Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Ökoloogia ja Maateaduste instituut
Geograafia osakond

Bakalaureusetöö geoinformaatikas

Andmekaeve Twitterist 2018. aasta Eurovisiooni näitel

Patrick Joan Thomson

Juhendaja: Evelyn Uuemaa

Kaitsmisele lubatud:
Juhendaja:
Osakonna juhataja:

Tartu 2018

Andmekaeve Twitterist 2018. aasta Eurovisiooni näitel

Töö uurib andmekaevet Twitteri sotsiaalmeediakeskkonnast. Teoreetilises osas tutvustatakse sotsiaalmeediaga kaardistamise võimalusi. Metoodikas käsitletakse andmekogumis- ning andmetöötlusprotsessi ning selgitatakse, kuidas eraldatakse geo-andmetega postitused teistest ning hiljem analüüsitakse neid. Esimene hüpotees, geo-andmetega säutsude osakaal on vähemalt 5%, ei leidnud kinnitust. Teine hüpotees, säutsude hulk tunnis on andmekogumisperioodi jooksul kõige suurem sündmuse ajal, leidis kinnitust. Kolmas hüpotees, mida populaarsem on sündmus riigis, seda rohkem postitatakse sellest riigist säutse, ei leidnud kinnitust.

Märksõnad: Andmekaeve, Twitter, Eurovisioon
CERCS: PS10 – Füüsikaline geograafia, geomorfoloogia, mullateadus, kartograafia, klimatoloogia
S230 – Sotsiaalne geograafia

Data mining from Twitter on the example of Eurovision Song Contest 2018

This paper analyzes data mining from Twitter. Theoretical background introduces social media mapping possibilities. The methodology part deals with data collection and processing and how geo-tagged tweets are separated from the datasets and later analyzed. The first hypothesis, geo-tagged tweets make up atleast 5% of all the collected tweets, was not confirmed. The second hypothesis, the highest number of tweets per hour during the data collection period is greatest at the time of the event, was confirmed. Third hypothesis, the more tweets are posted the more popular is the event in the country, was not confirmed.

Keywords: Data mining, Twitter, Eurovision
CERCS: PS10 – Physical geography, geomorphology, cartography, climatology
S230 – Social geography

Sisukord

Sissejuhatus 4

1. Teoreetiline alus 5

1.1. Ülevaade sotsiaalmeediaga kaardistamisest 5

1.2. Mis on Twitter? 6

1.3. Säutsu iseloomustus 7

2. Metoodika 10

2.1. Skripti väljatöötamine 10

2.2. Andmete andmebaasi viimine 11

2.3. Andmebaasis päringutega analüüs 11

2.4. Sündmuse valik ja andmete kogumine 12

3. Tulemused ja arutelu 13

Kokkuvõte 18

Summary 19

Tänusõnad 20

Kasutatud kirjandus 21

Sissejuhatus

Suurenenud nutiseadmete ja sotsiaalmeedia kasutamine on endaga kaasa toonud trendi, kus inimesed jagavad oma mõtteid või peavad vestlusi, mida varem peeti sõprade ringis, avalikult. Sellest tulenevalt toodetakse igapäevaselt suurel hulgal andmeid, millel on küljes ka ruumiandmed. Nende andmete kasutamine ruumianalüüsis jätab vahele kulukama ruumiandmete kogumise traditsioonilisel viisil ning võimaldab koguda andmeid ka sündmuste kohta, kus oleks muidu selliseid andmeid raske koguda. Uurimuses keskendutakse sotsiaalmeediaplatvormile Twitter ning uuritakse ruumianalüüsi võimalikkust suursündmuse ajal kogutud andmetest.

Töö eesmärk on uurida 2018. aastal toimunud Eurovisiooni lauluvõistluse näitel andmekaevet Twitterist. Töös vaadeldakse andmete üldist jaotust ning analüüsitakse ruumiandmetega säutse. Püstitati kolm hüpoteesi.

Hüpoteesid:

1) geo-andmetega säutsude osakaal on vähemalt 5%;
2) säutsude hulk tunnis on andmekogumisperioodi jooksul kõige suurem sündmuse ajal;
3) mida populaarsem on sündmus riigis, seda rohkem postitatakse sellest riigist säutse.

1. Teoreetiline alus

1.1 Ülevaade sotsiaalmeediaga kaardistamisest

Varasemalt väikses tuttavate ringkonnas jagatud arvamused on nüüd tänu sotsiaalmeediaplatvormidele nähtavad kõigile, kel vaid interneti ligipääs. Lisaks laienenud publikule on võimalik enda arvamust avaldada reaalajas ehk sündmuste toimumisega samaaegselt. Kommentaariumides, mikroblogides ja mujal on kasutajatel võimalus oma postitusele ruumiandmed lisada. Eriti lihtne ja rohkelt kasutatud on see variant nutitelefonidega. Geo-andmetega postitused sisaldavad kas füüsilisi koordinaate või ümbritseva ala kasti (inglise keeles *bounding box*) koordinaate. Postitused, millel on ruumiline komponent, võimaldavad andmete analüüsimist mitte ainult sisu vaid ka ruumilise jaotuse põhjal. Nutitelefonide populaarsuse kasvades tõuseb ka igapäevaselt genereeritud andmete hulk, mida on võimalik analüüsida.

Ruumianalüüs on geograafilise sisuga ülesannete lahendamine geoinfosüsteemi (GIS)- ja kaarditarkvaras matemaatiliste algoritmide abil. Ruumianalüüsi kombineerimisel sisuanalüüsiga on võimalik ennustada näiteks käimasolevate sündmuste tulemust. Brexiti toimumise ajal oli võimalik reaalajas jälgida sündmuse globaalset mõju ning ennustada säutsude sisu ning asukoha põhjal hääletuse tulemit (Agarwal jt, 2017).

Sotsiaalmeediast saadavad andmed on olulisel kohal ka katastroofiolukorras inimeste olukorra kindlaks tegemisel. Platvormid nagu Twitter ja Facebook aitavad kriisiolukorras kaardistada inimeste asukohta ja olukorda tänu n-ö kriisivastustele (inglise keeles *crisis response*), kus inimene saab märkida, kas ta on ohus või mitte (Lindsay, 2011). See meetod on juba neljas kõige parem infoallikas inimeste olukorra kohta kriisisituatsioonides (Lindsay, 2011). 34-st OECD riigist on Twitteri platvorm populaarne 23-s riigis, Facebook aga 21-s riigis (Kim ja Hastak, 2018). Twitterist on saadud andmeid maavärinate ja taifuunide tuvastamiseks ja nende asukoha määramiseks (Sakaki jt, 2010). Arthur jt (2018) leidsid Suurbritannia näitel, et üleujutuste tuvastamine ja asukoha määramine on võimalik tänu sotsiaalmeedia geo-andmetega postitustele. See meetod on potentsiaalselt kiirem alternatiiv erinevate meediaväljaannete läbi sirvimisele, et leida informatsiooni üleujutuste toimumise ja asukoha kohta. Küll aga tuleb meetodisse suhtuda skeptiliselt, sest hetkel on probleem geo-andmetega postituste vähesusega.

