

CS 751: Assignment 4

Bharath Kongara

Spring 2015

Contents

1	Question 1	2
	1.1 Solution	2
	1.2 Code Listing	2
2	Question 2	3
	2.1 Solution	3
3	Question 3	4
4	Question 4	17
	4.1 Solution	17

1 Question 1

- Using the pages from A3 that boilerpipe successfully processed, download those representations again & reprocess them with boilerpipe.:
- Document the time difference (e.g., $\text{Time}(A4) - \text{Time}(A3)$).
- Compute the Jaccard Distance x for each pair of pages (i.e., $P(A3)$ & $P(A4)$) for:
 - Unique terms (i.e., unigrams)
 - Bigrams
 - Trigrams
- See:http://en.wikipedia.org/wiki/Jaccard_index
- For each of the 3 cases (i.e., 1-, 2-, 3-grams) build a Cumulative Distribution Function that shows the % change on the x-axis & the % of the population on the x-axis
- See:http://en.wikipedia.org/wiki/Cumulative_distribution_function
- Give 3-4 examples illustrating the range of change that you have measured.

1.1 Solution

- I used same script `getText.py` from A3 which uses `jusText[1]` to remove templates from HTML files of 10000 URIs.
- `jusText[1]` is a tool for removing boilerplate content such as navigation links, headers and footers from HTML pages.
- Time Difference between processing boilerpipe of A3 and A4 is 27 days.
-

1.2 Code Listing

Below is the python code that is used to remove HTML templates from all HTML pages of 10000 URIs.

2 Question 2

- Using the pages from Q1 (A4), download all TimeMaps (including TimeMaps with 404 responses, i.e. empty or null TimeMaps)
 - Upload all the TimeMaps to github
- Build a CDF for # of mementos for each original URI (i.e., x-axis = # of mementos, y-axis = % of links)
- See: <http://timetravel.mementoweb.org/guide/api/>

2.1 Solution

Each line of text file are parsed by splitting words on space and constructed dictionary to get the unique word list. Here is the Python program which calculates top 50 unique word frequencies.

3 Question 3

- Using 20 links that have TimeMaps
 - With ≥ 20 mementos
 - Have existed ≥ 2 years (i.e., Memento-Datetime of first memento is April XX, 2013 or older)
 - Note: select from Q1/Q2 links, else choose them by hand
- For each link, create a graph that shows Jaccard Distance, relative to the first memento, through time
 - x-axis: continuous time, y-axis: Jaccard Distance relative to the first memento

