Praktikos Užduotis

Martynas Jašinskas

Turinys:

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

Pristatymas:

Šios paprastos programos funkcijos aprašymas:

1. Iš socialinio tinklo „Twitter“ rinkti, analizuoti ir duonbazėje išsaugoti žinutes, susijusias su naujienomis apie kibernetinę saugą.
2. Sėkmingai išgautas žinutes lengvai „išvalyti“ – pašalinti(URLs, #hash\_tags, @mentions) ir kitokias „Twitter“ žinučių atributikas
3. Lengvai 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:

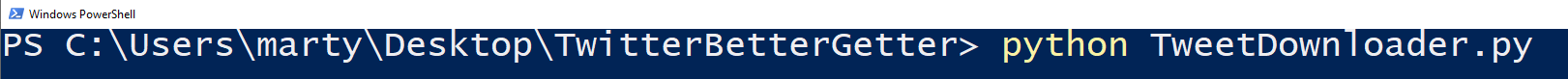
Šis projektas naudoja:

* 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.

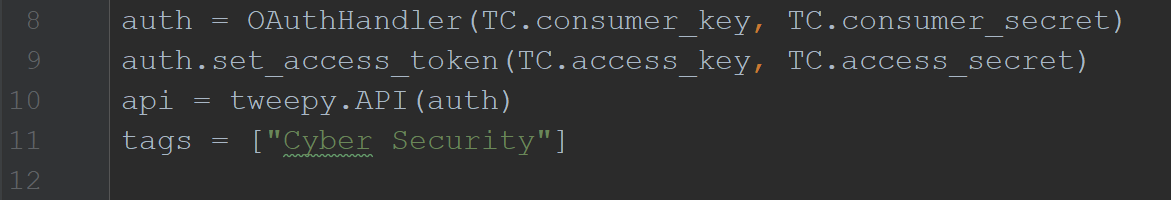
Detalus aprašymas:

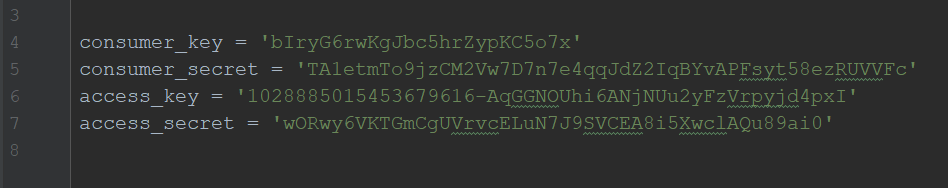
Čia aprašomas programos veikimo principas

Programa pradeda veikti paleidus „TweetDownloader.py“ failą:

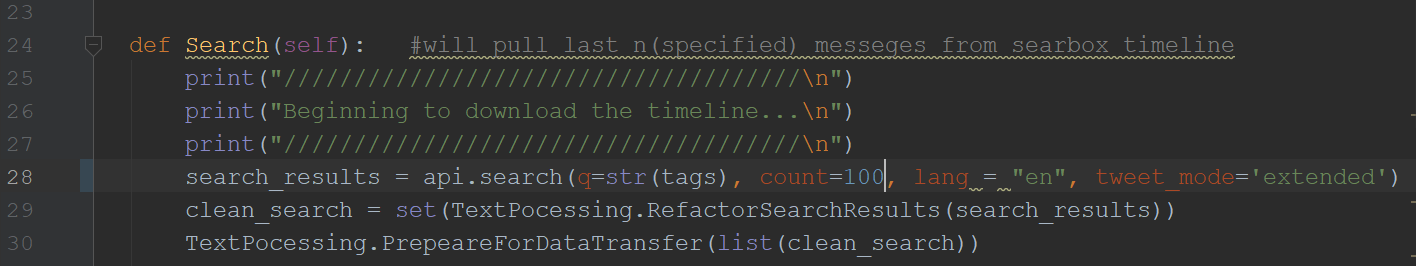


Tuomet, mes autentifikuojame savo Twitter API tokenus, ir nurodome savo raktinius žodžius, pagal kuriuos bus siučiamos žinutės:





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“)