Sotsiaalmeediaandmeid on kasutatud inimtegevuse aktiivsuse mõõtmiseks paikades, kus tradistiooniline teabe kogumine on kallis või ebaefektiivne, kuid oluline näiteks turundusele või juhtimisele. Uuringust selgus, et rahvusparkide populaarsus Soomes ja Lõuna-Aafrika

Vabariigis on tugevalt seotud nende kajastamisega sotsiaalmeedias, eriti Instagrami platvormil, kus põhifookus on piltide jagamisel. Antud andmetega on võimalik saada hea ülevaade rahvusparkide külastatavusest ja selle mustritest ning seeläbi korraldada rahvuspargi tööd efektiivsemalt. (Tenkanen jt, 2017)

Brasiilias kasutatakse sääskedega levivate haiguste, mille hulka kuulub ka nt Zika viirus, ennetamiseks ja nende vastu võitlemiseks VazaDengue platvormi. VazaDengue programmi eesmärk on kaardistada erinevad haigusjuhud ning ka potentsiaalsed haiguse levikualad. Populatsiooni laiemaks kaasamiseks jälgib programm pidevalt ka Twitteri infovoogu, et robustse sisuanalüüsi kaudu säutsud kategooriatesse jagada. Määratud märksõnade esinemisel toimub programmi sissekanne, mis on omakorda platvormi väärtuslik sisend. (Sousa jt, 2018)

Los Angelese linna näitel uuriti, kuidas muuta Twitteri *check-in* andmed erinevates linnaosades toimunud avariide ruumianalüüsi osaks. *Check-in* andmed aitavad kokkupõrgete analüüsi täpsemaks muuta. Üks suur avariide riskifaktor on rahvastiku tihedus. Lähedal asuvad lokaalid ning tihedad inimeste põhitegevused, mille alla käivad näiteks söömine, õppimine, puhketegevus, võivad soodustada avariiolukordade tekkimist. Ühendades Twitteri andmed tavapärase ruumianalüüsiga, on võimalik täpsemalt ennustada avariide teket ning seeläbi ka neid ära hoida. (Bao jt, 2017)

1.2 Mis on Twitter?

Twitter on sotsiaalmeediaplatvorm, mis võimaldab inimestel üle maailma omavahel lühisõnumite abil suhelda. Twitter loodi 12 aastat tagasi 2006. aasta maikuus. Twitter on üles ehitatud mikroblogidele, mille ühe postituse maht on piiratud. Kasutajad saavad enda arvamust või infot jagada säutsude (inglise k. *tweet*) abil. Säutsudel on limiteeriv mahupiirang, mis sunnib enda arvamuse või mõtted tavapärasest enam koondama. Algselt oli ühe säutsu mahupiirang 140 tähemärki, aga 2017. aasta novembris tõsteti see kõikides keeltes peale jaapani, korea ja hiina keele 270 märgini, muutes Twitteri keskkonna kasutajasõbralikumaks (Rosen, 2017).

Teiste kasutajate säutse on võimalik leida otsingumootori või teemaviidete kaudu. Postitusi klassifitseeritakse teemaviidete (inglise k. *hashtag*) järgi, milleks on sõnad või kokkukirjutatud fraasid, mille ees on #-märk. Kasutajaid on võimalik Twitteri keskkonnas ka jälgida ehk postitused ilmuvad jälgija infovoogu ning võimaldab saada teavitusi kui jälgitav on midagi postitanud. Vaikimisi on kõikide kasutajate säutsud avalikud, ent nende avalikkust on võimalik ka piirata muutes enda postitused privaatseks ehk nähtavaks ainult enda

jälgijatele. Kui kasutaja on muudetud privaatseks, ei ole võimalik avalikult säutse teemaviidete ega otsingumootori kaudu leida.

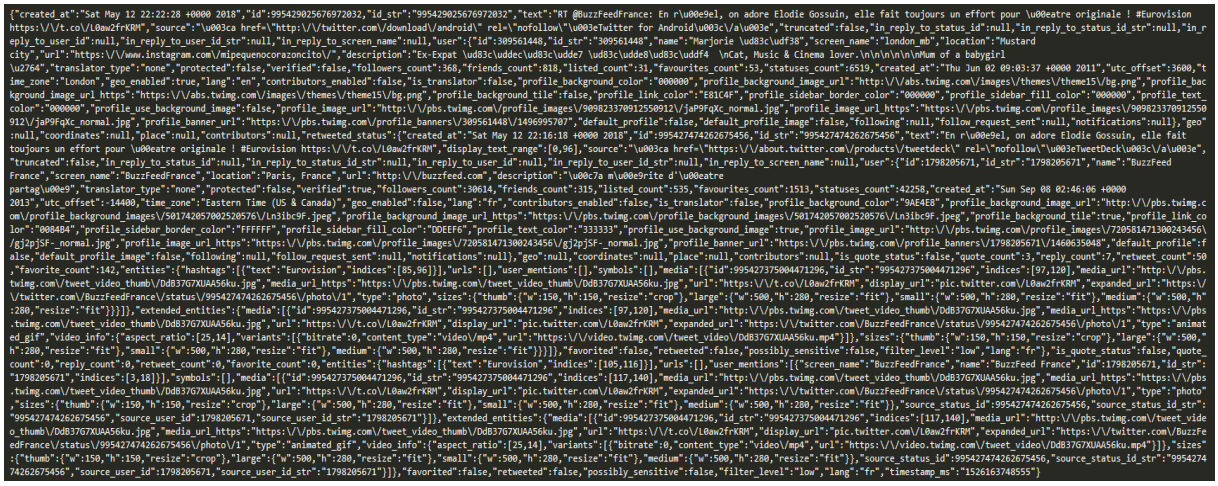
Teistele kasutajatele viitamiseks kasutatakse @-märki, millele järgneb kasutajanimi. Kui soovitakse teiste postitusele vastata või teisi oma postituses märkida kasutatakse samuti @-märki. Sellisel juhul tuleb postituses märgitud kasutajale teavitus. Lisaks on võimalik ka teiste kasutajate postitusi taaspostitada, mille lühendiks on RT (inglise k. *retweet*). Teiste kasutajate postitusi taaspostitades ilmuvad need ka taaspostitaja jälgijate infovoogu.

2018. aasta jaanuari seisuga on keskkonnas 330 miljonit aktiivset kasutajat, kellest 88% kasutavad Twitterit läbi mobiilirakenduse. Aktiivseid igakuiseid kasutajaid on enim Ameerika Ühendriikides (72,3 mil), millele järgneb Jaapan (50,9 mil). Euroopas on Twitter populaarne Suurbritannias (18,6 mil), Hispaanias (8,3 mil) ja Prantsusmaal (7,6 mil).

1.3 Säutsu iseloomustus

Säutsud on Twitteri alusosakesed, mida tuntakse ka olekuvärskendustena (*inglise k status update*) ja neid võib nimetada ka Twitteri keskkonna postitusteks. Säutsu objektil on pikk nimekiri juur-taseme atribuutidest. Säutsu objektid võivad olla seotud mitme emaobjektiga (*inglise k parent object*) ning mitmel objektil võib olla tütarobjekte (*inglisek child object*), näiteks geo-andmetega säutsudel on tütarobjektid. Säutsu *Javascript Object Notation* (.json) vorm on segu juur-taseme ja tütarobjektidest, mida eristatakse. Säuts Twitteri keskkonnas on nähtav joonisel 1, sama säuts kogu enda atribuutidega lahtikirjutatult joonisel 2.

Joonis 1. Säuts Twitteri keskkonnas.


