

**KAUNO TECHNOLOGIJOS UNIVERSITETAS**

**MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

**Vytautas Kraujalis**

Studijų modulio

P160M121 Duomenų tyrybos metodai ir programinės priemonės

**Laboratorinis darbas nr. 4**

Dėstytojas: V. Janilionis

**KAUNAS, 2022**

# UŽDUOTIS

1. Sukurkite savo Twiter paskyrą.

2. Sukurkite Twiter aplikaciją.

3. Pasirinkite laisvai prieinamą, kitų studentų dar nepasirinktą Twiter žinučių paskyrą (pvz. @sciam - Scientific American) ir ją užregistruokite L4 pasirinktų duomenų registravimo žurnale.

4. Iš pasirinktos paskyros importuokite 150-300 Twiter trumpų žinučių (angl. twiets) rinkinį.

5. Panaudoję teksto tvarkymo, filtravimo ir transformavimo priemones bei klasterizavimo metodus atlikite pasirinkto Twitter trumpų žinučių rinkinio analizę ir klasterizavimą. Sukurtą R programą pademonstruokite dėstytojui laboratorinio darbo metu ir atsakykite į pateiktus klausimus. Laboratorinio darbo ataskaitoje (PDF formate) pateikite uždavinio sprendimo eigą, R programą, gautus rezultatus ir komentarais bei išvadas. Parengtą ataskaitą iki dėstytojo nurodyto laiko pabaigos įkelkite į nurodytą Moodle katalogą.

**ATASKAITA**

Pasirinkta Twitter paskyra - @Tesla:

A picture containing graphical user interface

Description automatically generated

Susikūrę Twitter applikaciją nuskaitėme 300 @Tesla trumpų žinučių. Keletos žinučių fragmentas:

[[1]] Fun on tap <U+0001F3A2> <U+0001F3A5> @saradietschy https://t.co/o0lhc1Webd https://t.co/BStyuW1BZF

[[2]] In 2021, we received almost 3 million job applications. Join us! https://t.co/wOIwB0p5rf https://t.co/Z5MJo9khrt

[[3]] We created nearly 100k direct new jobs in just one decade – and are planning to grow for years to come <U+270C> https://t.co/xSAcfvyT0s

[[4]] At the end of their usable life, we recycle battery packs to build brand new batteries https://t.co/zfPt8MPMdU

[[5]] Battery longevity is critical. Every battery we make is designed to outlast the vehicle it’s in https://t.co/ZbTzQN55u4

Galime pastebėti, kad reikės nemažai tvarkymo šioms žinutėms. Turėsime pašalinti Unicode formato simbolius (tokius kaip <U+0001F3A2>), kitų žmonių/puslapių paminėjimus (pvz. @Tesla, @saradietschy), internetinių puslapių nuorodas (https://.....), atsakymų į kitas žinutes pradžias (prasideda pvz RT @elonmusk), “hashtag” žodžius (pvz. #energy).

Pateiksiu visą nuoseklų sąrašą, kokie žinučių tvarkymai buvo atlikti (skliaustuose pateikiu pavyzdį):

1. Pašalinau URLs (httpss://....)
2. Pašalinau kitų paskyrų paminėjimus (@my\_account)
3. Pašalinau „hashtag“ žodžius (#energy)
4. Pakeičiau & simbolius, kuriuos konvertavo į „&amp;” į žodį „and“
5. Pašalinau taškus, kablelius
6. Pašalinau žodį „RT:“ nuo žinučių pradžios, kas parodė atrašymą į kitą žinutę.
7. Pakeičiau naujos eilutės simbolį „\n“ tarpu
8. Pakeičiau visas raides mažosiomis
9. Pašalinau <U+???> unicode žodžius
10. Pašalinau ne ASCII formato raides, įtraukiant ir šypsenėles
11. Pakeičiau visus ne alfanumerinius simbolius į tarpus
12. Pakeičiau visus skaičius tarpais
13. Pašalinau 1-2 raidžių žodžius
14. Pakeičiau išsireiškimą „solar roof“ į „solar panel“. Vėliau pakomentuosiu šią vietą
15. Pakeičius visus tarpus, kad liktų tik vienas tarpas tarp žodžių.
16. Pašalinau anglų kalbos „stop“ žodžius bei šiuos žodžius: „tesla“, „just“, „one“, „every“, „deliveries“, „production“ kadangi šiame kontekse šie žodžiai mūsų nedomina.
17. Sutvarkiau žodžių kamienus

