term, insemnand cuvant, se calculeaza hashul acestuia si numarul de aparitii a cuvantului in text si aceste rezultate sunt pastrate intr-un SparseVector. Modelul de clasificare are nevoie de acest vector rar pentru a face predictia. Acest procedeu este detaliat in capitolul referitor la Clasificare.

Cat timp serverul detaliat in prima parte ramane pornit, acest al doilea script primeste tweeturi, le clasifica si le scrie intr-un fisier pentru a fi apoi folosit de celelalte scripturi.

In final, vom avea date care arata in felul urmator si pe care se poate lucra foarte usor spre a fi clusterizate dupa subiect si cartografiate estetic, scotand in evidenta tweeturile care au in componenta un anumit termen dat ca input.

	A	B	C
1	text	label	location
2	Heavy traffic in #Hillsborough on I-4 WB from Branch Forbes Rd to McIntosh Rd, incident	0	[-82.187, 28.02682]
3			
4	We're #hiring! Read about our latest #job opening here: Environmental Protection Assist	0	[-85.7584557, 38.2526647]
5			
6	Can you recommend anyone for this #job? Police Officer - https://t.co/RchjmUbfHV #se	0	[-98.3420118, 40.9263957]

2 Clasificare

2.1 Descrierea problemei

În statistica și învățare automată, clasificarea este problema identificării unei categorii din care face parte o nouă observație, pe baza unui set de date de antrenare, unde se cunoaște categoria pentru fiecare observație. În învățare automată, aceste observații se numesc instanțe, clasificarea fiind un procedeu de învățare supervizată. Un algoritm de învățare supervizată (supervised learning) analizează datele din setul de antrenare și inferează o funcție care poate fi folosită pentru maparea noilor instanțe. Un algoritm care implementează acest lucru se numește clasificator și deseori în statistica se folosește regresia logistică.

Această parte a aplicației are menirea de a clasifica binar un tweet, fiind categorisit fie ca fiind spam sau non-spam. O foarte mare importanță pentru a crea cu succes un clasificator de spam o are utilizarea unui dataset relevant și destul de extins. Cealaltă necesitate este aceea de a alege metoda cea mai potrivită pentru datele obținute. Câteva din metodele cele mai populare de clasificare binară sunt:

- arbori de decizie
- random forests
- support vector machines
- regresie logistică
- rețele bayesiene

Din cele enumerate mai sus, Apache Spark are implementări pentru toate, dar o parte din metode nu sunt potrivite pentru arhitectura taskului necesar de îndeplinit. Implementarea algoritmilor de arbori de decizie și random forests au nevoie de prea multă memorie, deoarece spațiul problemei este foarte extins. Numărul de features de care se folosesc acești algoritmi este de 2^{18} . Acest număr survine din necesitatea de a nu avea coliziuni când se face Hashing pe termi. Vocabularul limbii engleze fiind foarte bogat, acest lucru trebuie reflectat și în numărul maxim de features folosit în Hashing, deoarece fiecare cuvânt să aibă un hash unic și, deci, clasificarea și calculul erorii la clasificare să fie unul relevant.

Problema alegerii unui algoritm potrivit este destul de trivială, deoarece aceștia sunt foarte ușor de folosit, implementările fiind disponibile în framework. Multe din aceste clasificatoare se utilizează după un tipar asemănător. Se instantiază un obiect de tipul algoritmului, se antrenează acest algoritm pe datele de antrenare și apoi se testează pe datele de testare. Comparând acuratetea rezultată pentru fiecare algoritm, îl alegem pe cel cu acuratetea cea mai bună.

Problema gasirii unui dataset relevant este ceva mai dificila, deoarece avem nevoie de date care sa fie deja etichetate corect. Acest lucru poate fi facut manual, desigur, dar pentru foarte multe instante acest lucru devine impracticabil.

2.2 *Incercari ulterioare*

Initial, doream sa rezolv problema propagandei si miscarilor cu tente de instigare la instabilitate in Statele Unite ale Americii creata de utilizatori cu conturi false provenind din Rusia. Acest lucru parea facil la inceput, deoarece exista un dataset publicat de NBCNews [1] foarte interesant cu 200.000 de tweeturi din 2016 aparinand unor useri malitiosi ce doreau sa creeze instabilitate politica si sociala in randul cetatenilor, pentru a descreste popularitatea capitalismului si eventual pentru a impinge balanta sanselor castigarii alegerilor prezidentiale in favoarea unui participant sau altul.