{"created_at":"Sat May 12 22:22:28 +0000 2018","id":995429025676972032,"id_str":"995429025676972032","text":"RT @BuzzFeedFrance: En r\u00e9el, on adore Elodie Gossuin, elle fait toujours un effort pour \u00eatre originale ! #Eurovision https:\/\/t.co\/L0aw2frKRM","source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":309561448,"id_str":"309561448","name":"Marjorie \ud83c\udf38","screen_name":"london_mb","location":"Mustard city","url":"https:\/\/www.instagram.com\/mipequenocorazoncito\/","description":"Ex-Expat \ud83c\uddec\ud83c\udde7 \ud83c\udde8\ud83c\uddf4 \nCat, Music & Cinema lover.\n\n\n\n\n\nMum of a babygirl \u2764","translator_type":"none","protected":false,"verified":false,"followers_count":368,"friends_count":818,"listed_count":31,"favourites_count":53,"statuses_count":6519,"created_at":"Thu Jun 02 09:03:37 +0000 2011","utc_offset":3600,"time_zone":"London","geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"000000","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme15\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme15\/bg.png","profile_background_tile":false,"profile_link_color":"E81C4F","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/909823370912550912\/jaP9FqXc_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/909823370912550912\/jaP9FqXc_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/309561448\/1496995707","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Sat May 12 22:16:18 +0000 2018","id":995427474262675456,"id_str":"995427474262675456","text":"En r\u00e9el, on adore Elodie Gossuin, elle fait toujours un effort pour \u00eatre originale ! #Eurovision https:\/\/t.co\/L0aw2frKRM","display_text_range":[0,96],"source":"\u003ca href=\"https:\/\/about.twitter.com\/products\/tweetdeck\" rel=\"nofollow\"\u003eTweetDeck\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1798205671,"id_str":"1798205671","name":"BuzzFeed France","screen_name":"BuzzFeedFrance","location":"Paris, France","url":"http:\/\/buzzfeed.com","description":"\u00c7a m\u00e9rite d'\u00eatre partag\u00e9","translator_type":"none","protected":false,"verified":true,"followers_count":30614,"friends_count":315,"listed_count":535,"favourites_count":1513,"statuses_count":42258,"created_at":"Sun Sep 08 02:46:06 +0000 2013","utc_offset":-14400,"time_zone":"Eastern Time (US & Canada)","geo_enabled":false,"lang":"fr","contributors_enabled":false,"is_translator":false,"profile_background_color":"9AE4E8","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_images\/501742057002520576\/Un3ibc9F.jpeg","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/501742057002520576\/Un3ibc9F.jpeg","profile_background_tile":true,"profile_link_color":"008484","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/720581471300243456\/gj2pjSF-_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/720581471300243456\/gj2pjSF-_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/1798205671\/1460635048","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"quote_count":3,"reply_count":7,"retweet_count":50,"favorite_count":142,"entities":{"hashtags":[{"text":"Eurovision","indices":[85,96]}],"urls":[],"user_mentions":[],"symbols":[],"media":[{"id":995427375004471296,"id_str":"995427375004471296","indices":[97,120],"media_url":"http:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","media_url_https":"https:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","url":"https:\/\/t.co\/L0aw2frKRM","display_url":"pic.twitter.com\/L0aw2frKRM","expanded_url":"https:\/\/twitter.com\/BuzzFeedFrance\/status\/995427474262675456\/photo\/1","type":"photo","sizes":{"thumb":{"w":150,"h":150,"resize":"crop"},"large":{"w":500,"h":280,"resize":"fit"},"small":{"w":500,"h":280,"resize":"fit"},"medium":{"w":500,"h":280,"resize":"fit"}}}]},"extended_entities":{"media":[{"id":995427375004471296,"id_str":"995427375004471296","indices":[97,120],"media_url":"http:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","media_url_https":"https:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","url":"https:\/\/t.co\/L0aw2frKRM","display_url":"pic.twitter.com\/L0aw2frKRM","expanded_url":"https:\/\/twitter.com\/BuzzFeedFrance\/status\/995427474262675456\/photo\/1","type":"animated_gif","video_info":{"aspect_ratio":[25,14],"variants":[{"bitrate":0,"content_type":"video\/mp4","url":"https:\/\/video.twimg.com\/tweet_video\/DdB37G7XUAA56ku.mp4"}]},"sizes":{"thumb":{"w":150,"h":150,"resize":"crop"},"large":{"w":500,"h":280,"resize":"fit"},"small":{"w":500,"h":280,"resize":"fit"},"medium":{"w":500,"h":280,"resize":"fit"}}}]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"fr"},"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{"text":"Eurovision","indices":[105,116]}],"urls":[],"user_mentions":[{"screen_name":"BuzzFeedFrance","name":"BuzzFeed France","id":1798205671,"id_str":"1798205671","indices":[3,18]}],"symbols":[],"media":[{"id":995427375004471296,"id_str":"995427375004471296","indices":[117,140],"media_url":"http:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","media_url_https":"https:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","url":"https:\/\/t.co\/L0aw2frKRM","display_url":"pic.twitter.com\/L0aw2frKRM","expanded_url":"https:\/\/twitter.com\/BuzzFeedFrance\/status\/995427474262675456\/photo\/1","type":"photo","sizes":{"thumb":{"w":150,"h":150,"resize":"crop"},"large":{"w":500,"h":280,"resize":"fit"},"small":{"w":500,"h":280,"resize":"fit"},"medium":{"w":500,"h":280,"resize":"fit"}},"source_status_id":995427474262675456,"source_status_id_str":"995427474262675456","source_user_id":1798205671,"source_user_id_str":"1798205671"}]},"extended_entities":{"media":[{"id":995427375004471296,"id_str":"995427375004471296","indices":[117,140],"media_url":"http:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","media_url_https":"https:\/\/pbs.twimg.com\/tweet_video_thumb\/DdB37G7XUAA56ku.jpg","url":"https:\/\/t.co\/L0aw2frKRM","display_url":"pic.twitter.com\/L0aw2frKRM","expanded_url":"https:\/\/twitter.com\/BuzzFeedFrance\/status\/995427474262675456\/photo\/1","type":"animated_gif","video_info":{"aspect_ratio":[25,14],"variants":[{"bitrate":0,"content_type":"video\/mp4","url":"https:\/\/video.twimg.com\/tweet_video\/DdB37G7XUAA56ku.mp4"}]},"sizes":{"thumb":{"w":150,"h":150,"resize":"crop"},"large":{"w":500,"h":280,"resize":"fit"},"small":{"w":500,"h":280,"resize":"fit"},"medium":{"w":500,"h":280,"resize":"fit"}},"source_status_id":995427474262675456,"source_status_id_str":"995427474262675456","source_user_id":1798205671,"source_user_id_str":"1798205671"}]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"fr","timestamp_ms":"1526163748555"}

Joonis 2. Säutsu objekt tekstina.

Säutsu atribuudid:

1) created_at – UTC aeg (maailmaaeg), millal säuts loodi;

2) id – unikaalne säutsu id, mis on suurem kui 53 bitti;

3) id_str – säutsu id väljendatud sõnena;

4) text – säutsu sisu, mis on vaikimisi UTF-8 formaadis;

5) source – platvorm, mille kaudu säuts tehtud on, internetibrauseris koostatud säutsudel on vaikimisi valik *web*;

6) truncated – kahendväärtus, mis näitab, kas tekstiparameetri väärtus lühendati, näiteks taaspostitusega, mille algupärase teksti pikkuse piir on 140 tähemärki;

7) in_reply_to_status_id – kui tegemist on taaspostitusega, siis sisaldab väärtust, milleks on originaase säutsu id;

8) in_reply_to_status_str – originaalsäutsu id sõnena;

9) in_reply_to_user_id - taaspostituse puhul taaspostitatud kasutaja id väärtus;

10) in_reply_to_user_id_str – taaspostituse puhul taaspostitatud kasutaja id väärtus sõnena;

11) in_reply_to_screen_name – taaspostituse puhul taaspostitatud kasutaja kasutajanimi;

12) user – kasutaja objekt, mis sisaldab hulka teavet kasutajast;

13) coordinates – kui on säutsutud kasutades geolokatsiooni, siis jäävad väärtusesse pikkus ja laiuskraadid;

14) place – näitab, mis piirkonnaga säutsu seostatakse, võib esineda polügonina, mingi linna või asutusüksuse piirkonnas;

15) quoted_status_id – esineb juhul kui säutsus on tsitaat, number näitab tsiteeritud säutsu id’d;

16) quoted_status_id_str - esineb juhul kui säutsus on tsitaat; number näitab tsiteeritud säutsu id’d sõnena;

17) is_quote_status – näitab kahendväärtust ning kas säutsus on tsitaat;

18) quoted_status – näitab originaalse säutsu objekti, kui säuts on tsitaat;

19) retweet_status – näitab säutsu objekti, mida on taaspostitatud.