Gavome žodžių žemėlapį:

Text

Description automatically generated

Pav. 1 Žodžių žemėlapis

Žemėlapis jau parodo mūsų pasirinktą temą – Teslą. Nepašalinę “Tesla” žodžio, turėtume šį žodį pati didžiausią. Dabar gauname su šia kompanija susijusią informaciją:

Model – Tesla kompanija prekiauja elektriniais automobiliais, kurie vadinasi Tesla model X, Tesla model Y ir t.t. Todėl šiuo raktažodžiu identifikuojame kuomet kalbama apie automobilius.

Solar, panel – Tesla kompanija ant automobilių turi saulės kolektorius, todėl apie juos daug kalbą, taip pat kita įmonė prekiauja saulės kolektoriais ant namų, todėl tai čia irgi būna paminima.

Powerwall – Tesla kompanija prekiauja namų stotelėmis, kurias įsirengus galima pasikrauti automobilį.

Giga, berlin – Tesla kompanija įsteiginėja Gigafactory gamyklas, viena iš paskutinių – Berlyne.

Plaid – Teslos automobilių funkcija, automobilių grupės dalis, daug reklamuojamas funkcionalumas.

Taip pat sudarę žodyną iš šių žinučių, turime 993 požymius su “Sparsity” lygiu 99%. Žodyną sumažinome iki 20 požymių ir tuomet gauname „Sparsity“ lygį 95%.

Atlikome hierarchinį žodžių klasterizavimą:

Diagram, schematic

Description automatically generated

Pav. 2 Hierarchinio klasterizavimo dendograma

Galim pastebėti kelias įdomias vietas: akivaizdžiai žodžiai “solar” ir “panel” būna kartu, nepakeitus išsireiškimo “solar roof” į “solar panel”, šioje vietoje būtume mate “solar” ir “roof”. Tačiau šiose žinutėse tai reiškia vieną ir tą patį – saulės kolektorius ant automobiliu stogo, tačiau būna tik skirtingas išsireiškimas. Taip pat “giga” kartu su “berlin”, kadangi šiuo metu statoma Gigafactory (vadinama giga) gamykla būtent Berlyne. Likę žodžiai jau ne taip susiklasterizuoja

Taip pat galime pamatyti, šiek tiek prasčiau, jog klasterizuojasi “Texas” ir “safety” žodžiai, paieškojus informacijos pamačiau, jog buvo įvykusi avarija Texas valstijose susijusi su automobilio saugumu.

Atlikę medoidų klasterizavimą gavome rekomenduojamus 8 klasterius:

cluster 1 :

cluster 2 : solar panel

cluster 3 : model

cluster 4 : giga texas

cluster 5 : giga berlin

cluster 6 : now

cluster 7 : model plaid

cluster 8 : powerwall

Tačiau klasteriai pagal interpretaciją vienas su kitais persipinę. Pavyzdžiui “giga Texas” ir “giga berlin” iš esmės kalba apie Gigafactory gamyklas, todėl vertėtų pabandyti apjungti. “model” ir “model plaid” kalba apie automobilius, tik pastarasis išskiria konkretų modelį, todėl vertetų apjungti irgi. Powerwall asocijuojasi su “solar panel”, kadangi šios stotys paprastai maitinamo nuo saulės kolektorių, esančių ant pastato, todėl galbūt pavyktų apjungti. Sumažinau klasterių skaičių iki 4:

cluster 1 :

cluster 2 : solar panel

cluster 3 : model

cluster 4 : giga berlin

Gauname klasterius, panašius į jau apkalbėtus, kuomet darėme žodžių klasterizavimą hierarchiniu metodu.