Avand atat de multe date cu tweeturi fake, aveam nevoie sa gasesc un dataset cu tweeturi ale unor useri de buna credinta si cu continut relevant si curat. Acest lucru s-a dovedit a fi fiind destul de dificil deoarece majoritatea dataseturilor urmaresc ori un subiect anume, ori tweeturi cu continut negativ, astfel ca dataseturile cu tweeturi ce discuta subiecte aleatoare in mod non-spam sunt putine. De asemenea, folosind un dataset non-spam cu subiecte non-politice ar fi dus la o falsă foarte buna acuratete la testare, deoarece, antrenand algoritmul pe doua dataseturi cu topicuri diferite, adica unul cu materiale politice spam si unul cu materiale non-politice non-spam se ajunge la o clasificare a subiectului tweetului si nu neaparat a apartenentei la o categorie spam sau non-spam. Acest lucru s-a si intamplat de altfel cu un dataset de genul acesta. In incercarea de a folosi un alt dataset de tweeturi non-spam, dar politice, am ajuns la concluzia ca in cazurile reale, clasificatorul dadea dovada de un comportament de underfitting, clasificand toate tweeturile non-politice ca fiind non-spam, iar cele cateva tweeturi politice intalnite ca fiind aleatoriu spam sau non-spam. Acesta lucru se datoreaza naturii datasetului propus de NBCNews, tweeturile continute in acesta fiind aproape imperceptibil de asemanatoare cu tweeturile politice si non-spam obtinute pe parcurs.

Astfel, a trebuit sa renunt la incercarea de a rezolva problema clasificarii tweeturilor de propaganda sau instigare la instabilitate politica deoarece textul din acele tweeturi nu ofera destule informatii relevante. In acest impas se afla si mari organizatii, guvernamentale sau nu, si deci ramane o preocupare deschisa pentru viitor.

2.3 *Solutia*

Reluand analiza imaginii de ansamblu, am ajuns la concluzia ca solutia ideala este folosirea unui dataset cu tweeturi cu subiecte aleatoare, etichetate cu spam sau non-spam. Twitter este o platforma in care oricine poate avea o voce referitoare la orice, acest lucru ducand la o impresionanta diversitate a subiectelor abordate. Pentru domeniul de lucru al acestei aplicatii, care este gasirea unor subiecte bine definite in aceasta mare de tweeturi aleatoare, clasificatorul nostru trebuie sa ne permita se renuntam la acele tweeturi care nu ar avea nicio relevanta pentru niciun subiect. Ne referim aici la tweeturi fara sens, cu caractere iligibile, enunturi incorrigibile, vulgaritate fara menire, reclame si vanzari de factura malitioasa, vouchere, phishing, spam.

Un studiu facut pe acest domeniu, de analiza a detectiei tweeturilor cu continut de slaba calitate [2] pune la dispozitie un dataset cu 100.000 de instante etichetate, tweeturi de continut aleatoriu, fie de slaba sau inalta calitate. Acest fisier de tip CSV contine doar ID-ul tweetului si eticheta acestuia, incat, in mod oficial, dataseturile mari de tweeturi nu pot fi distribuite in mod public, cu textul si celelalte attribute in plaintext.

Pentru a extrage tweetul folosind API-ul Twitter, avand la dispozitie ID-ul tweetului, avem nevoie de o functie care utilizeaza key-urile descrise in primul capitol.

```
def get_tweet_from_id(id):  
    url = "https://api.twitter.com/1.1/statuses/show.json?id=" + str(id)  
    response = requests.get(url, auth=auth_worker, stream=True)  
    for line in response.iter_lines():  
        my_full_tweet = json.loads(line)  
    return my_full_tweet
```

Mai departe, se parseaza CSV-ul oferit in articolul [2] si se apeleaza aceasta functie pentru fiecare ID de acolo. Pentru fiecare tweet care este inca valabil, adica a caror continut returnat nu incepe cu 'error', se apeleaza o functie care adauga un JSON intr-un CSV, pentru usurinta folosirii ulterioare.