20) quote_count – näitab mitu korda on säutsi tsiteeritud.

21) reply_count – näitab, mitu korda on säutsule vastatud.

22) retweet_count – näitab, mitu korda on säutsi taaspostitatud;

23) favorite_count – näitab, mitu korda on teised kasutajad säutsu meeldivaks märkinud;

24) entities – sisu, mis ei ole säutsu tekst, nagu hüperlingid, teemaviited, kasutajate mainimised, meedia, sümbolid, küsitlused;

25) extended_entities – kui säutsus on 1-4 pilti, video või GIF, siis sisaldab metaandmete massiivi;

26) favorited – kahendväärtus, mis näitab, kas säutsu loonud kasutaja on seda meeldivaks märkinud;

27) retweeted – kahendväärtus, mis näitab, kas säutsu loonud kasutaja on seda taaspostitanud;

28) possibly_sensitive – kahendväärtus kui säuts sisaldab hüperlinki ning seetõttu viitab millelegi muule, kui säutsule endale;

29) filter_level – näitab maksimaalset parameetrit, mida kasutades säuts striimides esineb, nt filter_level medium väärtusega striimitakse säutse, mille filter levelid on none, low ja medium;

30) lang – automaattuvastuse poolt määratud säutsu keel;

31) matching_rules – esinevad filtreeritud keskkondades nagu Twitter Search ja PowerTrack, pakub ID-d ja *tag*-i, mis on seotud säutsule vastavale reeglistikuga, PowerTrackiga võib rohkem kui üks reegel sobida säutsuga.

Twitter API võib sisaldada ka säutsu lisaatribuute:

1) current_user_retweet - atribuut on nähtav siis, kui kasutaja taaspostitab oma postitust;

2) scopes - võtmeväärtuspaaridena esinev komplekt, mida kasutavad Twitteri poolt reklaamitud tooted;

3) withheld_copyright - kui väärtus on tõene, siis on säutsu sisu DMCA kaebuse tõttu eemaldatud;

4) withheld_in_countries - kui väärtus on tõene, siis näitab riikide kahetähelisi koode, kus sisu on DMCA tõttu eemaldatud;

5) withheld_scope - näitab, kas eemaldatud on postituse sisu või postituse teinud kasutaja.

2. Metoodika

2.1 Skripti väljatöötamine

Andmed koguti Pythoni programmeerimiskeeles kirjutatud skriptis, millel oli ühendus Twitteri rakendusliidesega. Skript on programmeerimiskeeles kirjutatud lähtekood, mis kompileeritakse ümber masinkoodiks (Github, 2018). Twitteri rakendusliides (inglise k *Application Programming Interface*) võimaldab programmiga teha automaatseid päringuid sotsiaalvõrgustiku andmetest. Selle jaoks peab registreerima oma rakenduse Twitteri lehel, mille kaudu saab autentimiseks vajalikud koodid. Andmete kogumise meetodi valikul olid olulised Twitteri-poolsed piirangud. Varem postitatud säutsude kogumisel on Twitteri rakendusliidese poolne piirang, millest tulenevalt saab 15 minutilise akna sees teha 15 päringut varasemalt postitatud andmete kohta, et vähendada serverite koormust. Statistiliselt töödeldavate andmete kogumiseks on see meetod liiga aeglane. Seega otsustati jooksvalt andmeid koguva skripti kasuks.

```
import tweepy
from tweepy import OAuthHandler
consumer_key =   'näide1'
consumer_secret = 'näide2'
access_token =    'näide3'
access_secret = 'näide4'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
```

Joonis 3. Twitteri rakendusliidesega ühendamise koodi osa.

Twitteriga ühenduse loomiseks kasutati teeki *Tweepy*, mis on välja töötatud Python programmeerimiskeelele, et saada ühendust rakendusliidesega. Teegist *Tweepy* kasutati kolme moodulit: *Stream*, millega sai ühenduse Twitteri infovoole; *Streamlistener*, millega sai jälgida enda valitud termineid infovoos; *OAuthHandler*, millega autenditi kasutatava rakenduse ühendus Twitteri rakendusliidese serveriga.

Skripti tööpõhimõte oli jooksvalt oodata kuni Twitteri infovoost tuleb läbi otsitav termin. Seejärel salvestatakse see .json faili uuele reale. Vea korral väljastas programm veateatekoodi. Iga säuts oli oma eraldi json objektis. Andmete salvestamiseks valiti .json failivorming, sest Twitteri rakendusliides väljastab andmeid selles formaadis. Antud formaadi eelis on faili sisu mõistetavus nii inimesele kui ka masinale.

2.2. Andmete andmebaasi viimine

Andmetöötluseks valiti PostgreSQL andmebaas, sest tegemist on vabavaralise andmebaasiga, mille jaoks on ka kohandatud PostGIS laiend, mis võimaldab ruumiandmete töötlemist andmebaasis. Lisaks on andmebaasis sisseehitatud väli json-andmetüübi jaoks. Eelnevalt kirjeldatud skriptiga kogutud andmete andmebaasi üle viimiseks kasutati teeki *psycopg2*, mis on Pythoni jaoks kohandatud ligipääs PostgreSQL andmebaasile ning võimaldab andmebaasis baastoiminguid teha ka otse läbi Pythoni skripti. Lisaks kasutati teeki *codecs*, mis võimalds Pythoni baaskäskude kodeerimist ja dekodeerimist (Github, 2018).

Andmete ülekandmisel PostgreSQL andmebaasi jäid *Unicode*, mis on rahvusvaheline standard tähtede kodeerimiseks, märgid oma algsele kujule. Päringute tegemisel tekitas probleeme *\u0000*, mis tähistab tühja kohta (‘’). Selle eemaldamiseks kasutati andmebaasis päringut, mis muutis json objekti sisu tekstiks, asendas *\u0000* tühja kohaga ning muutis väljatüübi tagasi json’iks.

2.3. Andmebaasis päringutega analüüs

Andmebaasis kasutati PostgreSQL sisseehitatud json-väljatüübist päringu tegemiseks kasutatavaid operaatoreid. Võrreldes tavapärase päringuga andmebaasis on json objektis sees võimalik veel päringuid teha. Geo-andmete eraldamine toimus kahes osas. Väljadel, millel olid koordinaatandmed, eraldati punktobjekt pikkus- ja laiuskraadide kaudu. Näide ühest punktist GeoJson formaadis: {"type": "Point", "coordinates": [4.72033, 50.87216]}.

Väljadel, kus oli märgitud ainult place-atribuut, oli asukoha eraldamiseks vaja kõigpealt välja saada märgitud ala ümbritsev kast (inglise k *bounding box*). Twitteri rakendusliidese poolt lisatud ümbritseva kasti geomeetria on vigane, sest seal on antud ainult nelja nurga koordinaadid. Geomeetriliselt korrektne polügon algab ja lõppeb oma alguspunktiga.