Chart

Description automatically generated

Pav. 3 Silueto plotis

Vidutinį silueto plotį gavome 0,53, kas jau rodo neblogą rezultatą. Pirmasis klasteris turi visas šiukšles, kurių negalėjo suklasterizuoti, į jį įtraukė 213 žinučių su vidutiniu 0,60 silueto pločiu. Antrasis klasteris, sudarytas iš “solar” ir “panel” turi 25 žinutes su vidutiniu 0,27 silueto pločiu, trečiame klasteryje turime “mode” su 43 žinutėmis ir 0,37 vidutiniu silueto pločiu, o ketvirtas klasteris – “giga berlin” – turi 19 žinučių bei 0,43 vidutinį silueto plotį.

Apžvelgsiu sudarytų klasterių žinutes:

Klasteris 2:

*[[1]][1] "Tesla: In the past decade, our solar panels produced more than enough clean energy to power every Tesla factory and every… https://t.co/DB03lEdN8W"*

*[[2]][1] "Tesla: RT @TeslaSolar: Solar Roof lets you become your own renewable energy utility. Generate clean energy and drive on solar.\n📸: @bethlyp on IG h…"*

*[[3]][1] "Tesla: Solar Roof keeps snow &amp; ice off your roof\n\n📸: @Xenius101 https://t.co/Xfj9XGRUX1"*

*[[4]][1] "Tesla: RT @TeslaSolar: New Solar Roof tiles can deliver a 22% increase in power, resulting in more energy offset https://t.co/pxxPPgeB8s"*

*[[5]][1] "Tesla: Solar Roof generating clean energy in Aspen, CO 🏔\n\n📸 @weddleandsons https://t.co/5Ar8ndjoXF"*

*[[6]][1] "Tesla: Solar Roof powers Buffalo Heritage Carousel in NY 🌞 https://t.co/cxFFwmuQ2e"*

*[[7]][1] "Tesla: Energy generation from Tesla Solar products vastly exceeds energy consumption of Tesla factories https://t.co/VchhrfJPKc"*

*[[8]][1] "Tesla: Our product line-up covers renewable energy from end to end, from solar panels to vehicles &amp; software designed to m… https://t.co/CZYC8td1bC"*

*[[9]][1] "Tesla: Solar Roof is designed to withstand heavy storms, hail and even medium-size tree branches https://t.co/IvTROrPC0j"*

*[[10]][1] "Tesla: RT @elonmusk: Solar + Powerwall battery ensures that your home never loses power"*

*[[11]][1] "Tesla: RT @elonmusk: Tesla Solar Roof + Powerwall in major new housing development in Austin\nhttps://t.co/u7NJAswFW2"*

*[[12]][1] "Tesla: RT @elonmusk: Tesla Solar + Powerwall battery enables consumers to be their own utility"*

*[[13]][1] "Tesla: RT @elonmusk: @AustinTeslaClub The Tesla Solar Roof is slightly hydrophobic, so water, snow or ice slide off easily"*

*[[14]][1] "Tesla: Solar Roof even clears off the snow"*

*[[15]][1] "Tesla: RT @TonyChoMrMagic: I just installed the largest (44KW) solar roof in Florida @teslaenergy @ ChoZen Retreat @livechozen, Thank you @elonmu…"*

*[[16]][1] "Tesla: Order solar panels before the solar tax credit drops at the end of this year. Pair with Powerwall to experience no… https://t.co/SRmsKlvQrx"*

*[[17]][1] "Tesla: There's no place like home..especially if it has a Solar Roof.\n\nThis Kansas ranch got theirs installed in 4 days."*

*[[18]][1] "Tesla: RT @UndecidedMF: In Tesla’s Q2 earnings call they dropped the little bombshell that Tesla's solar panels cost $1.49 per watt after incentiv…"*

*[[19]][1] "Tesla: Giga New York built 4MW of Solar Roof last week, enough for up to 1000 homes!"*

*[[20]][1] "Tesla: Solar Roof is taking off across the US. Become a Solar Roof installer today, no experience needed!… https://t.co/uRTQW0hkDB"*

*[[21]][1] "Tesla: Solar Roof uses durable glass solar tiles to generate electricity.\n \nOur latest version only takes a few days to se… https://t.co/LdC3Nzv5Cx"*

*[[22]][1] "Tesla: Submit questions for our Solar Roof Q&amp;A ahead of our 2pm launch:"*

*[[23]][1] "Tesla: Join us at 2pm PT today to learn all about our V3 Solar Roof product. \n\nVisit https://t.co/rd0ILdpzRL to listen in."*