Valabilitatea tweeturilor depinde de sansa; articolul de unde provine datasetul fiind publicat in 2017, exista posibilitatea ca unele din acestea sa fi fost sterse de pe Twitter. Acest fapt este unul foarte extins, dar din fericire nu unul complet. Din 100.000 de tweeturi totale, circa 1214 tweeturi cu label-ul 'low-quality' sunt valabile, iar cele cu label-ul 'not low-quality' sunt in numar de 15942. Pentru un dataset balansat, se vor folosi aproximativ acelasi numar de instante pentru fiecare categorie, fiind destule observatiile in numar de aproximativ 1200.

In final, distributia datasetului se face in doua fisiere , cate unul pentru fiecare categorie. Deoarece exista diferente intre encodingul acestor fisiere si encodingul acceptat de interpretorul Python, fara ca sa existe caractere iligibile, trebuie folosita o functie care curata datasetul.

De asemenea, pe Twitter exista conceptul de retweeting care permite utilizatorilor sa distribuie

pe propriul cont anumite postari ale altor persoane. Pentru a semnala acest lucru, Twitter adauga un substring de forma “RT @utilizator_cu_postarea_originala: “ respectivului tweet. Avem nevoie sa eliminam acest tip de substring din orice tweet ce il contine.

```
def printAndSaveTweetTextFromCsv(file):
    newfile = open(os.path.splitext(file)[0] + "_cleanLOWERCASE.txt", 'w', encoding='ascii')
    with open(file, encoding='latin-1') as csvfile:
        reader = csv.DictReader(csvfile)
        for row in reader:
            normal = [x.lower() for x in row['text'].split()]
            normal_stringed = ' '.join(map(str, normal))
            cleaned = unicodedata.normalize('NFKD', normal_stringed).encode('ascii', 'ignore')
            tokenized = tokenizer.tokenize(cleaned)
            if len(tokenized) > 0:
                cut = tokenized[2:]
                full = tokenized
                #print( cut if tokenized[0] == "RT" else full )
                newfile.write(' '.join(cut)) if tokenized[0] == "rt" else newfile.write(' '.join(full))
            newfile.write('\n')
```

Pe langa acest lucru, functia “printAndSaveTweetTextFromCsv” tokenizeaza tweeturile folosind biblioteca nltk.tokenize, procedeu descris in primul capitol, transpune orice caracter din tweet in echivalentul lowercase si codifica totul din format latin-1 in format ascii normalizand unicode in format NFKD cu funtia 'normalize' din biblioteca unicodedata, pentru a elimina caracterele iligibile datorata codificarii utilizate de Twitter. Noile tweeturi sunt salvate intr-un nou fisier de tip text cu un anumit suffix.

Fisierele rezultate sunt citite si incarcate in memorie cu functia textFile a unui obiect de tip SparkContext.

```
fake = sc.textFile("a1newCSVFullTweets_cleanLOWERCASE.txt")
real = sc.textFile("a0newCSVFullTweets_cleanLOWERCASE.txt")
```

Fiecare tweet este transformat intr-o lista de cuvinte folosind functia 'map' din Python.

```
fake_words = fake.map(lambda sentence: sentence.split())
real_words = real.map(lambda sentence: sentence.split())
```

```
[['same', 'https://t.co/1ghdvqvzrc']]
[['heads', 'low', 'hopes', 'high', '~']]
```

*Exemplu de tweeturi low-quality (sus) si
non-low-quality (jos)*

Se hash-uieste fiecare term si pentru fiecare tweet va rezulta un SparseVector, un vector rar, ce contine numarul de feature-uri (in fiecare caz, 2*18), hash-ul fiecarui term si numarul de aparitii a acestuia.

```
tf = HashingTF(numFeatures=2**18)
fake_features = tf.transform(fake_words)
real_features = tf.transform(real_words)
```

```
[SparseVector(262144, {179060: 1.0, 232159: 1.0})]
[SparseVector(262144, {5995: 1.0, 18426: 1.0, 100779: 1.0, 162531: 1.0, 170314: 1.0})]
```

Vectorul de features pentru tweeturile de mai sus.

Folosind LabeledPoint din `pyspark.mllib.regression`, putem adauga eticheta pentru fiecare astfel de SparseVector, respectiv eticheta 1 pentru low-quality si 0 altfel.

Folosind functia randomSplit, impartim aleatoriu setul de date in set de date de antrenare si set de date de testare. Datele de antrenare vor reprezenta 80% din total, iar datele de testare vor reprezenta 20% din total.