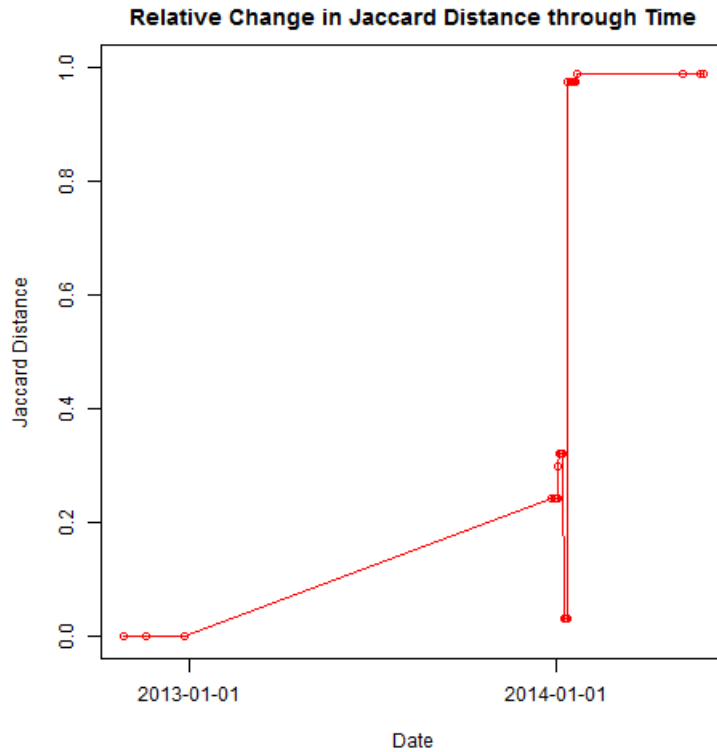


Figure 1: Jaccard Distance relative to the first memento url1

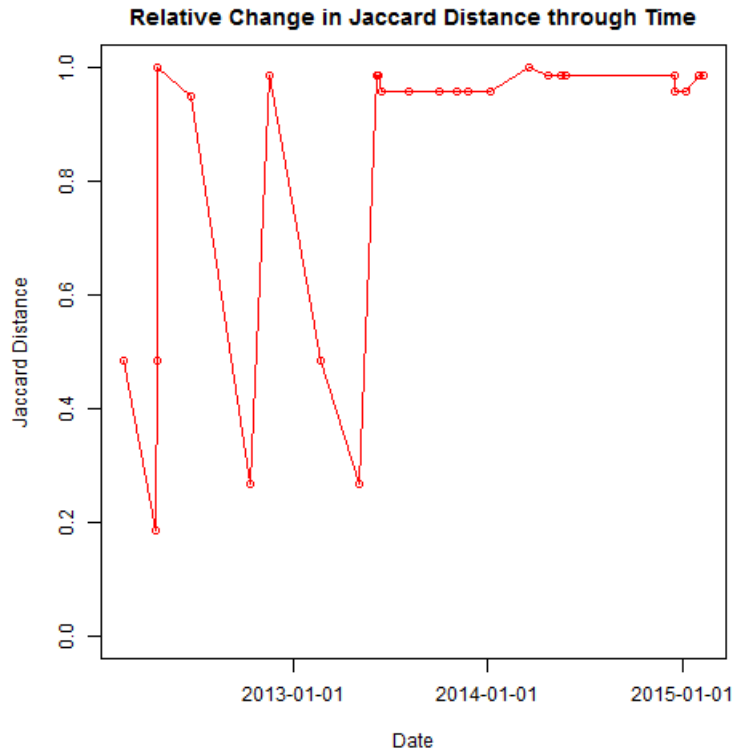


Figure 2: Jaccard Distance relative to the first memento url2

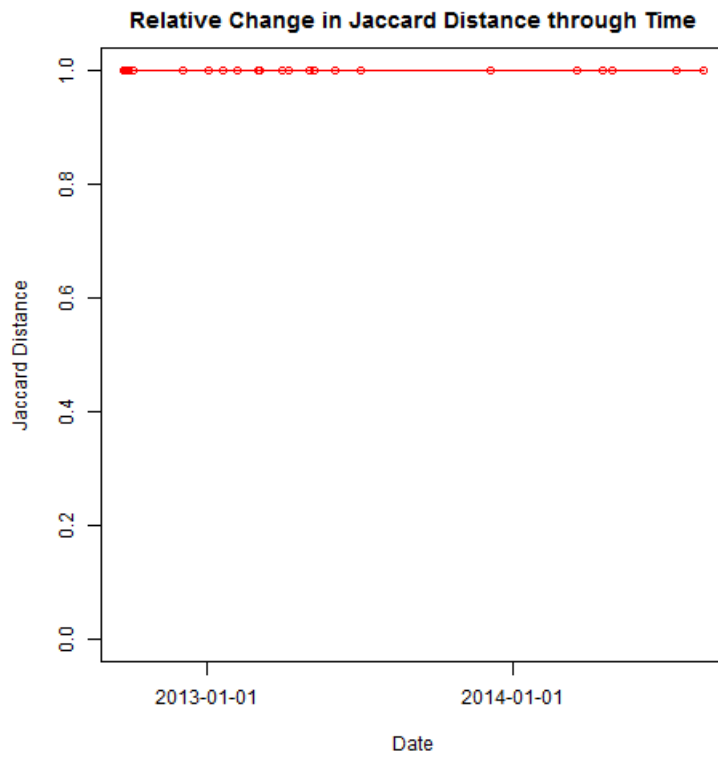


Figure 3: Jaccard Distance relative to the first memento url3

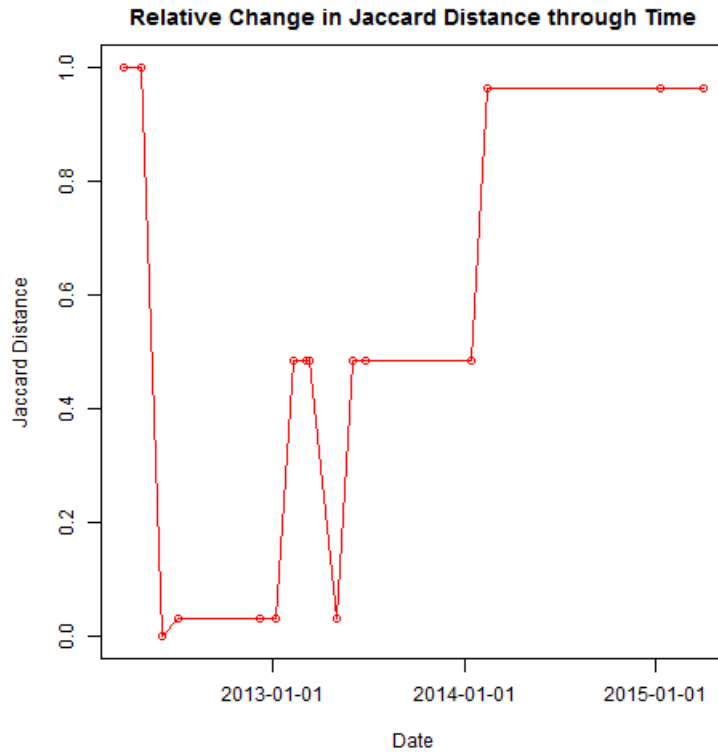