*[[24]][1] "Tesla: RT @elonmusk: Get Tesla Solar plus Powerwall battery for 24/7 clean power &amp; no more blackouts! https://t.co/mDoPO17YB9"*

*[[25]][1] "Tesla: Also great news if you want to power your commute with the sun instead of toxic &amp; polluting gas: solar panels are n… https://t.co/nmRc5OKYPL"*

Kaip matome, klasterizavimas tikrai gana neblogai atliko darbą, šios visos žinutės yra susijusios arba su “solar roof” arba su “solar panel” arba dar “powerwall”, arba kitaip tariant, su saulės kolektoriais. Todėl pagal interpretaciją, šis klasteris tikrai gerai atskiria vieną iš Tesla kompanijos sričių. Sumažinę klasterių kiekį, žinutes, susijusias su saulės kolektoriais metodas apjungė į vieną klasterį ir taip išvengiame specifinės saulės kolektorių srities.

Klasteris 3:

*[[1]][1] "Tesla: The single-piece rear casting in Model Y replaces 70+ underbody parts https://t.co/u1anB2c4xd"*

*[[2]][1] "Tesla: First Model Y deliveries in the UK https://t.co/BRs4mmmOGL"*

*[[3]][1] "Tesla: RT @elonmusk: Tesla Model S &amp; X Plaid are so fun!"*

*[[4]][1] "Tesla: In late 2021, we started building Model Ys at Giga Texas. After final certification of Austin-made Model Y, we plan… https://t.co/VFEsqKSiQX"*

*[[5]][1] "Tesla: First deliveries of new Model X https://t.co/WfccyUXcFD"*

*[[6]][1] "Tesla: NYC as seen through Model Y glass roof\n\n📸: @Virus\_City\_ https://t.co/qY8TODtIQC"*

*[[7]][1] "Tesla: Model Y arrives in Europe https://t.co/nh1noiyMeR"*

*[[8]][1] "Tesla: Model S Plaid sets official Nürburgring EV world record!"*

*[[9]][1] "Tesla: RT @elonmusk: Tesla Model S Plaid just set official world speed record for a production electric car at Nurburgring. Completely unmodified,…"*

*[[10]][1] "Tesla: RT @DurvidImel: What it feels like to ride in a @Tesla Model S Plaid. Thanks @MKBHD for making all the blood in my body rush straight into…"*

*[[11]][1] "Tesla: RT @elonmusk: Review of Model S Plaid by Dan Neil\nhttps://t.co/JY45vn8jNN"*

*[[12]][1] "Tesla: To activate, tap Software → long press model name → type holidays or ModelXMas"*

*[[13]][1] "Tesla: All Model X can dance\n\n📸 @thechristina99 @michellehellman https://t.co/Qjh8qLJzwK"*

*[[14]][1] "Tesla: Playing Cyberpunk in Model S Plaid https://t.co/y9Ev4F6EaU"*

*[[15]][1] "Tesla: RT @DMC\_Ryan: Model S Plaid launch arguably felt even more intense from the backseat when I got a second test ride. I just laughed; I could…"*

*[[16]][1] "Tesla: RT @elonmusk: Order Tesla Model S Plaid at https://t.co/LuUhICqQEI"*

*[[17]][1] "Tesla: Model S Plaid delivery event at our Fremont factory will be streamed live on June 10, 7pm Pacific https://t.co/V7c77ySFti"*

*[[18]][1] "Tesla: RT @elonmusk: Model S goes to Plaid speed this week"*

*[[19]][1] "Tesla: Model 3 deliveries in Europe https://t.co/fiBNp2a6pG"*

*[[20]][1] "Tesla: Backseat views from Model Y\n\n📸 by loghi007 on Instagram https://t.co/4ss4Hns7fn"*

*[[21]][1] "Tesla: Model 3 earns Top Safety+ award from IIHS for third year in a row https://t.co/rr0QtV20al"*

*[[22]][1] "Tesla: RT @elonmusk: Plaid Model S ships next month https://t.co/HFUPTnQiPB"*

*[[23]][1] "Tesla: Plaid Model X has the quickest acceleration of any SUV by far, at 2.5 seconds 0-60"*