Mai departe, putem deja antrena si testa un model astfel.

```
algorithm = LogisticRegressionWithLBFGS()
model = algorithm.train(training_data)
print('logistic regression with lbfgs:', score(model))
```

Functia 'score' calculeaza acuratetea modelului cu urmatoarea formula [3] (wikipedia).

Accuracy (ACC)

$$ACC = (TP + TN) / (P + N)$$

Aceasta formula reprezinta raportul dintre instantele True Positives + True Negatives si Positives + Negatives, adica numarul de instante a caror predictie a fost corecta supra totalul instantelor. O predictie true positive descrie o instanta pozitiva a carei predictie a fost calculata ca fiind pozitiva. O predictie true negative descrie o instanta negativa a carei predictie a fost calculata ca fiind negativa.

```
def score(model):
    features = []
    for element in test_data:
        features.append(element.features)

    predictions = model.predict(features)

    labels = []
    for element in test_data:
        labels.append(element.label)

    labels_with_predictions = zip(labels, predictions)

    elements_gotten_right = []
    for element in labels_with_predictions:
        if element[0] == element[1]:
            elements_gotten_right.append(element)
    return len(elements_gotten_right) / float(len(test_data))
```

In total au fost testate 6 modele, dintre care 2, decision tree si random forests aveau nevoie de prea multa memorie datorita numarului de features prea mare. Celealte patru au oferit rezultate interesante si destul de apropiate.

```
logistic regression sgd: 0.9186079953983319
logistic regression with lbfgs: 0.913718723037101
naive bayes: 0.9350014380212827
svm with sgd: 0.9194708081679609
```

Acuratetea modelelor

Surprinzator, algoritmul naiv al lui Bayes ofera acuratetea cea mai mare dintre cele 4, deci acesta ramanand a fi folosit pentru a clasifica binar apartenenta unui tweet la una dintre categoriile 'low-quality' si 'non-low-quality'. Clasificatorul este folosit in cadrul scriptului ce se ocupa cu colectarea datelor. Fiecare tweet este adaugat intr-un fisier CSV ce contine textul, locatia de unde a fost trimis tweetul si eticheta pusa de clasificator.

3 Clusterizare

Clusterizarea este metoda de invatare nesupervizata ce are ca scop gruparea unor obiecte astfel incat instantele din acelasi grup, numit cluster, sa fie mai asemanatoare intre ele decat fata de alte instantele din alte clustere.

[4]“Este unul din obiectivele principale ale minării de date si o tehnica comuna in analiza statistica datelor. Se foloseste in multe domenii, cum ar fi machine learning, recunoasterea pattern-elor, analiza imaginilor, bioinformatica, compresia datelor si grafica pe calculator.”

Exista multi algoritmi ce se ocupa cu clusterizarea unor date, deoarece exista multe interpretari a ceea ce poate insemna un cluster sau cum se poate crea si modela un cluster in cadrul implementarilor

Modelele de cluster pot fi urmatoarele:

- modele de conectivitate
- modele bazate pe centroizi
- modele de distributie
- modele de densitate
- modele subspatiu
- modele de grup
- modele bazate pe grafuri
- modele neurale

Pentru fiecare din aceste modele exista numeroare exemple de algoritmi. De exemplu, pentru modelele de clustere bazate pe conectivitate, exista clusterizare ierarhica, ce urmareste sa cladeasca o ierarhie de clustere. De obicei, este un algoritm greedy ce poate fi de tip aglomerativ, “bottom up”, (unde orice observatie porneste in propriul cluster si perechi de cluster se imbina o data ce un cluster urca in ierarhie) sau diviziv, “top down” (unde toate observatiile pornesc intr-un singur cluster si se efectueaza splituri incepand cu acest cluster), iar rezultatele clusterizarii, adica ierarhia sunt descrise intr-o dendrograma.

Un algoritm ce lucreaza cu clusteri din modelul bazat pe centroizi este algoritmul k-means, ce reprezinta un cluster ca fiind un vector de elemente ce au un centroid, centrul acelor instantele, descris ca medie a pozitiilor fiecarui element.

