Praktikos Užduotis

Martynas Jašinskas

Turinys:

* [Trumpas pristatymas](#Pristatymas)
* [Įrankiai](#Irankiai)
* [Detalus aprašymas](#Detalus_aprasymas)
* [Rezultatai](#Rezultatai)

Pristatymas:

Šios programos trumpas funkcijos aprašymas:

1. Iš socialinio tinklo „Twitter“ rinkti žinutes, susijusias su naujienomis apie kibernetinę saugą.
2. Sėkmingai išgautas žinutes lengvai „išvalyti“ – pašalinti(URLs, #hash\_tags, @mentions) ir kitokias „Twitter“ žinučių atributikas
3. Apdorotoms žinutėms atlikti sentimentų analizę, skirtą kategorizuoti žinutes pagal nuspėjamą turinio nuotaiką.
4. Baigtinius duomenis išsaugoti „Hadoop“ duomenų bazėje

Projektas galimas: <https://github.com/Miautawn/TwitterBetterGetter>

Įrankiai:

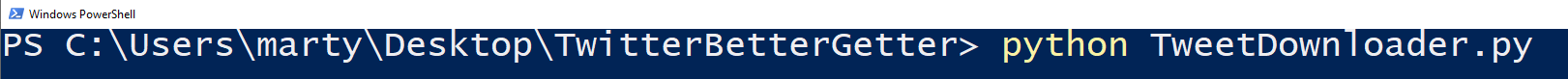
Šis projektas naudoja:

* Python3 bibliotekas: tweepy(Twitter API), nltk, pandas, hdfs(Hadoop distributed file system)
* Twitter Developer Apps – API <https://developer.twitter.com/en.html>
* VADER (Valence Aware Dictionary for sEntiment Reasoning) sentimentų analizės modelį.
* Hadoop duomenų bazę.

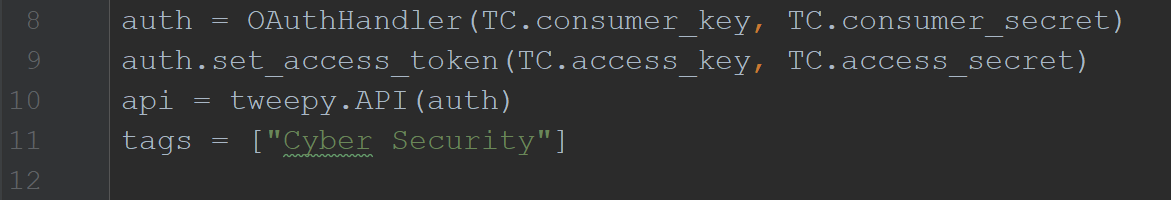
Detalus aprašymas:

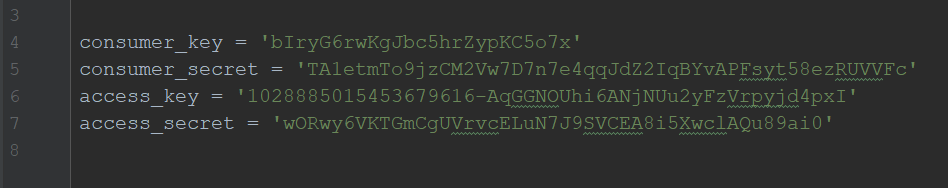
Čia aprašomas programos veikimo principas:

Programa pradeda veikti paleidus „TweetDownloader.py“ failą:

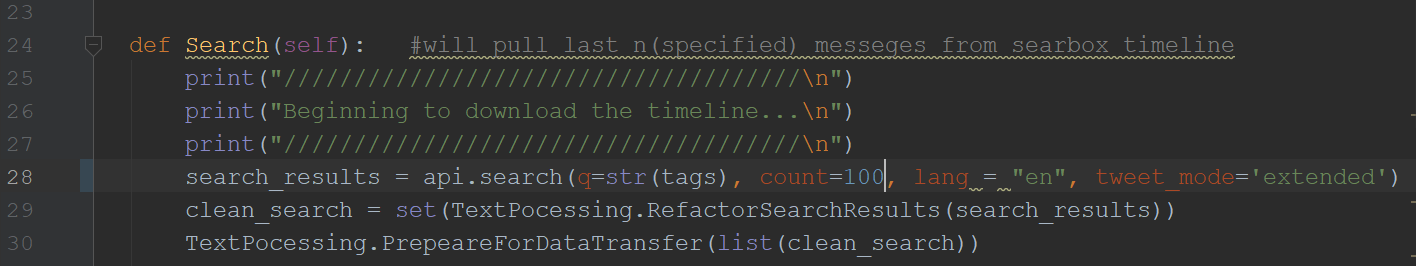


Tuomet, mes autentifikuojame savo Twitter API tokenus, ir nurodome savo raktinius žodžius, pagal kuriuos bus siučiamos žinutės:





Baigus autentifikavimą, Twitter API atsiunčia n skaičių žinučių iš nurodytų raktažodžių paieškos (funkcija „Search“):



Čia, atsiųstos žinutės yra apdorojamos:

* ištraukiamas tekstas(funkcija „RefactorSearchResults“)



* išvalomi URLs(funkcija „ExtractURLs“)



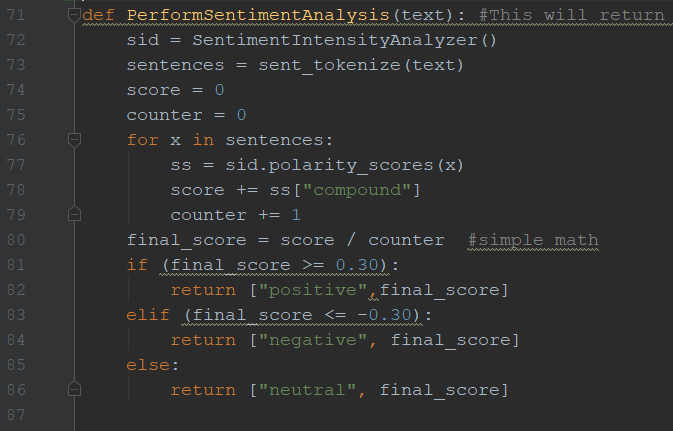
* išvalomos Twitter žinučių atributikos(funkcija „TextCleaning“)



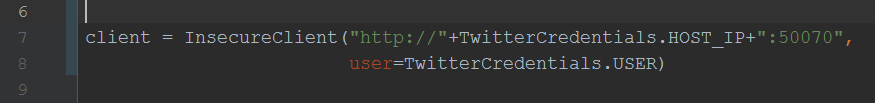
Galiausiai apdorotam tekstui yra daroma sentimentų analizė, pasitelkiant VADER modelį. Šis modelis nėra mašininio mokymosi tipo, bet verčiau iš anksto nustatytų gramatikos taisyklių ir žodymo specifikacijų rinkinys.

<https://www.nltk.org/modules/nltk/sentiment/vader.html>

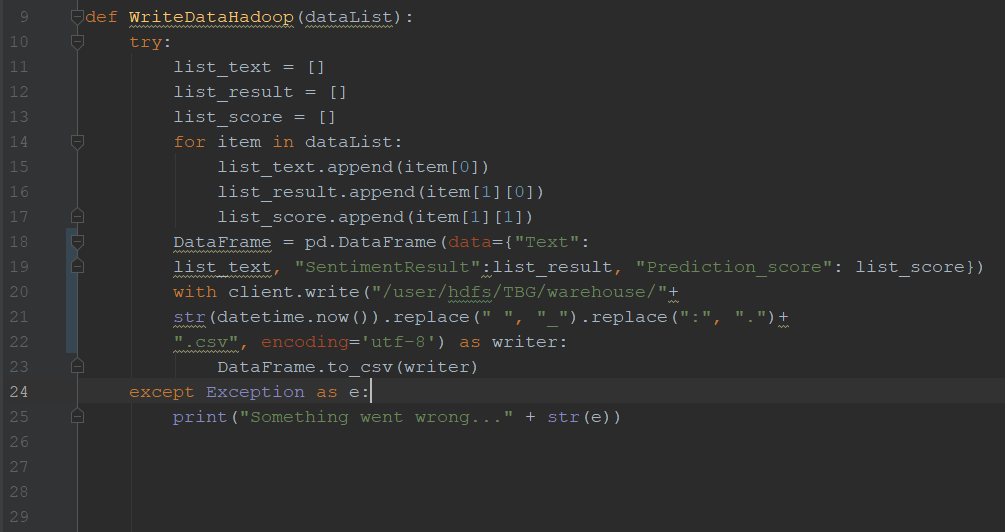
Nors šis modelis nėra tikslesnis už kitus ML modelis, jo privalumai pasireiškia, kai informacija keliauja tiesiogiai iš duomenų kanalų, kadangi jis yra daug greitesnis už kitus. (funkcija „PerformSentimentAnalysis“):