Figure 4: Jaccard Distance relative to the first memento url4

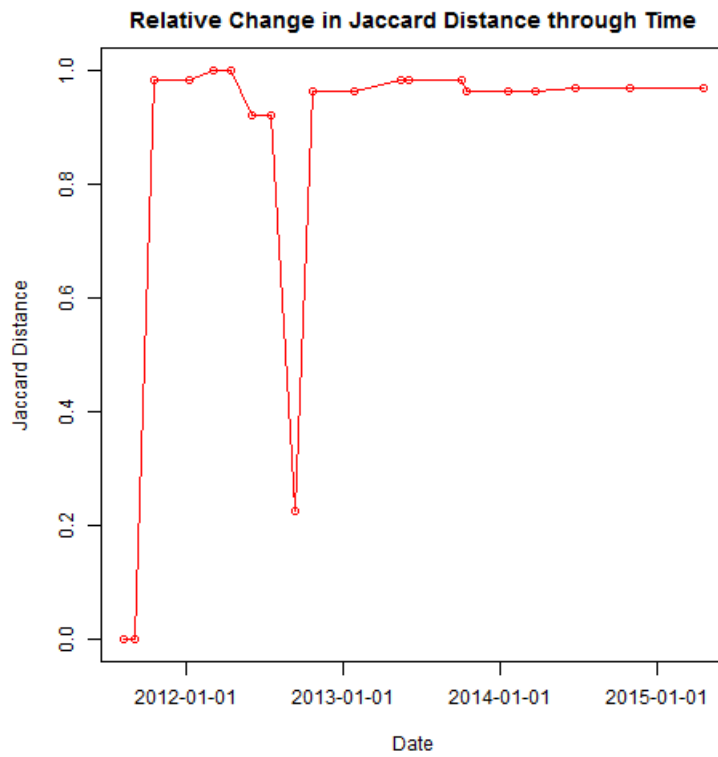


Figure 5: Jaccard Distance relative to the first memento url5

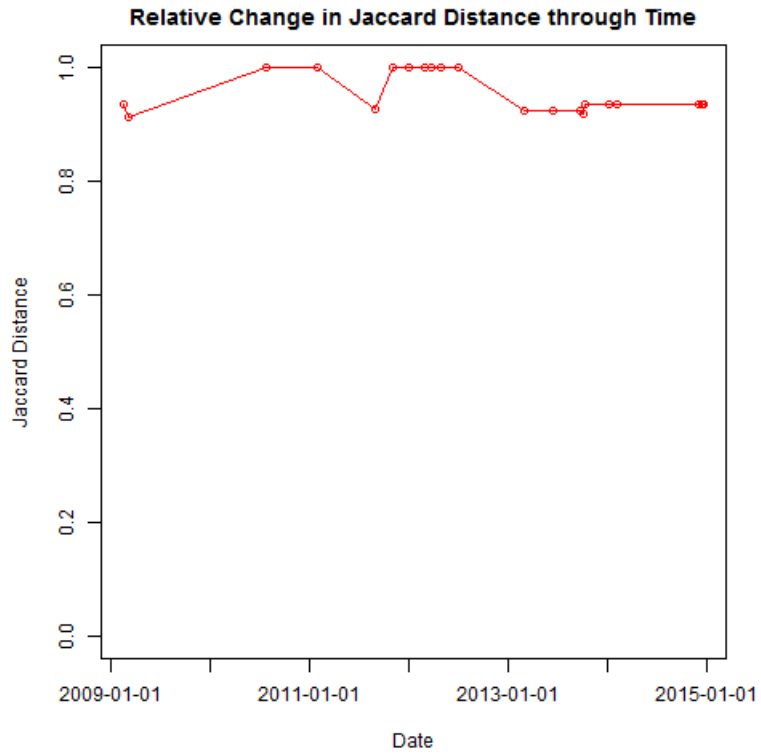


Figure 6: Jaccard Distance relative to the first memento ur65

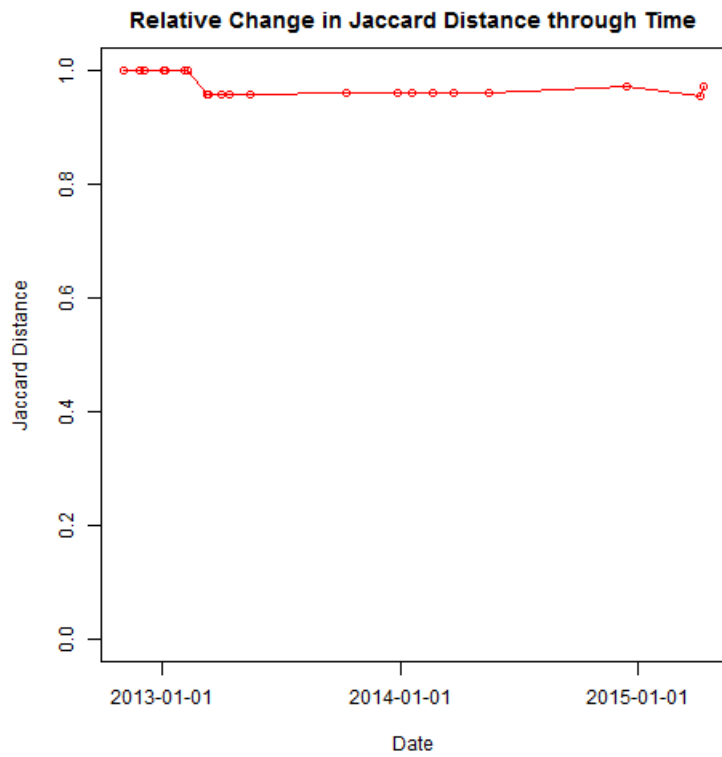