Näide säutsu ümbritseva kasti geoandmetest GeoJson formaadis: {"id": "c1430b24da8e9229", "url": "https://api.twitter.com/1.1/geo/id/c1430b24da8e9229.json", "place_type": "city", "name": "Lisbon", "full_name": "Lisbon, Portugal", "country_code": "PT", "country": "Portugal", "bounding_box": {"type": "Polygon", "coordinates": [[[-9.229826, 38.691375], [-9.229826, 38.795853], [-9.090164, 38.795853], [-9.090164, 38.691375]]]}, "attributes": {}}. Sellest formaadist punktandme kättesaamiseks kasutati PostGIS laienduse funktsioone, mis moodustasid punktidest joonobjekti ning ühendasid selle lõpu algusega, et tekitada polügoni, seejärel võeti tsentroid (joonis 4).

```
SELECT
  ST_Centroid(
    ST_MakePolygon(
      ST_MakeLine(
        b,
        ST_PointN(b,1)
      )
    )
  )
FROM (
  SELECT
    ST_LineMerge(
      ST_GeomFromGeoJson(
        (
          '{"type": "MultiLineString", "coordinates": ' || ((
            tweet->>'place'
          )::jsonb->'bounding_box'->'coordinates')::text || '}')
      )
    ) AS b from schema.table where tweet->>'place' is not null
) AS sq;
```

Joonis 4. *Place* atribuudist punkti kättesaamiseks kasutatud päring.

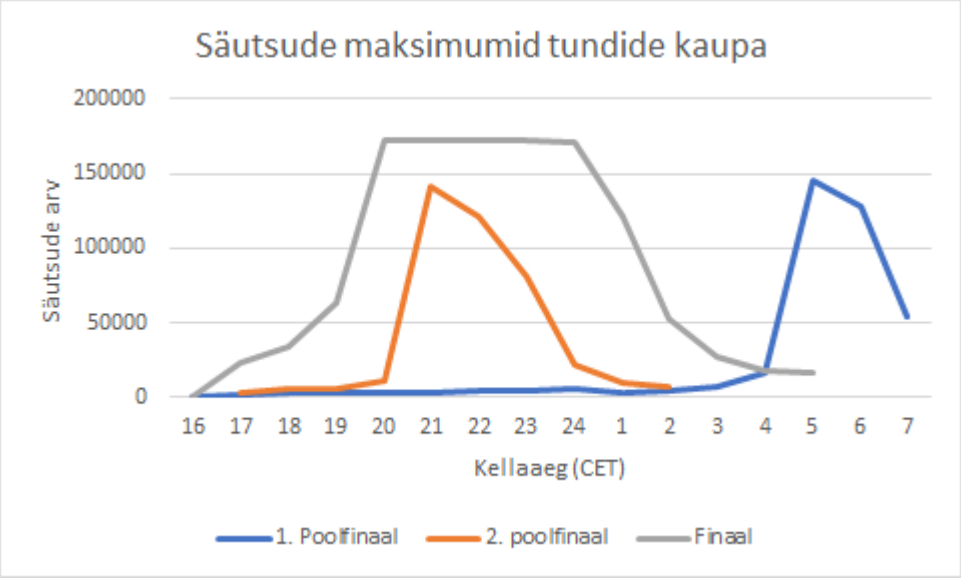
2.4. Sündmuse valik ja andmete kogumine

Andmekaeve jaoks valiti välja sündmus, mille kajastus Twitteri platvormil on eelduste kohaselt suur. Välja valiti 2018. aasta Eurovisiooni lauluvõistluse kolm peasündmust: esimene ja teine poolfinaal ning finaal. Seega koguti andmeid kolmes osas kolmel erineval päeval ühe nädala jooksul. Eurovisioon osutus valituks rahvusvahelisuse, pika ajaloo ja populaarsuse pärast. Selle aasta finaali jälgis Eestis ETV kanali vahendusel reaalajas keskmiselt 233 000 inimest, mis moodustab 19.3% nelja-aastastest ja vanematest Eesti elanikest. Eesti esineja etteaste ajal oli aga jälgimas korraga 300 000 inimest, mis moodustas 78.7% televaatajatest. Võrdluseks saab tuua, et 2018. aasta vaadatuim ETV saade on olnud 24. veebruari "Aktuaalne kaamera", millel oli keskmiselt 304 000 üle nelja-aastast vaatajat. (Eesti Rahvusringhääling, 2018)

Märksõnad, mida säutsudest otsiti oli “Eurovision” ning teemaviited #Eurovision ning #esc. Teemaviide #esc on lühend Eurovsion Song Contestist ehk eestikeelselt Eurovisiooni lauluvõistlus. Andmete kogumiseks valiti 22-25-tunnine ajavahemik. Skript pandi tööle 10 tundi enne võistluse algust ning peatati 10 tundi pärast vastava (pool)finaali lõppu. Ajapuhvri eesmärk oli koguda andmeid umbkaudu ööpäeva jooksul nii, et sündmus ise jääks puhvri sisse ning võtaks arvesse ka ajavahet toimumiskoha ja kaugemate osalejate vahel, milleks on vastavalt Portugal ja Austraalia. Samuti on võimalik niimoodi näha, milline on lauluvõistluse säutsude jaotus ajas.

3. Tulemused ja arutelu

Kolme päeva jooksul koguti kokku 2 109 557 säutsu, millest 400 608 esimese ja 417 281 teise poolfinaali ning 1 291 668 finaali ajal. Säutsude postitamise maksimumi poolfinaalide graafikujooned olid visuaalselt sarnased, nagu on nähtav joonisel 5. Ühel tunnil oli nähtav postituste lagi ning pärast seda langes tunnis postitatud säutsude arv kiirelt. Esimese poolfinaali puhul säutsuti enim alles kuus tundi pärast poolfinaali lõppu, teise poolfinaali puhul postitati enim võistluse keskel. Esimese poolfinaali säutsude ajalise jaotuse nihke võis põhjustada sündmuse algus, mistõttu polnud veel välja kujunenud üldiste postituste temaatika ning kasutajad jagasid oma emotsioone, mis tekkisid poolfinaali tulemustest. Finaali puhul oli säutsude voog kõige suurem. Umbkaudu viis tundi ehk terve võistluse aja postitati keskmiselt 170000 säutsu tunnis.



Joonis 5. Säutsude ajaline jaotus ööpäeva jooksul.

Esimese poolfinaali andmestikus oli 202 (0.05%) koordinaatandmetega ning 13 716 (3.42%) asukohaandmetega (inglise k *place attributes*) säutsu. Teise poolfinaali andmestikus oli 196 (0.05%) koordinaatandmetega säutsu ning 12 893 (3.09%) asukohaandmetega säutsu. Finaali andmestikus oli 407 (0.03%) koordinaatandmetega säutsu ning 22 776 (1.76%) asukohaandmetega säutsu.

