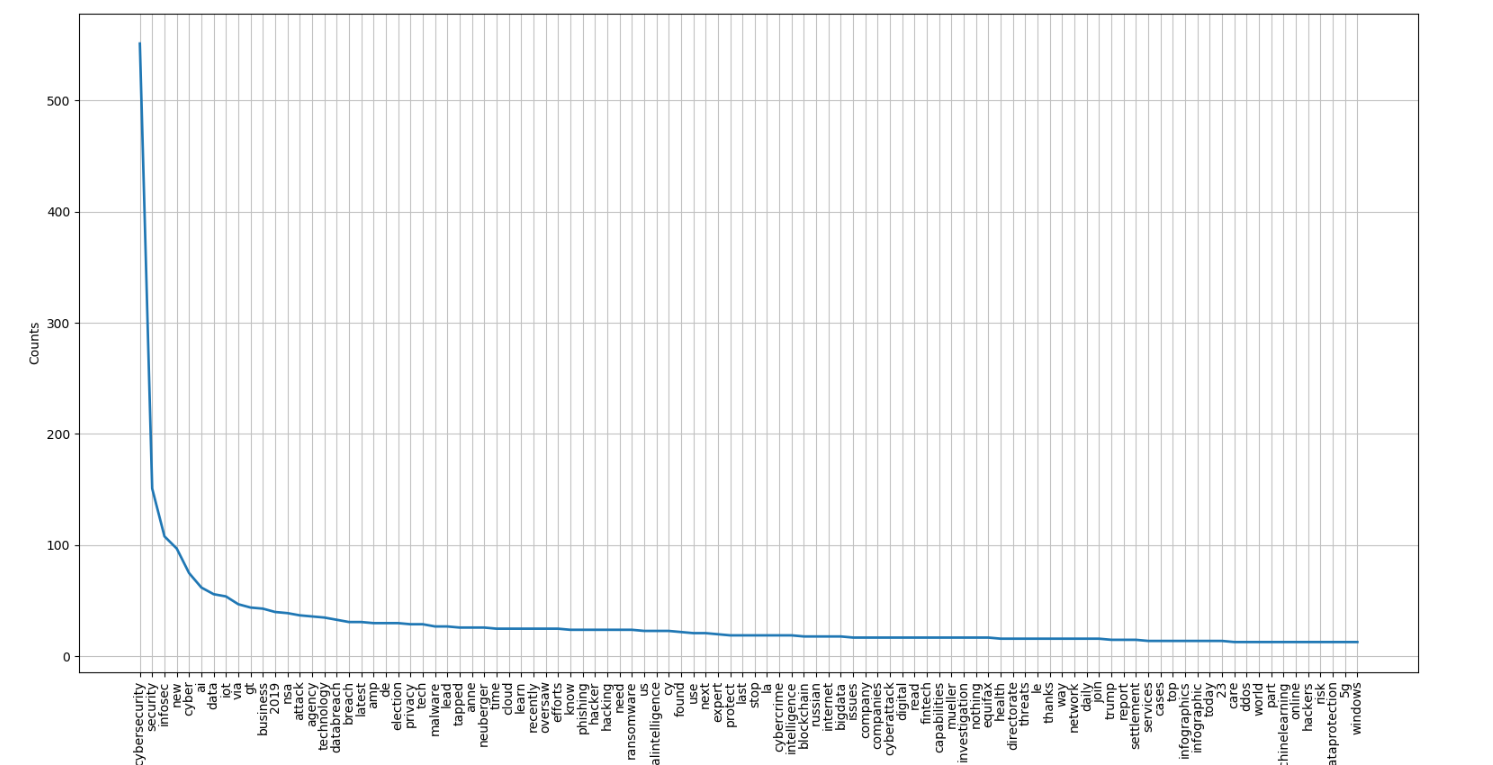
Praktikos užduotis (toliau – Projektas)

Martynas Jašinskas



Turinys:

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

Pristatymas:

Projekto trumpas 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: <https://github.com/Miautawn/TwitterBetterGetter>

Įrankiai:

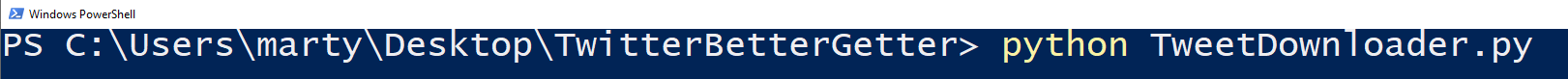
Projekto programa naudoja:

* 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.