Figure 7: Jaccard Distance relative to the first memento url7

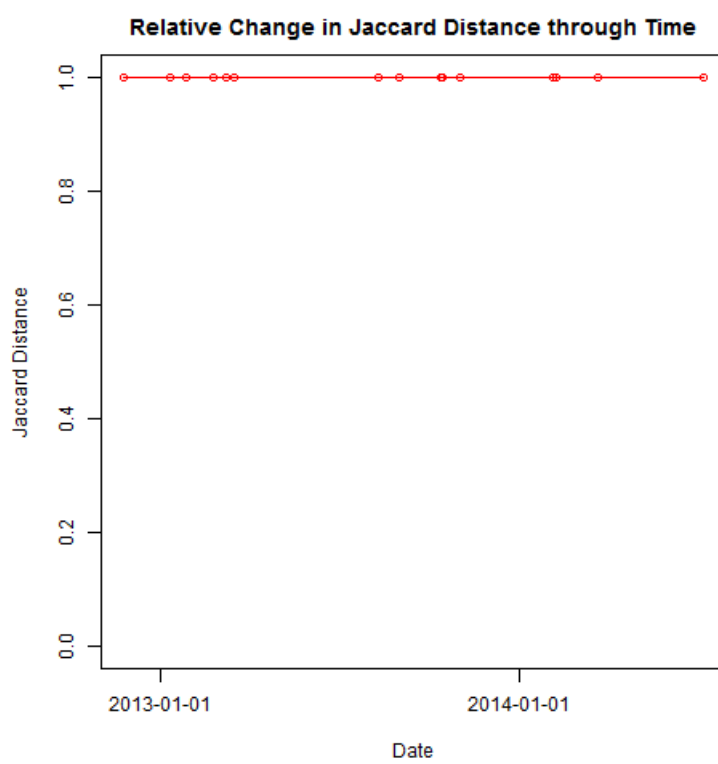


Figure 8: Jaccard Distance relative to the first memento url8

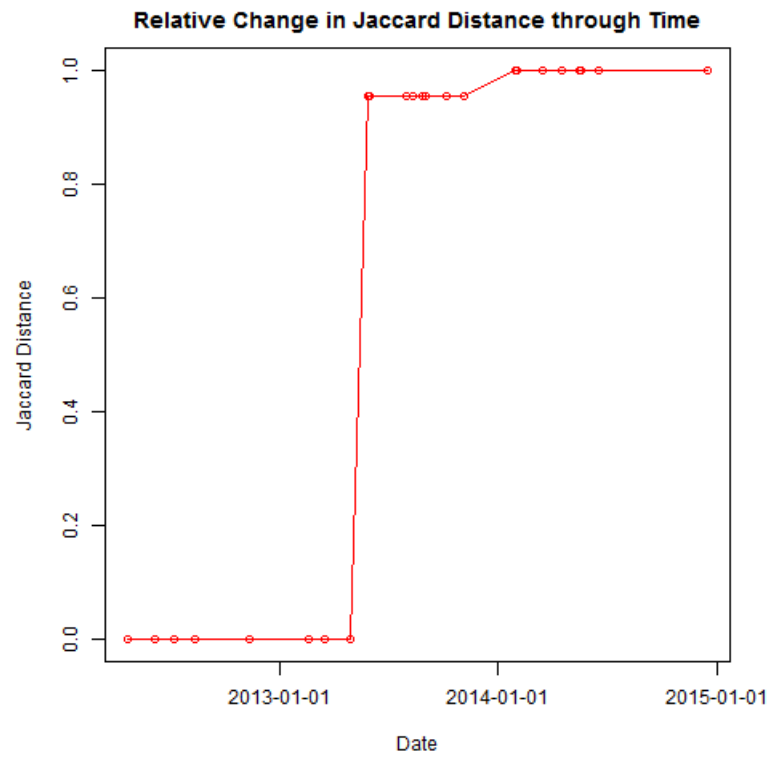


Figure 9: Jaccard Distance relative to the first memento url9

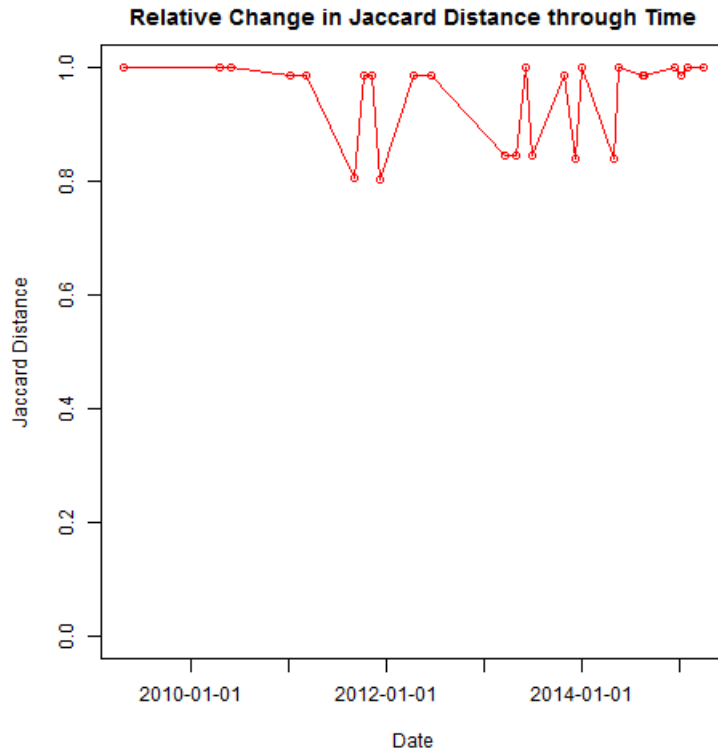


Figure 10: Jaccard Distance relative to the first memento url10