Globaalsel tasandil populatsioonile üldistatava ruumianalüüsi läbi viimiseks ei ole võimalik veel piisavalt andmeid koguneda. Geo-andmetega säutsude osakaal kõikide säutsude seas oli keskeltläbi vaid 1.4%, niiet selleks, et koguda selle meetodiga piisavalt üldistatavaid geo-andmeid, peab koguma suures koguses säutse. 1.4% geoandmetega säutse on võrreldes teiste töödega vähe, tuues näiteks Boulton jt. (2016) uuringu, kus geoandmetega postitused

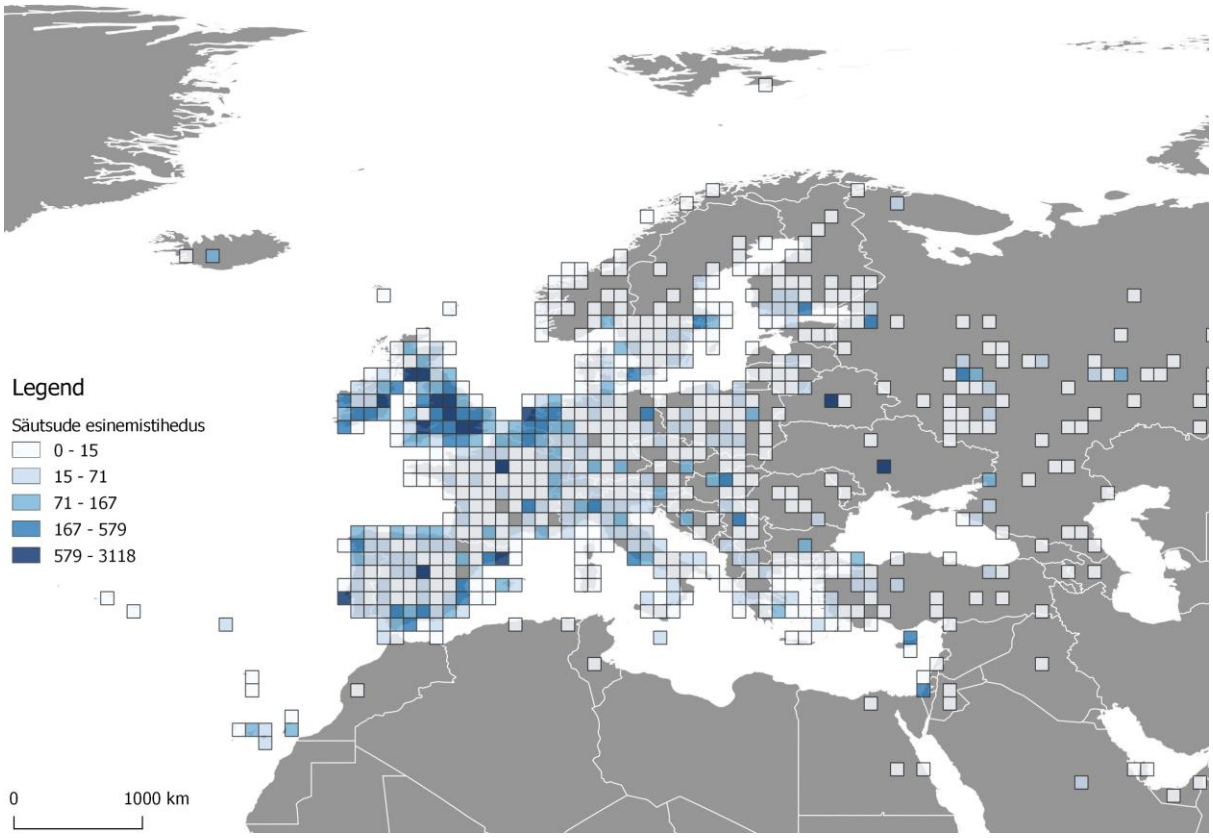
moodustasid 12% kogu andmestikust. Üks variant selle probleemi lahendamiseks on ainult geo-andmetega säutsude kogumine, ent see oleks laias plaanis ebamõistlik, sest kogudes vastava teema kohta kokku kõik säutsud olenemata ruumiandmetest, on andmekogumi kasutusvõime laiem ning lisaks ruumianalüüsile on võimalik teha ka sisuanalüüsi.

Valdav osa esimese poolfinaali andmestikust moodustasid kasutajate enda säutsud. Taaspostitatud säutsud moodustasid vaid 3.42% postitustest. Teise poolfinaali andmestikust umbkaudu 40% moodustasid taaspostitatud säutsud. Finaali madalamat ruumiandmete osakaaluga andmestikku seletab suur taaspostitatud säutsude osa (60.68%). Taaspostituste osakaal oli esimesest poolfinaalist finaalini kasvava tendentsiga. Esimese poolfinaali ajal olid enim taaspostitatud autorid ametlikud kasutajad, näiteks Eurovisooni enda Twitteri konto või BBC. Teise poolfinaali ja finaali puhul oli enim taaspostitatud autorite seas rohkem inimeste erakasutajaid. Taaspostituste suur osakaal võib tuleneda sellest, et sellise sündmuste käigus tekivad konkreetsed üksikud säutsud, mida palju taaspostitatakse. Selle töö raames on taaspostitatud säutsud probleemsed. Valdavalt ei lisata taaspostitatud säutsule geo-andmeid. Kuna selle uurimuse raames oli analüüsiks oluline postitatud säutsude asukoht, jäävad taastpostitatud säutsud ruumiandmete analüüsist välja.

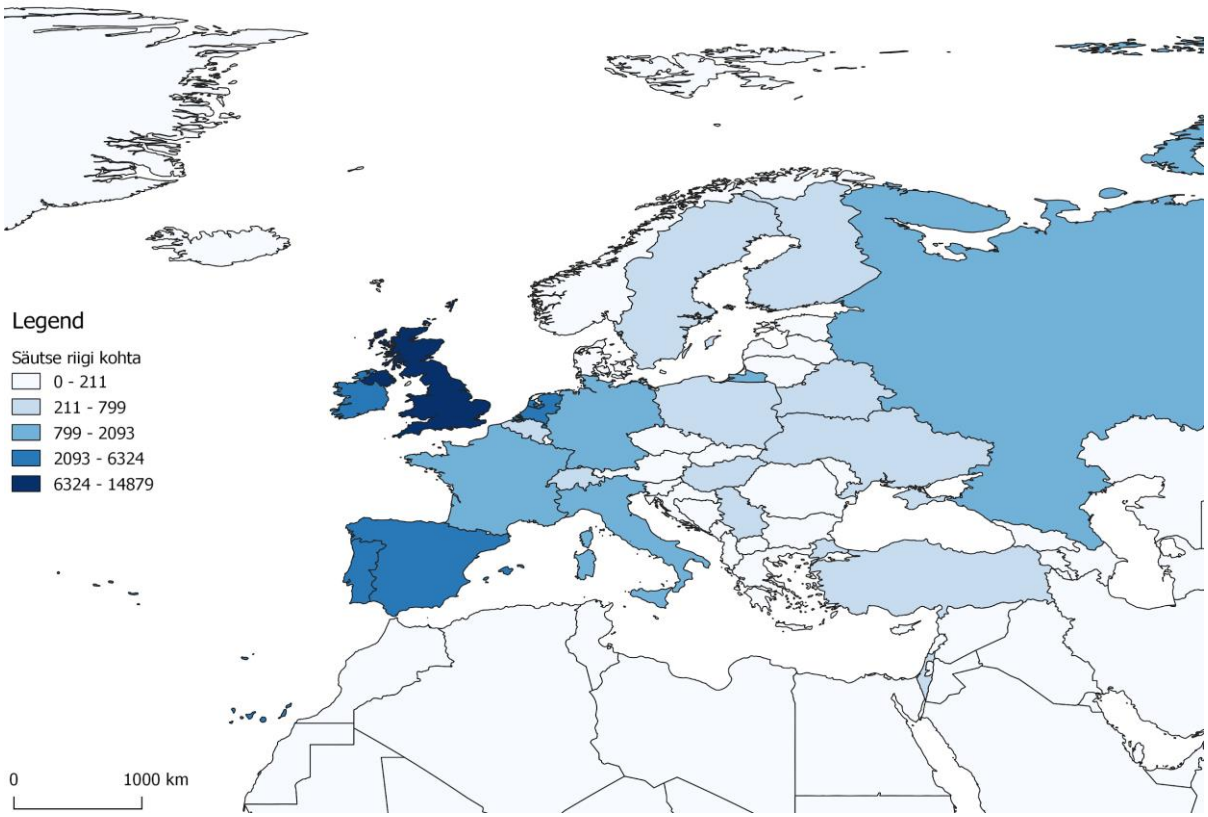
Eurovisiooni finaali jälgis see aasta umbkaudu 52 miljonit televaatajat üle Euroopa, Austraalia ja Ameerika Ühendriikide. Koguseliselt enim oli nende seas vaatajaid Saksamaalt (16%), Hispaaniast (14%), Suurbritanniast (13%) ning Prantsusmaalt (10%). Eurovisiooni võib populaarseks pidada riikides, kus võistlust vaatab suur osa üldpopulatsioonist. Kõige populaarsem on Eurovisioon Islandil, kus lausa 42% rahvastikust võistlust jälgis. Ühtlaselt teist kuni neljandat kohta jagavad omavahel Rootsi, Serbia ja Norra (26% populatsioonist), kelle järel tuleb 21% Küpros ja seejärel 20% Eesti ja Taani. (Eurovisionworld, 2018)