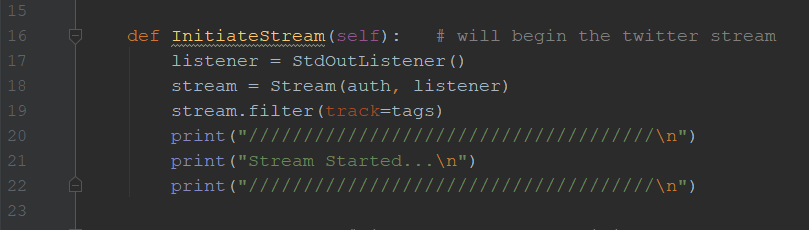
Gavus sentimentų analizės rezultatus, belieka įkelti duomenis į hadoop. Tai padarome su hdfs biblioteka, kuri mums leidžia prisijungti bei naršyti/skaityti HDFS. (Reikia nurodyti savo duombazės IP adresą ir prisijungimo duomenis)



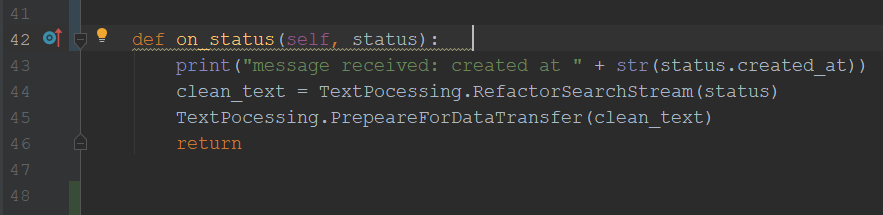
Gavus kliento prisijungimą, duomenis įkelsime csv formatu, 3 skyriais: tekstas, sentimentų analizės rezultatas, spėjimo tikimybė. (funkcija WriteDataHadoop):



Baigus siųsti senas žinutes, įsijungia tiesioginis Twitter duomenų kanalas(funkcija InitiateStream):

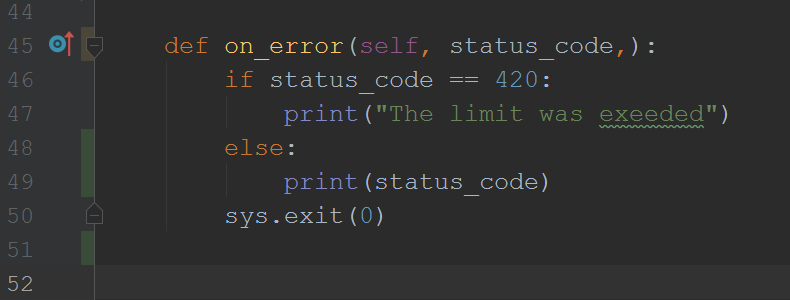


Gavus žinutes, suveikia funkcija „on\_status“, kuri mums duoda naujų Twitter žinučių duomenis:



Tuomet žingsniai kartojasi su kiekviena nauja teksto žinute.

Jeigu tiesioginio duomenų kanalo srauto metu įvyks klaida, suveiks funkcija „on\_error“, kuri terminuos programos ciklą:



Rezultatai

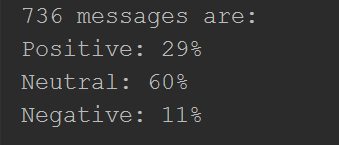
Galutiniai duomenys, kurie laikomi duomenų bazėje atrodo taip:

**Text:** „*Mueller investigation did nothing to stop the next Russian attack*“

**SentimentResult:** „negative“

**Prediction\_score:** „-0.2987“

Iš per ~30min surinktų duomenų galime matyti, kad:

*Iš 736 žinučių:*

*29% jų yra pozityvios*

*60% jų yra neutralios*

*11% jų yra neigiamos*

Taip pat patikrinę daugiausiai pasikartojančius žodžius, galime pamatyti kokios temos labiausiai domina arba yra aktualiausios tarp išrinktų žinučių.

Žiūrėti vizualizaciją: <https://github.com/Miautawn/TwitterBetterGetter/blob/master/FrequencyCheck.png>

Čia galime pamatyti tokių terminų kaip: AI, nsa, cybercrime, blockchain it t.t.

Frazės kaip: attack, databreach, ransomeware, phising, hacker, dataprotection ir Russian 😊, gali rodyti kibernetinių atakų ir nusikaltimų padaugėjimą.

Kol žodžiai kaip: IoT (Internet of Things), cloud, Big Data ir 5G gali sufleruoti kokiomis technologijomis žmonės domisi.

Šiuo metodu galime pamatyti koks buvo popouliariausias linkas, kurį vartotojai nurodydavo savo žinutėse:

<https://twitter.com/i/web/status/1153641331253813249>

Pavyzdžiui šis nukreipia į twitterio skelbimą, kuriame rašoma apie NSA norus kurti nauja JAV kibernetinio saugumo vadybą.