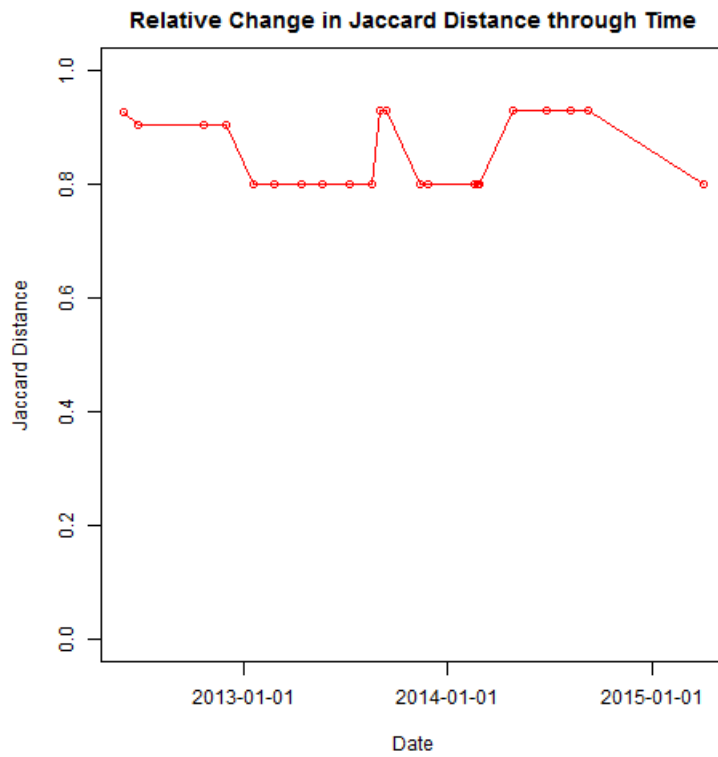


Figure 11: Jaccard Distance relative to the first memento url11

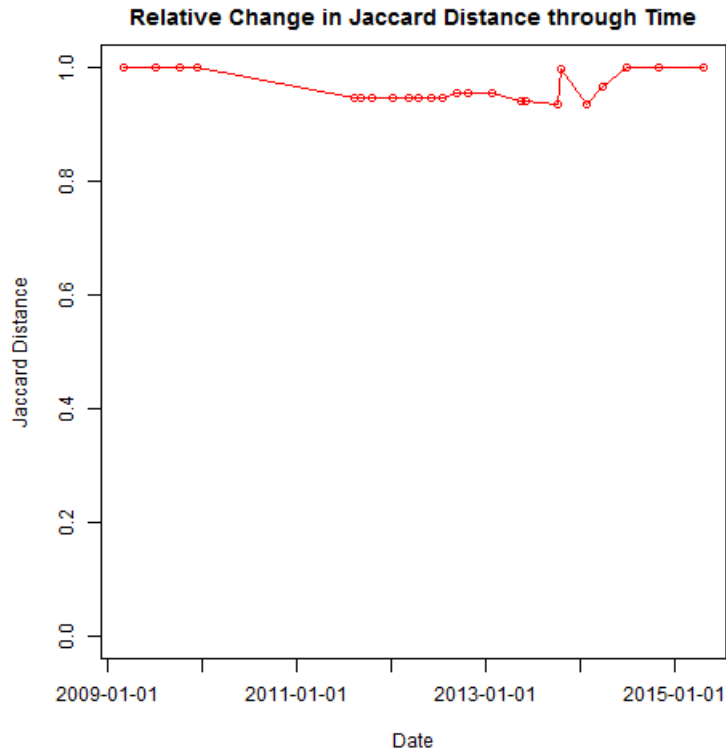


Figure 12: Jaccard Distance relative to the first memento url12

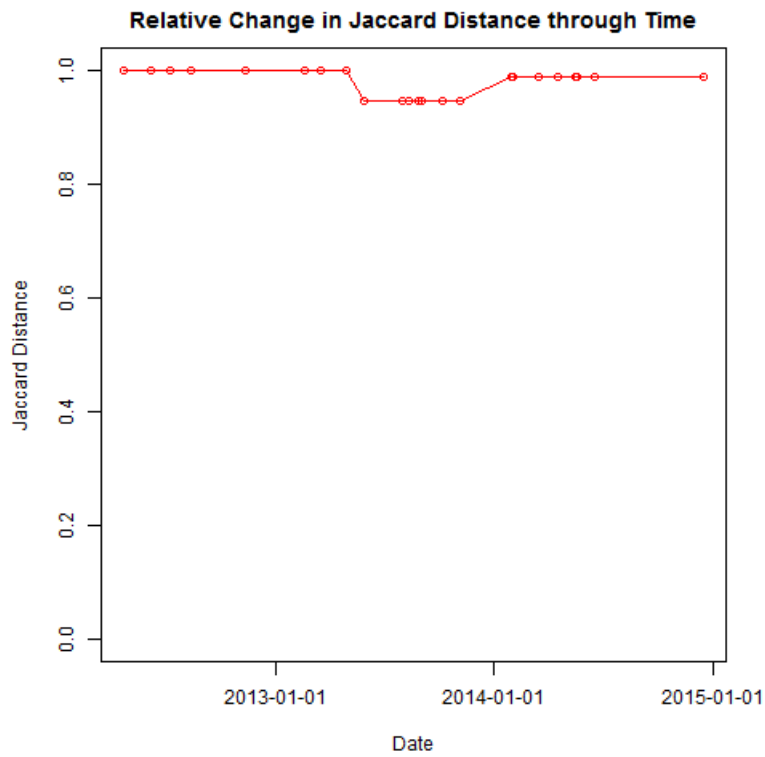


Figure 13: Jaccard Distance relative to the first memento url13