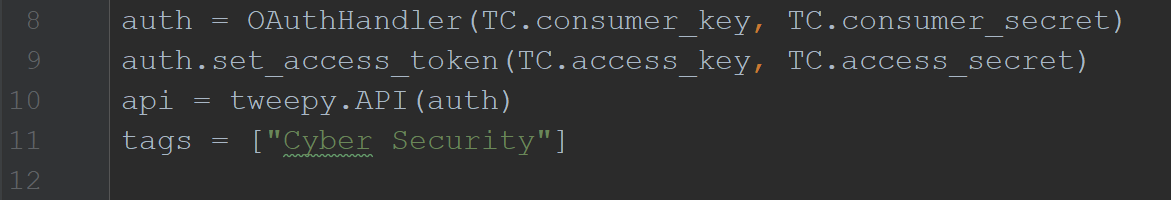
Detalus aprašymas:

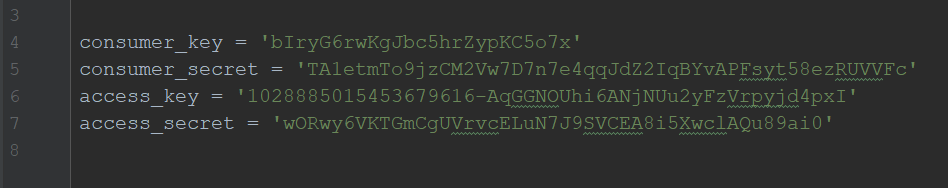
Šioje dalyje aprašomas programos veikimo principas:

Programa pradeda veikti paleidus „TweetDownloader.py“ failą:

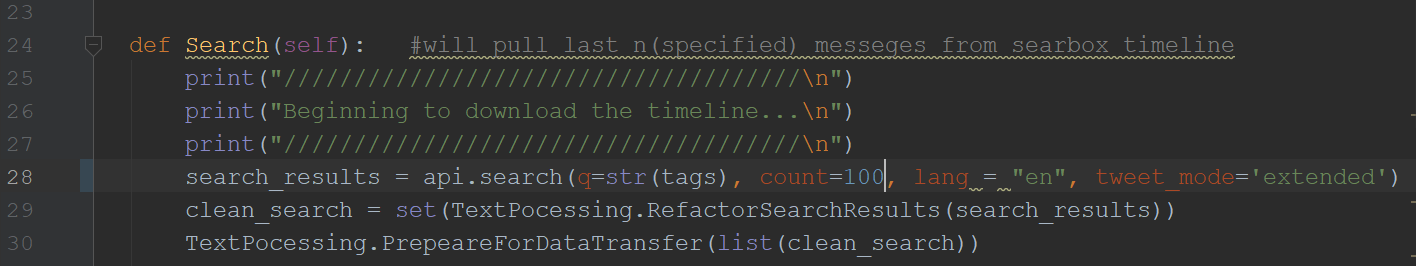


Tuomet, mes autentifikuojame savo Twitter API tokenus, ir nurodome savo raktinius žodžius, pagal kuriuos bus siunčiamos žinutės:





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



Čia, atsiųstos žinutės yra apdorojamos:

* išgaunamas tekstas (funkcija „RefactorSearchResults“)



* išvalomi URLs (funkcija „ExtractURLs“)



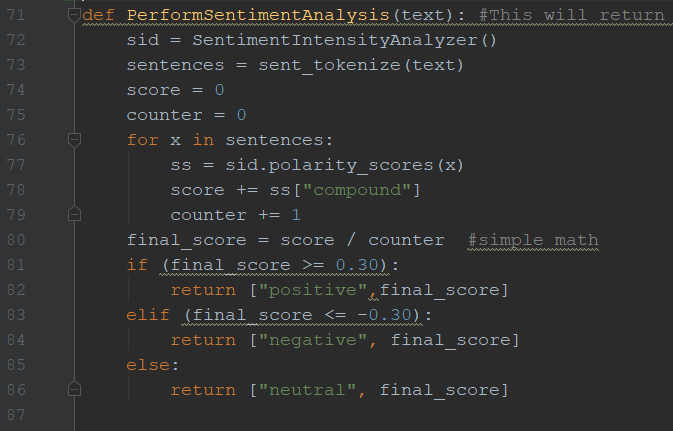
* išvalomos Twitter žinučių atributikos (funkcija „TextCleaning“)



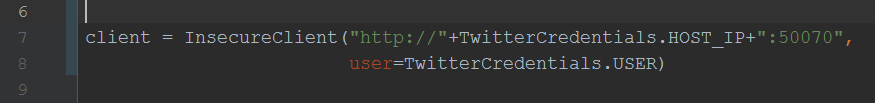
Toliau apdorotam tekstui yra daroma sentimentų analizė, pasitelkiant VADER modelį. Šis modelis nėra mašininio mokymosi tipo, bet iš anksto nustatytų gramatikos taisyklių ir žodyno specifikacijų rinkinys.