Kõige rohkem geo-andmetega säutse postitati Suurbritanniast, Londoni ümbrusest, teisteks kuumkohtadeks olid Madalmaad (joonis 6, 7). Palju ruumiandmetega säutse tuli ka Portugalist, Lissaboni piirkonnast, kus sel aastal Eurovisioon aset leidis. Väljaspoolt Euroopat oli asukohaga säutsude arvu poolest esirinnas Ameerika Ühendriigid ja Austraalia, mis olid vastavalt 9. ja 11. kohal riikidest, kust tuli enim geo-andmetega säutse (joonis 8). Asukoha poolest kõige eksootilisemad säutsud postitati Antarktikast ja Okeaaniast.

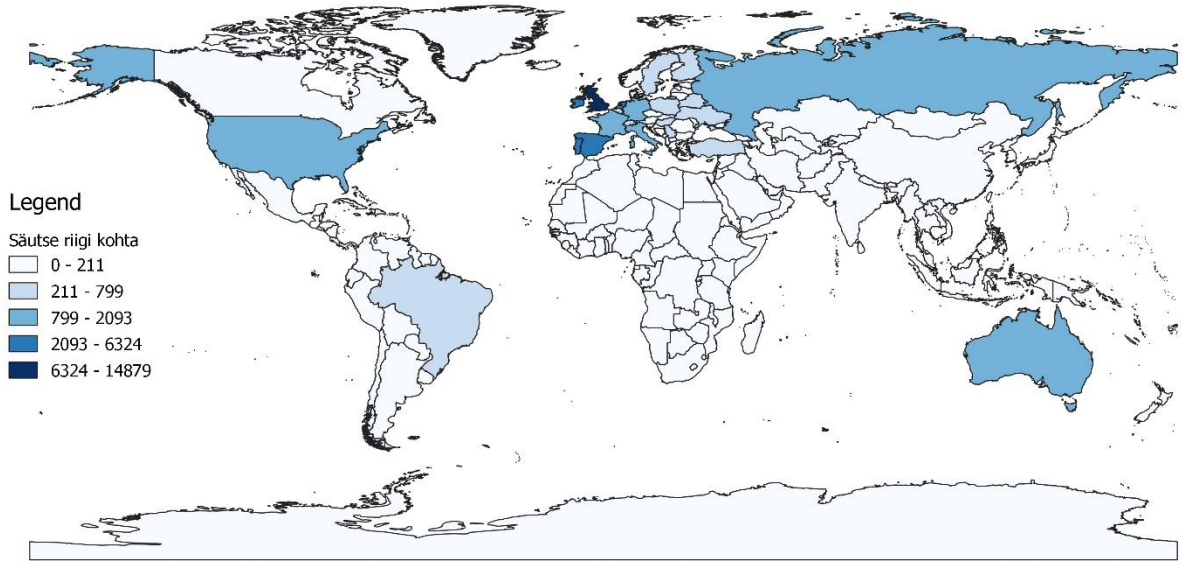
Säutsude keeleline jaotus oli poolfinaalides sarnane. Poolfinaalide jooksul säutsuti valdavalt inglise keeles, järgmisteks olid hispaania ning vene keel. Finaalis säutsuti enim hispaania keeles (~501 000 säutsu), inglise keeles säutsuti veidi vähem (~445 000 säutsu).



Joonis 6. Säutsude esinemistiheduse kuumkaart (*inglise k heatmap*) Euroopa ümbruses.



Joonis 7. Geo-andmetega säutsude kogus riigiti Euroopas.



Joonis 8. Geo-andmetega säutsude kogus riigiti üle maailma.

Andmetest selgus, et geo-andmetega säutsude põhjal ei saa järeldada seost säutsude koguse ja võistluse populaarsuse vahel riigi tasandil. Kõige rohkem geoandmetega säutse riigi kohta tuli Suurbritanniast, kus aga vaid 10% elanikest Eurovisiooni vaatavad. Riikidest, kus tegelikult sel aastal Eurovisioon populaarne oli ehk Island, Rootsi, Serbia ja Norra, ei jõudnud ükski riigiti säutsude arvu poolest kõrgele kohale, pigem jäid lausa edetabeli teise otsa. 42% Islandi populatsioonist jälgis võistlust, ent vaid 6,62% populatsioonist omab Twitteris kasutajat (StatCounter, 2018). Rootsis, Serbias ja Norras oli protsentuaalselt 26% populatsioonist võistlust vaatamas, ent ka nendes riikides on Twitter pigem ebapopulaarne. Rootsis täheldati Twitteri populaarsuse langust juba aastal 2016, mil see aastaga 4% langes, jõudes 18%-ni, kusjuures igapäevaseid kasutajaid on vaid 3% rahvastikust (The Internet Foundation of Sweden, 2016). Serbias on vaid 2,18% ja Norras 4,65% rahvastikust Twitteris (StatCounter, 2018). Rohkem tundub säutsude tihedus olevat sõltuv pigem vastava sotsiaalmeediaplatvormi populaarsusega riigis. Eeldust kinnitab asjaolu, et Euroopas enim aktiivseid Twitteri kasutajaid ongi Suurbritannias. Samuti on Twitteri populaarsuse poolest esikohal olev Ameerika Ühendriigid Eurovisiooni kohta säutsumise poolest 9. kohal, olgugi et nende riik isegi võistluses ei osale.

Andmete tõlgendamisel ei saa sotsiaalmeedia kontekstis jätta mainimata ka ealisi erinevuseid. Twitter ja muud sotsiaalmeediaplatvormid on populaarsed pigem noorte, kuni 25-aastaste seas. Selle uuringu andmete kogumisel kasutaja vanust ei registreeritud ja seega ei saa teha ka tegelikke järeldusi. Eurovisiooni lauluvõistluse kohta andmeid kogudes ei peaks see siiski

olema suureks takistuseks, sest ka seda võistlust jälgivad pigem nooremad, ent kasutades sarnast andmete kogumismeetodit teiste sündmuste tarbeks, tuleb olla tulemuste üldistamisel üldpopulatsioonile ettevaatlik.

Andmete kogumise faasis oleks võinud skriptis kasutada *psycopg2* teeki, et andmeid salvestada otse andmebaasi, mis oleks andmeanalüüsi aega vähendanud. Suuremate andmemahtude puhul tuleb optimeerida skripte, et vahemälu täitumine ei peataks skripti tööd. Andmete indekseerimine kiirendab päringuid. Säutsude asukoha määramine punktidena tekitab probleeme place-atribuudi puhul, sest Twitteri sisseehitatud atribuut on vigase geomeetriaga. Kokku oli kirjutatud umbes 600 rida koodi, millest kasulikuks osutus 100. Varasemad skriptid, mis käsitlesid andmete tõmbamist varasematest säutsudest jäeti kõrvale, et koguda suuremal määral andmeid.