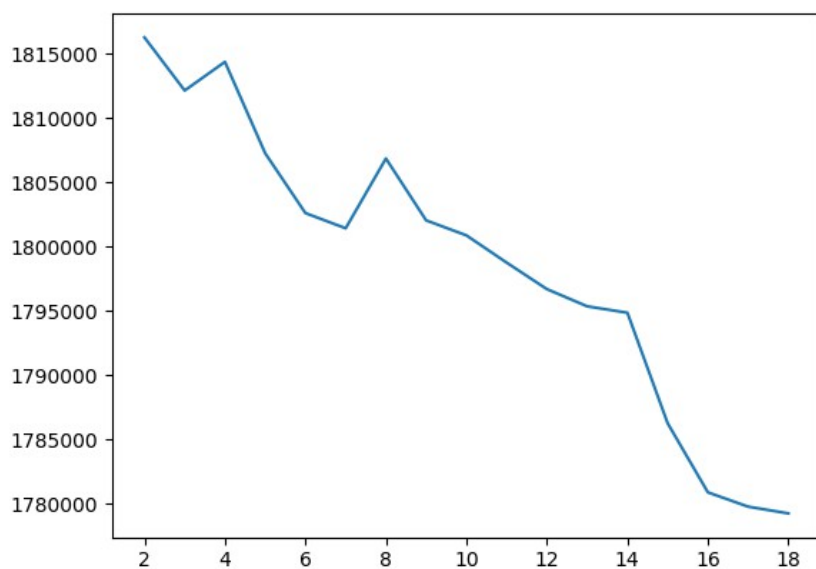
In modelul de distributie, clusterelor sunt descrise folosind distributii statistice, cum ar fi distributii normale multivariate, in cadrul algoritmului EM (expectation-maximization), care este o metoda iterativa de a gasi parametrii unui model statistic.

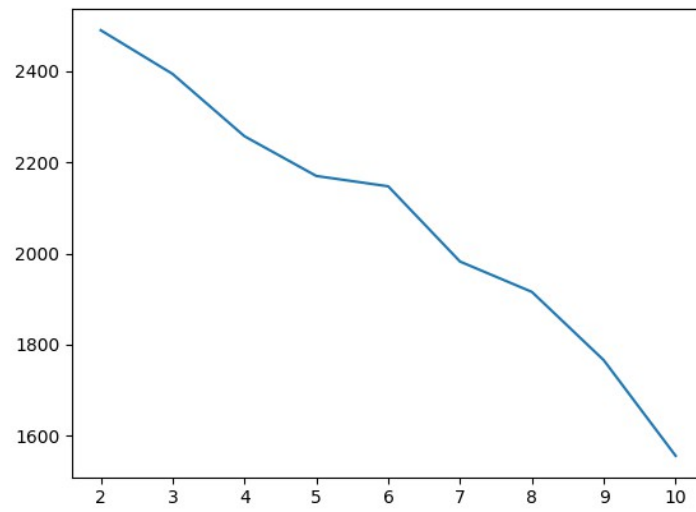
Modelele de densitate caracterizeaza clusterelor ca fiind niste zone dense de regiune in spatiul datelor si sunt utilizate in algoritmi DBSCAN si OPTICS.

Modelele subspatiu sunt folosite in biclusterizare, unde clusterelor sunt modelate si cu membrii clusterelor, si cu attributele relevante.

Algoritmi ce folosesc modelele cu grupuri nu produc doar informatia despre cum se face gruparea, si nu un model rafinat pentru rezultate.

„Clusterizarea, sau analiza de tip cluster, nu se refera la un algoritm specific, ci la obiectivul general ce trebuie atins. Acesta poate fi indeplinit de diferiti algoritmi ce difera destul de mult intre ei, prin prisma faptului ca clusterelor pot fi create si interpretate foarte diferit, in functie de implementare. Anumite interpretari ale clusterelor include grupuri cu distante mici intre membrii clusterelor, arii dense in spatiul datelor, intervale sau distributii statistice particulare. Deci, metoda de clustering poate fi formulat ca o problema de optimizare cu mai multe obiective. Algoritmi potriviti si parametrii alesi depind de datasetul problemei si utilizarea rezultatelor. Clusterizarea nu este deci o sarcina automata, ci un proces iterativ de knowledge discovery (procesul automatizat de cautare a tiparelor in volume mari de date) sau optimizare interactiva cu mai multe obiective ce implica mai multe incercari.”





Bibliografie

Links

- [1] <https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731>
- [2] <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182487>
- [3] https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers
- [4]
- [5]