*[[24]][1] "Tesla: Plaid Model S is the first production car to achieve 0-60 in under 2 seconds"*

*[[25]][1] "Tesla: New Model S &amp; Model X just launched\n\nDetails https://t.co/7Ol1BvJoj8"*

*[[26]][1] "Tesla: Also, Model 3 frunks https://t.co/Up848EP3mk"*

*[[27]][1] "Tesla: Stamping Model Y body parts at our Fremont factory https://t.co/qbLrYowJRC"*

*[[28]][1] "Tesla: 欢迎! Model Y deliveries in China have officially begun 🚘🇨🇳 https://t.co/fG5aax1k2b"*

*[[29]][1] "Tesla: Model Y awarded 5-star safety rating in every category!\nhttps://t.co/dgbQWofpjf"*

*[[30]][1] "Tesla: The 7-seater option is now available for order on Model Y https://t.co/XLu84G3GJc"*

*[[31]][1] "Tesla: Model Y Standard Range is now available to order for $41,990\n \n- 244 mile range (EPA est.)\n- Rear-wheel drive\n- 0-6… https://t.co/FI1VWXAxQX"*

*[[32]][1] "Tesla: In-car footage from Sunday’s Plaid Model S run at Laguna Seca https://t.co/rwQDBFCWoH"*

*[[33]][1] "Tesla: Big updates in 🇨🇦 \n\n– Trans-Canada Supercharging ✅ \n– Model Y deliveries ✅ \n– Leasing for Model Y &amp; Model 3 🎯\nhttps://t.co/gyFBEIsjSP"*

*[[34]][1] "Tesla: Looking up in Model Y https://t.co/iAfpJ491Zz"*

*[[35]][1] "Tesla: Model Y deliveries begin!\n\nhttps://t.co/ZhuiM5MTOf https://t.co/3gX6MBPmhp"*

*[[36]][1] "Tesla: Track Mode V2 for Model 3 Performance introduces more advanced handling customization and lap data recording."*

*[[37]][1] "Tesla: Model 3 recently earned the 2020 IIHS Top Safety Pick+ Award. Winning cars are selected on their ability to protect… https://t.co/bQ9gbGlbJo"*

*[[38]][1] "Tesla: Model X just earned 5 stars from Euro NCAP in safety testing"*

*[[39]][1] "Tesla: Do you want a sneak peek of our crash lab?\n\nOf course you do. \n\nIt’s one of the reasons why Model 3 is among the sa… https://t.co/BoFyISHNRo"*

*[[40]][1] "Tesla: Until then, auf Wiedersehen Germany 👋\n\nHere’s some of our initial Model S Plaid data to keep you buzzing until we r… https://t.co/0JDcfznVCa"*

*[[41]][1] "Tesla: Data from our track tests indicates that Model S Plaid can achieve 7:20 at the Nürburgring. \n\nWith some improvement… https://t.co/r84mF9sScL"*

*[[42]][1] "Tesla: Model 3 has earned the IIHS 2019 Top Safety Pick+ Award, earning top scores across all eight tests covering crashwo… https://t.co/SUOQV0mTvY"*

*[[43]][1] "Tesla: Our Model X towed a plane once and now we're in the 2020 Guinness World Records book.\n\nSo…how's your day going? https://t.co/MTfi6Wyh7y"*

Peržiūrėję šias žinutes, matosi akivaizdus interpretuojamas klasteris – klasteris, kurio žinutėse yra kalbama apie Tesla kompanijos automobilius (vadinamus model). Sumažinę klasterių kiekį taip pat išvengėme konkrečių modelių išskyrimo į atskirus klasterius.