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

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

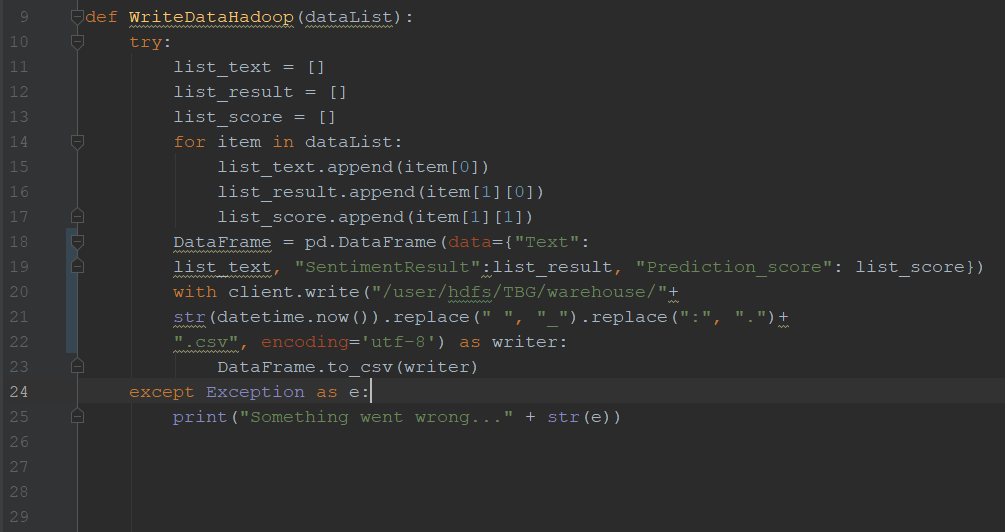


Gavus sentimentų analizės rezultatus, belieka įkelti duomenis į hadoop. Tai padarome su hdfs biblioteka, kuri mums leidžia prisijungti bei naršyti/skaityti HDFS. Reikia nurodyti savo duomenų bazės IP adresą ir prisijungimo duomenis:

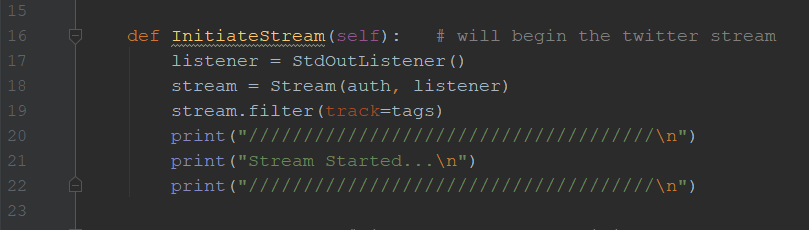


Gavus kliento prisijungimą, duomenis įkeliame csv formatu, 3-is skyriais (funkcija „WriteDataHadoop“):

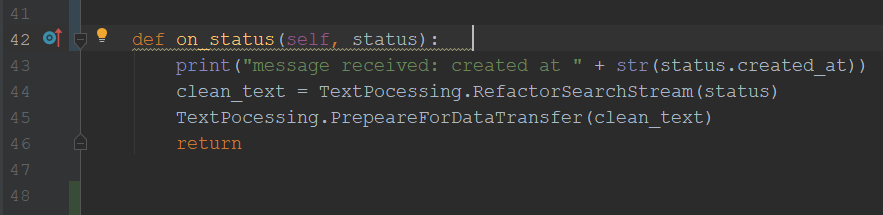
* tekstas;
* sentimentų analizės rezultatas;
* spėjimo tikimybė.



Baigus siųsti senas žinutes, įsijungia tiesioginis Twitter duomenų srautas (funkcija „InitiateStream“):

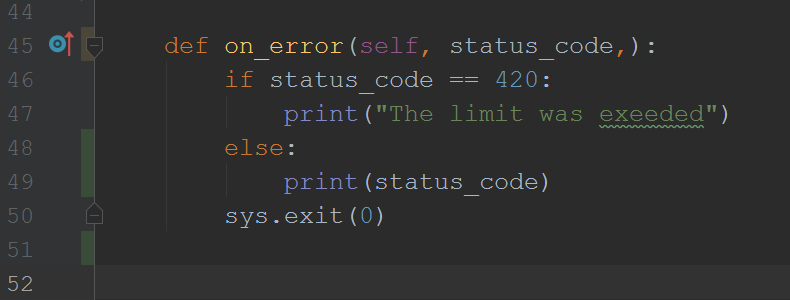


Gavus naujas žinutes, suveikia funkcija „on\_status“, kuri pateikia Twitter žinučių duomenis:



Tuomet žingsniai kartojasi su kiekviena nauja teksto žinute.

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



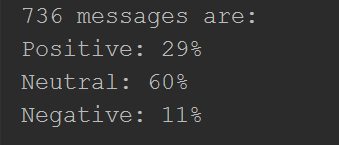
Rezultatai

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 duomenų, surinktų per ~30 min., matyti, kad:

*Iš 736 žinučių:*

*29% yra pozityvios*

*60% yra neutralios*

*11% yra neigiamos*

**Išvados**

Įvertinę dažniausiai pasikartojančius žodžius, galime pamatyti kokios temos labiausiai domina arba yra aktualiausios tarp surinktų žinučių.

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

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

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

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

Šiuo metodu taip pat galime pamatyti koks buvo popouliariausia nuoroda (*URL*), kurį vartotojai nurodydavo savo žinutėse:

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

Ši nuoroda nukreipia į twitterio skelbimą, kuriame rašoma apie NSA ketinimus kurti naują JAV kibernetinio saugumo padalinį.