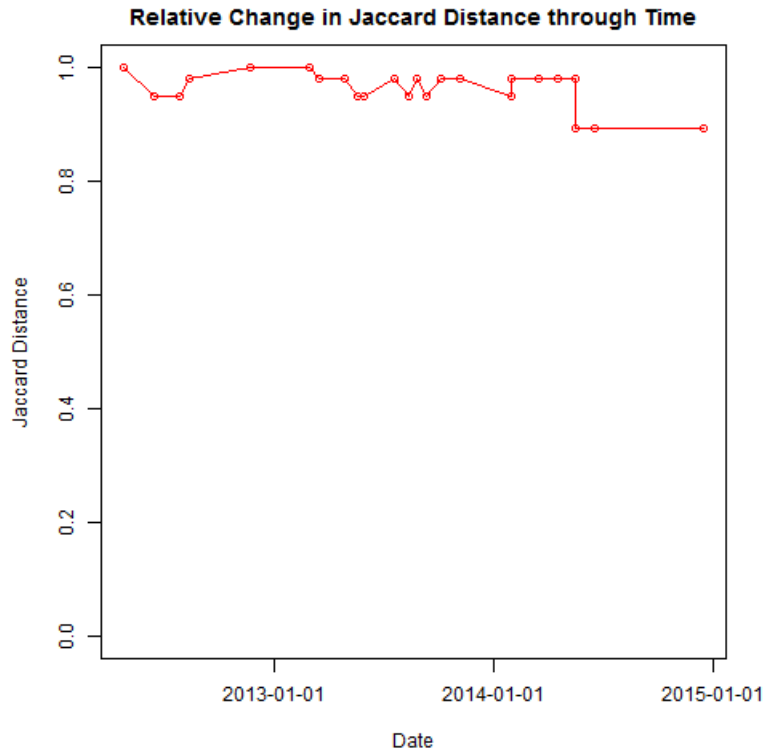


Figure 14: Jaccard Distance relative to the first memento url14

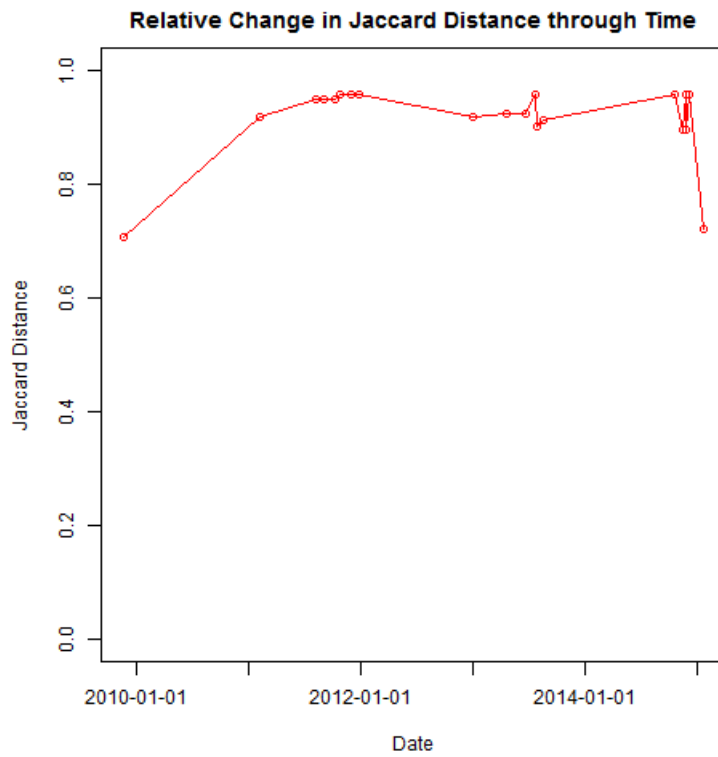


Figure 15: Jaccard Distance relative to the first memento url15

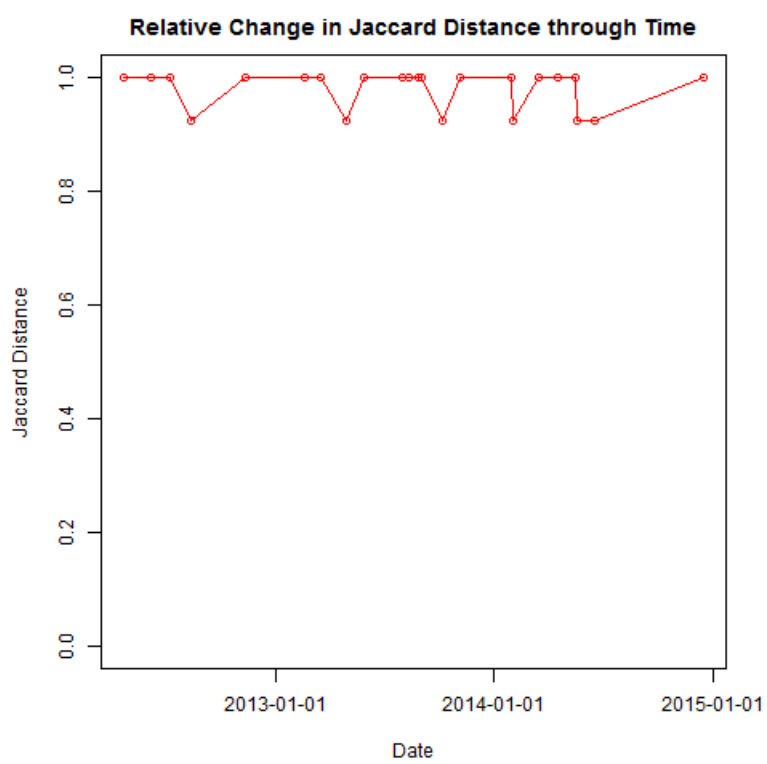


Figure 16: Jaccard Distance relative to the first memento url16

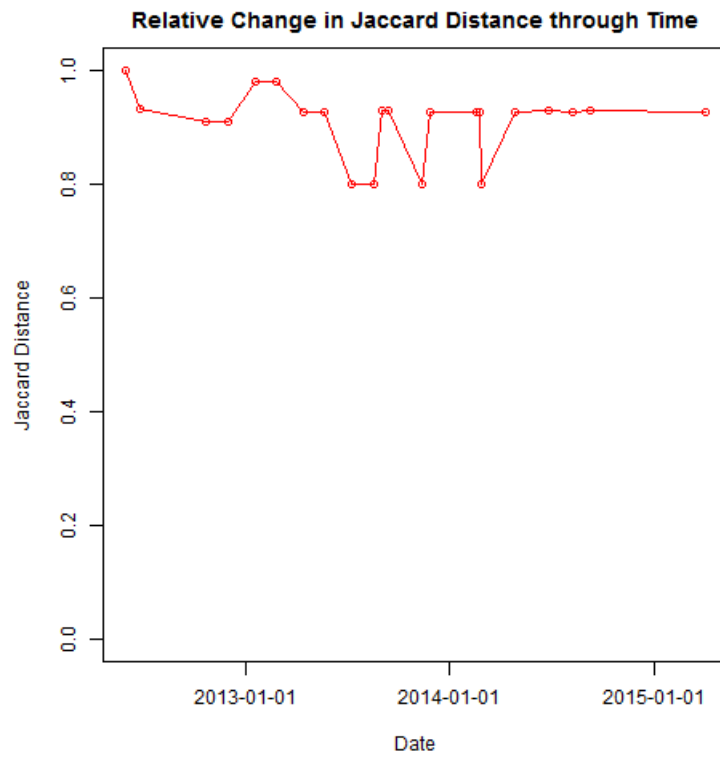