Klasteris 4:

*[[1]][1] "Tesla: Thank you to everyone who came out to celebrate the first Giga Texas deliveries and party with us. Yeehaw! https://t.co/w4LJR79l5b"*

*[[2]][1] "Tesla: Giga Texas y’all https://t.co/Mrsb7jRyhr"*

*[[3]][1] "Tesla: Flying Through Giga Berlin"*

*[[4]][1] "Tesla: Come join the Giga Berlin team! https://t.co/p1PVbG8c6x https://t.co/AyBsX36lNN"*

*[[5]][1] "Tesla: First Model Y deliveries at Giga Berlin https://t.co/JCneGrBJRA"*

*[[6]][1] "Tesla: 🎨 at Giga Berlin https://t.co/ojIvG9cg4F"*

*[[7]][1] "Tesla: Giga Berlin darf eröffnen 🇩🇪🚘 \n\nCome join the team! https://t.co/OdNeEGt5eB"*

*[[8]][1] "Tesla: We started equipment testing at Giga Berlin through the vehicle production process in late 2021 https://t.co/lIcDY8BPJB"*

*[[9]][1] "Tesla: Giga Shanghai continues to be our main export hub https://t.co/a9QB3oCmvv"*

*[[10]][1] "Tesla: Graffiti art at Giga Berlin https://t.co/aOj4kZNcbU"*

*[[11]][1] "Tesla: Giga Berlin-Brandenburg County Fair https://t.co/s25GaYuujl"*

*[[12]][1] "Tesla: RT @elonmusk: Giga Berlin-Brandenburg county fair &amp; factory tour on 9 October!"*

*[[13]][1] "Tesla: Giga Texas: 6 months ago vs this week https://t.co/CAgmIJ3wKQ"*

*[[14]][1] "Tesla: If you want to help cover Giga Berlin in awesome graffiti art, send us your work at GigaBerlinArt@Tesla.com"*

*[[15]][1] "Tesla: Cybertruck at Giga Texas https://t.co/c1RuektPnN"*

*[[16]][1] "Tesla: Winter at Giga Texas https://t.co/eqL8lPICKe"*

*[[17]][1] "Tesla: Winter at Giga Berlin https://t.co/ZyUwqWWY7S"*

*[[18]][1] "Tesla: RT @elonmusk: Giga Berlin progress https://t.co/ekpG5qcbUi"*

*[[19]][1] "Tesla: Come work on cell / battery production at Giga Texas &amp; Giga Berlin! https://t.co/PhsI5bb0uD https://t.co/jEUatMLDmE"*

Šiame klasteryje turime žinutes, kur pagrinde kalbame tik apie Gigafactory gamyklas (nėra užsimenama apie saulės kolektorius ar automobilius pvz). Turime daugiausiai paminėtų “Giga Berlin” gamyklų Berlyne, tačiau taip pat rasime mažiau paminėtų ir “Giga Texas” arba “Giga Shanghai”. Klasteris interpretuojamas irgi neblogai, kadangi atskiriame visas žinutes kalbančias tik apie šias gamyklas.

Į pirmąjį klasterį įtraukėme daugiausiai žinučių, kuriu nepavyko suklasterizuoti interpretuojamai. Šių visų žinučių nepateiksiu, vietoj to pateiksiu kelis pavyzdžius:

*[[1]][1] "Tesla: Fun on tap 🎢\n\n🎥 @saradietschy https://t.co/o0lhc1Webd https://t.co/BStyuW1BZF"*

*[[2]][1] "Tesla: In 2021, we received almost 3 million job applications. \n\nJoin us! https://t.co/wOIwB0p5rf https://t.co/Z5MJo9khrt"*

*[[17]][1] "Tesla: RT @elonmusk: Show starts in ~20 mins"*

*[[31]][1] "Tesla: View a breakdown of your charging history and gas savings in the Tesla app https://t.co/9zPCjBZjoQ"*

*[[35]][1] "Tesla: RT @elonmusk: Tesla merch buyable with Dogecoin"*

*[[40]][1] "Tesla: Merry Christmas to all! 🎄 https://t.co/QhOMe02pxy"*

Ir t.t.

Peržvelgus šio klasterio žinutes pasimato, jog sukrito visos šiukšlės. Padidinę klasterių skaičių žinučių šiek tiek pašalintume iš šio klasterio, tačiau reikšmingų žinučių klasteryje jau būna tik 1-3, kas nėra labai intepretuojama. Šios žinutės yra atsitiktinės, atrašymai į kitas žinutes, nuotraukos, kelių žodžių žinutės apie nieką. Šias žinutes būtų gana sunku suklasterizuoti ir žmogui (turint tik šį 300 žinučių rinkinį). Padidinę žinučių kiekį, manau pavyktų atrasti mažesnių specifinių klasterių, tačiau tuo pačiu įtrauktume ir daugiau šiukšlių.