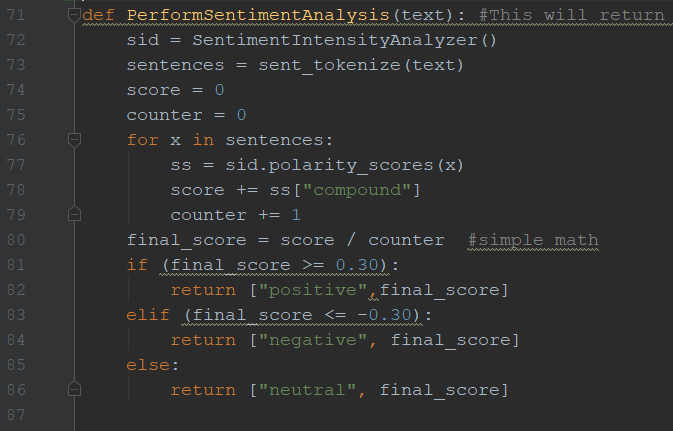
* išvalomos Twitter žinučių atributikos(funkcija „TextCleaning“)



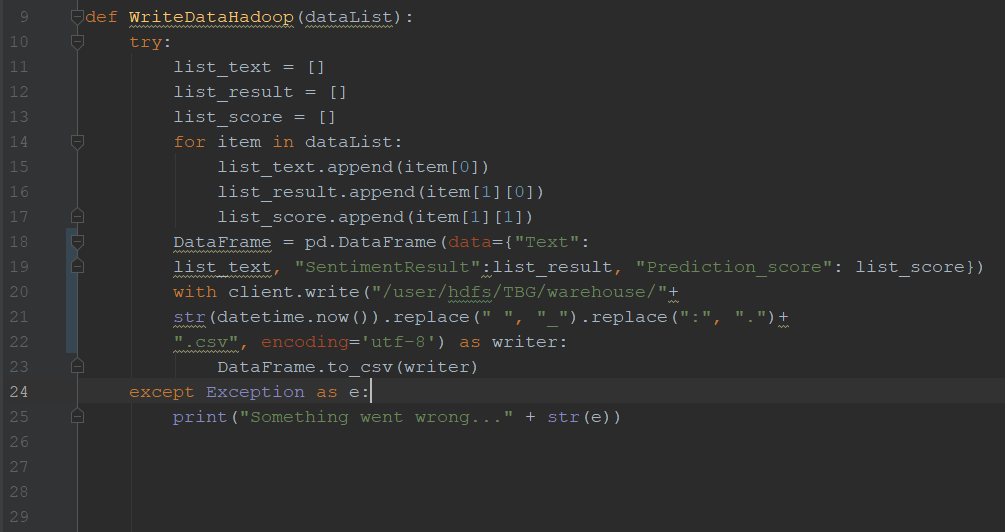
Galiausiai apdorotam tekstui yra daroma sentimentų analizė, pasitelkiant VADER modelį. Šis modelis nėra mašininio mokymosi tipo, bet verčiau iš anksto nustatytų gramatikos taisyklių ir žodymo specifikacijų rinkinys.

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

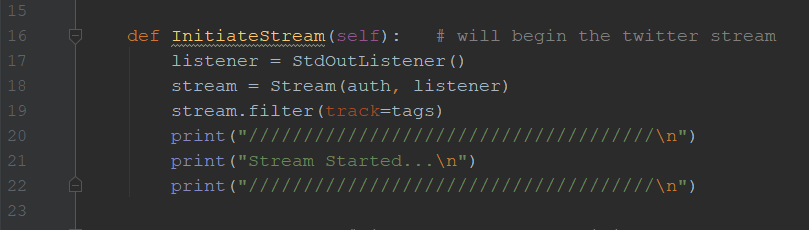
Nors šis modelis nėra tikslesnis už kitus ML modelis, jo privalumai pasireiškia, kai informacija keliauja tiesiogiai iš duomenų kanalų, kadangi jis yra daug greitesnis už kitus. (funkcija „PerformSentimentAnalysis“):



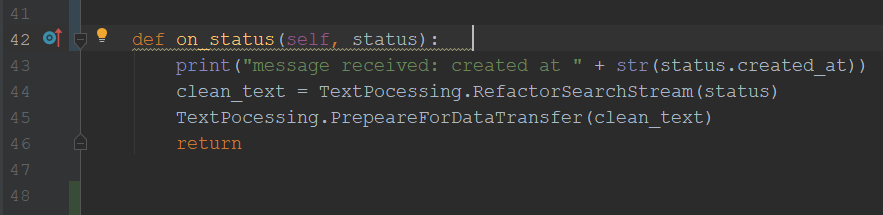
Gavus sentimentų analizės rezultatus, belieka įkelti duomenis į hadoop. Tai padarome su hdfs biblioteka, kuri mums leidžia prisijungti bei naršyti/skaityti HDFS. Duomenis įkelsime csv formatu, 3 skyriais: tekstas, sentimentų analizės rezultatas, spėjimo tikimybė. (funkcija WriteDataHadoop):



Po to, 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.