Selleks, et sõelale oleks jäänud veelgi rohkem säutse, mida üritusest säutsuti, oleks võinud kasutada andmete kogumisel otsitava märksõnana ka selle aasta lauluvõistluse ametlikku teemaviidet #AllAboard, sest teemaviite kasutamist reklaamiti kogu võistluse käigus väga aktiivselt. Ilmselt postitasid paljud teemaga kursis olevad inimesed ja aktiivsed jälgijad just pigem selle teemaviitega kui lihtsalt mainisid Eurovisiooni. Täpsema ruumianalüüsi ja üldise sisuanalüüsi raames on oluline olla sündmusega, mille ajal andmeid kogutakse, hästi kursis, et valida andmete kogumisel õiged märksõnad.

Andmekaeve Twitterist 2018. aasta Eurovisiooni näitel

Patrick Joan Thomson

Kokkuvõte

Bakalaureusetöös uuriti 2018. aastal toimunud Eurovisiooni lauluvõistluse näitel andmekaevet Twitterist. Vaadeldi andmete üldist jaotust ning analüüsiti ruumiandmetega säutse. Püstitati kolm hüpoteesi: geo-andmetega säutsude osakaal on vähemalt 5%; säutsude hulk tunnis on andmekogumisperioodi jooksul kõige suurem sündmuse ajal; mida populaarsem on sündmus riigis, seda rohkem postitatakse sellest riigist säutse.

Uurimuses selgus, et geo-andmetega säutsude hulk on üldiselt madal ning taaspostituste arv vähendab seda veelgi. Esimene hüpotees ei leidnud kinnitust. Leidis kinnitust aga teine hüpotees, et on säutsude hulk tunnis on andmekogumisperioodi jooksul kõige suurem sündmuse ajal. Lisaks selgus, et säutsude arv riigiti sõltub enim vastava sotsiaalmeediaplatvormi populaarsusest riigis. Sündmuse enda populaarsus riigis ei ole tingimata seotud selle suurema kajastusega sotsiaalmeedias, seega kolmas hüpotees kinnitust ei leidnud. Geoandmetega säutsude kogum ei ole üldistatav kogupopulatsioonile sõltuvalt nende andmete vähesusest ja teadmatusest vanuselise jaotuvuse kohta.

Bakalaureusetöö annab ülevaate Eurovisiooni lauluvõistluse kajastusest sotsiaalmeediaplatvormil Twitter. Töö on vajalik, et anda ülevaade järjest rohkem kasutavate andmeallikate analüüsi võimaluste kohta. Kasutades töö käigus välja töötatud skripte on võimalik andmekogumisprotsessi rakendada ka teiste sündmuste jälgimisele.

Data mining from Twitter on the example of Eurovision Song Contest 2018

Patrick Joan Thomson

Summary

In this Bachelors' thesis the research subject is data mining in the example of the Eurovision Song Contest of 2018. In the paper the overall distribution of tweets with spatal data was analyzed. Three hypotheses were set: geo-tagged tweets make up at least 5% of all the collected tweets, the highest number of tweets per hour during the data collection period is greatest at the time of the event and the more popular the event in the country the more tweets are posted.

The study revealed that the amount of geo-tagged tweets is relatively low and high numbers of retweets lower it even more. First hypothesis was not confirmed. Second hypothesis was confirmed, the tweets per hour are the greatest during the event. In addition it was found that the amount of tweets is more dependent on the use of social platform in the country not the popularity of the event, which means that the third hypothesis was not confirmed. Geo-tagged tweets are not universally applicable to the population as a whole depending on the lack of data and unknown age distribution.

The Bachelors' thesis gives an overview of the Eurovision Song Contest coverage from Twitter. The paper is needed to provide an overview of one of the possible ways of analyzing increasingly used data sources.

Tänusõnad

Soovin avaldada tänu juhendajale Evelyn Uuemaale, Erki Saluveerile ja Mairo Puusepale mitmekülgse abi ja nõu eest. Samuti sooviksin välja tuua Kärt Puusepa igakülgse abi eest töö kirjutamise perioodil.

Kasutatud kirjandus

Agarwal, A., Singh, R., Toshniwal, D. (2018). Geospatial sentiment analysis using twitter data for UK-EU referendum, Journal of Information and Optimization Sciences, 39:1, 303-317, DOI: 10.1080/02522667.2017.1374735

Arthur, R., Boulton, C. A., Shotton, H., Williams, H. T. P. (2018). Social sensing of floods in the UK. PLoS ONE. 13(1): e0189327. https://doi.org/ 10.1371/journal.pone.0189327

Bao, J., Liu, P., Yu, H., Xu, C. (2017). Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accident Analysis & Prevention. 106, 358-369.

Boulton, C., Shotton, H., Williams, H. T. P. (2016). Using Social Media to Detect and Locate Wildfires. The Workshops of the Tenth International AAAI Conference on Web and Social Media Social Web for Environmental and Ecological Monitoring: Technical Report WS-16-20.

Daniel, M., Neves, R. F., Horta, N. (2017). Company event popularity for financial markets using Twitter and sentiment analysis. Expert Systems with Applications. 71, 111-124.

Eesti Rahvusringhääling (2018). ERR-i telepilti Eurovisiooni finaalist vaatas 406 000 inimest. Kasutatud 22.05.2018 aadressil https://menu.err.ee/831623/err-i-telepilti-eurovisiooni-finaalist-vaatas-406-000-inimest.

Eurovisionworld (2018). 186 million watched Eurovision 2018. Kasutatud 24.05.2018 aadressil https://eurovisionworld.com/esc/186-million-watched-eurovision-2018.

Github (2018). Bakatweet. Kasutatud 27.05.2018 aadressil https://github.com/patrickjoan/bakatweet.

Kim, J., Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. International Journal of Information Management. 38(1), 86-96.

Lindsay, B. R. (2011). Social media and disasters: Current uses, future options and policy considerations. Congressional research service reports, 13. Kasutatud 27.05.2018 aadressil http://fas. org/sgp/crs/homesec/R41987.pdf.

Rosen, A. (2017). Tweeting Made Easier. Twitter Blog. Kasutatud 20.05.2018 aadressil https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

Sakaki, T., Okazaki, M., Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web, ACM, 851-860.

Sousa, L., Mello, R., Cedrim, D., Garcia, A., Missier, P., Uchôa, A., ...Romanovsky, A. (2018). VazaDengue: An information system for preventing and combating mosquito-borne diseases with social networks. Information Systems. 75, 26-42.

StatCounter (2018). Social Media Stats Iceland. Kasutatud 27.05.2018 aadressil http://gs.statcounter.com/social-media-stats/all/iceland.

StatCounter (2018). Social Media Stats Norway. Kasutatud 27.05.2018 aadressil http://gs.statcounter.com/social-media-stats/all/norway.

StatCounter (2018). Social Media Stats Serbia. Kasutatud 27.05.2018 aadressilhttp://gs.statcounter.com/social-media-stats/all/serbia.

Tenkanen, H., Minin, E.D., Heikinheimo, V., Hausmann, A., Herbst, A., Kajala, L., Toivonen, T., (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Scientific Reports. 7, 17615. doi:10.1038/s41598-017-18007-4.

The Internet Foundation of Sweden (2016). Facebook dominates and Twitter drops – how Swedes use social media 2016. Kasutatud 27.05.2018 aadressil https://www.iis.se/english/press/pressreleases/facebook-dominates-and-twitter-drops-how-swedes-use-social-media-2016/.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Patrick Joan Thomson,

annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Andmekaeve Twitterist 2018. aasta Eurovisiooni näitel

mille juhendaja on Evelyn Uuemaa,

reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni; üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 28.05.2018