Todėl atrodo, jog iš šios kompanijos puslapio prieš tai aptarti 3 klasteriai yra geriausiai interpretuojami atsižvelgus į kompanijos veiklą.

Matematiniu požiūriu sudaryti klasteriai irgi atrodo neblogai (vidutinis silueto plotis >0.20 visuose klasteriuose).

**PRIEDAS 1. R kodas**

library(twitteR)

library(ROAuth)

library(NLP)

library(Rcpp)

library(tm)

library(SnowballC)

library(fpc)

library(RColorBrewer)

library(wordcloud)

library(ggplot2)

library(tidyverse)

library(rtweet)

# RETRIEVING tweets from Twitter

# consumer\_key <- "tmrCmRkUIxnmL3lpD76dbXnVV" #API Key

# consumer\_secret <- "nDPRBbaUiR1aTK2NXqO7I44HGr55Zd03VY1jSlpL9yLz3uK6sU" # API Key secret

# access\_token <- "738658682653790209-TcBVcp8BHbNQ05sahnMMCaqiW3Asmp5"

# access\_token\_secret <- "PDD3hiIeXSlmlSApzGty9IO9Q99rwTpeUYAEznnn88eCj"

# setup\_twitter\_oauth(consumer\_key, consumer\_secret, access\_token, access\_token\_secret)

# rdmTweets <- userTimeline("tesla", includeRts=TRUE, n = 300)

# saveRDS(rdmTweets, file="rdmTweets.RData")

# RETRIEVING tweets from file

rdmTweets <- readRDS("rdmTweets.RData")

(length(rdmTweets))

# convert tweets to a data frame

df <- twListToDF(rdmTweets)

dim(df)

# Sudaromas dokumentu rinkinys (CORPUS) -->

myCorpus <- Corpus(VectorSource(df$text))

# Check the tweets

for (i in 1:5) {

cat(paste("[[", i, "]] ", sep=""))

writeLines(strwrap(myCorpus[[i]], width=73))

}

# We need to remove:

# unicode emojis (<U+0000>),

# mentions (@??????),

# URLS (https://...)

# retweet start (RT @?????)

# symbols??? (&amp;) - 89 row

# escape characters (\n)

# hashtags (#)

clean\_tweets <- function(x) {

x %>%

# Remove URLs

str\_remove\_all(" ?(f|ht)(tp)(s?)(://)(.\*)[.|/](.\*)") %>%

# Remove mentions e.g. "@my\_account"

str\_remove\_all("@[[:alnum:]\_]{4,}") %>%

# Remove hashtags

str\_remove\_all("#[[:alnum:]\_]+") %>%

# Replace "&" character reference with "and"

str\_replace\_all("&amp;", "and") %>%

# Remove puntucation, using a standard character class

str\_remove\_all("[[:punct:]]") %>%

# Remove "RT: " from beginning of retweets

str\_remove\_all("^RT:? ") %>%

# Replace any newline characters with a space

str\_replace\_all("\\\n", " ") %>%

# Make everything lowercase

str\_to\_lower() %>%

# Remove <U+???> instances

str\_replace\_all("[<].\*[>]", "") %>%

# Remove non-ASCII letters (including emojis)

str\_replace\_all("[^\x01-\x7F]", "") %>%

# Remove non-alphanumeric characters

str\_replace\_all("[^[:alnum:]]", " ") %>%

# Remove numbers

str\_replace\_all("[[:digit:]]+", " ") %>%

# Remove 1-2 letter words

str\_replace\_all(" \*\\b[[:alpha:]]{1,2}\\b \*", " ") %>%

# Replace solar roof with solar panel

str\_replace\_all("solar roof", "solar panel") %>%

# Remove any trailing whitespace around the text

str\_trim("both")

}

myCorpus <- tm\_map(myCorpus, content\_transformer(clean\_tweets))

# Remove mystopwords from corpus

myStopwords <- c(stopwords("en"), "tesla", "just", "one", "every", "deliveries", "production")

# Remove mystopwords from corpus

myCorpus <- tm\_map(myCorpus, removeWords, myStopwords)

myCorpusCopy <- myCorpus

myCorpus2<-tm\_map(myCorpus, stemDocument)

stemCompletion2 <- function(x, dictionary) {

x <- unlist(strsplit(as.character(x), " "))

# Unexpectedly, stemCompletion completes an empty string to

# a word in dictionary. Remove empty string to avoid above issue.

x <- x[x != ""]

x <- stemCompletion(x, dictionary=dictionary)

x <- paste(x, sep="", collapse=" ")

PlainTextDocument(stripWhitespace(x))

}

myCorpus3 <- lapply(myCorpus, stemCompletion2, dictionary=myCorpusCopy)

myCorpus4<-myCorpus

for (i in 1:length(myCorpus)) {

myCorpus4[[i]] <-(myCorpus3[[i]]$content)

}

# Check the orginial tweets and steemed tweets

for (i in 1:300) {

cat(paste("ORIGINAL", "[[", i, "]] ", sep=""))

writeLines(strwrap(myCorpus[[i]], width=73))

cat(paste("STEMMED", "[[", i, "]] ", sep=""))

writeLines(strwrap(myCorpus4[[i]], width=73))

}

tdm <- TermDocumentMatrix(myCorpus4, control=list(wordLengths=c(1,Inf)))

tdm

m <- as.matrix(tdm)

# calculate the frequency of words and sort it descendingly by frequency

wordFreq <- sort(rowSums(m), decreasing=TRUE)

# colors

pal <- brewer.pal(9, "BuGn")

pal <- pal[-(1:4)]

# word cloud

set.seed(375) # to make it reproducible

grayLevels <- gray( (wordFreq+10) / (max(wordFreq)+10) )

wordcloud(words=names(wordFreq), freq=wordFreq, min.freq=5,

random.order=F, colors=pal)

# remove sparse terms

tdm2 <- removeSparseTerms(tdm, sparse=0.975) #0.975

# Kuo sparse arciau vieneto tuo daugiau zodziu lieka matricoje (didesne dimensija).

tdm2

m2 <- as.matrix(tdm2)

# cluster terms

distMatrix <- dist(scale(m2))

fit <- hclust(distMatrix, method="ward.D")

plot(fit)

# cut tree into 4 clusters

rect.hclust(fit, k=4)

(groups <- cutree(fit, k=4))

# transpose the matrix to cluster documents (tweets)

m3 <- t(m2)

# set a fixed random seed

set.seed(122)

# k-means clustering of tweets

k<- 4

kmeansResult <- kmeans(m3, k)

# cluster centers

round(kmeansResult$centers, digits=3)

#

# To make it easy to fnd what the clusters are about,

# we then check the top three words in every

# cluster.

for (i in 1:k) {

cat(paste("cluster ", i, ": ", sep=""))

s <- sort(kmeansResult$centers[i,], decreasing=T)

cat(names(s)[1:5], "\n")

# print the tweets of every cluster

# print(rdmTweets[which(kmeansResult$cluster==i)])

}

# transpose the matrix to cluster documents (tweets)

m3 <- t(m2)

pamResult <- pamk(m3, metric="manhattan")

# number of clusters identified

(k <- pamResult$nc)

pamResult <- pamResult$pamobject

# print cluster medoids

for (i in 1:k) {

cat(paste("cluster", i, ": "))

cat(colnames(pamResult$medoids)[which(pamResult$medoids[i,]==1)], "\n")

# print tweets in cluster i

# print(rdmTweets[pamResult$clustering==i])

}

# partitioning around medoids with estimation of number of clusters

pamResult <- pamk(m3, krange=4, metric="manhattan")

# number of clusters identified

(k<-pamResult$nc)

pamResult <- pamResult$pamobject

# print cluster medoids

for (i in 1:k) {

cat(paste("cluster", i, ": "))

cat(colnames(pamResult$medoids)[which(pamResult$medoids[i,]==1)], "\n")

# print tweets in cluster i

print(rdmTweets[pamResult$clustering==i])

}

# set layout to two graphs per page matrix 2x1

# layout(matrix(c(1,2),2,1)) # set to two graphs per page

# plot clustering result

plot(pamResult, color=F, labels=4, lines=0, cex=.8, col.clus=1,

col.p=pamResult$clustering)

# change back to one graph per page

layout(matrix(1))