Figure 17: Jaccard Distance relative to the first memento url17

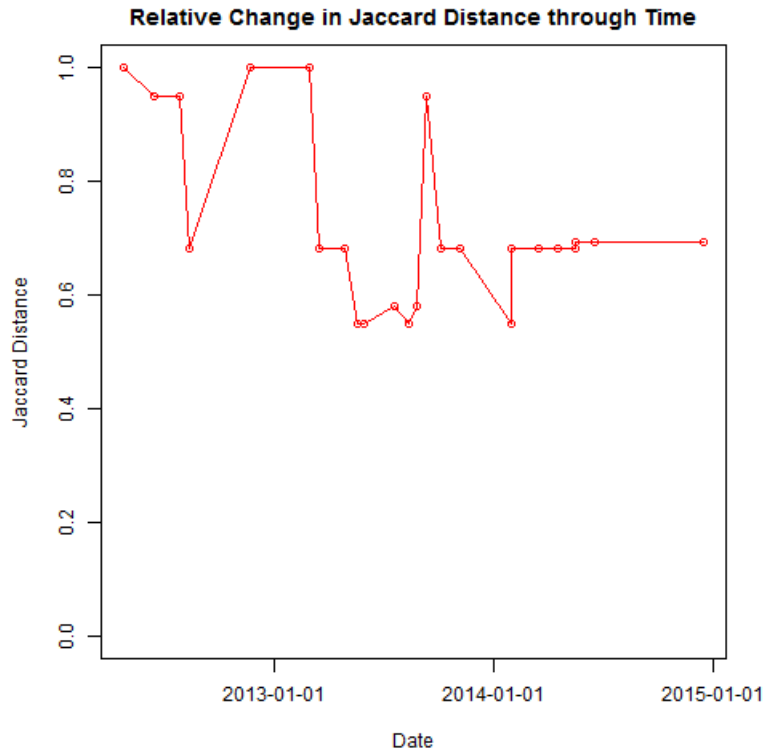


Figure 18: Jaccard Distance relative to the first memento url18

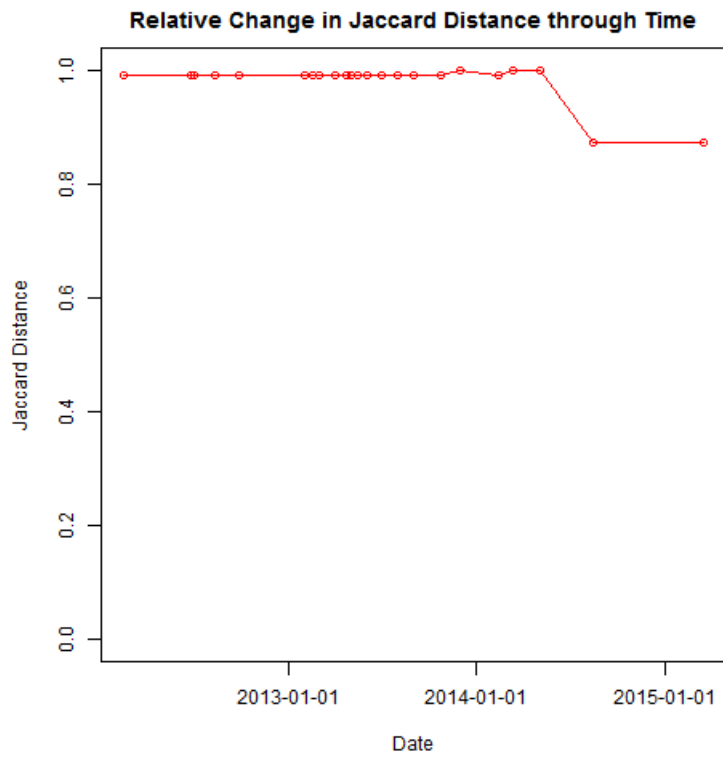


Figure 19: Jaccard Distance relative to the first memento url19

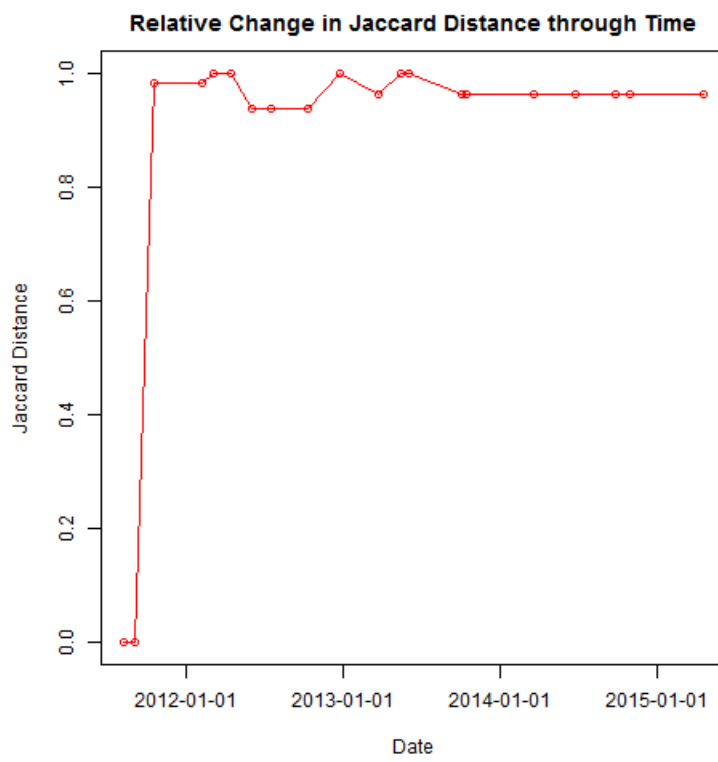


Figure 20: Jaccard Distance relative to the first memento url20

4 Question 4

- Choose a news-related event
- Use twarc.py to collect 1000 tweets, every day for 5 different days
 - See: <https://github.com/edsu/twarc>
- For each day:
 - Create a wall
 - Build a tag/word cloud for each day
 - Create a map using GeoJSON & Github
 - * <https://help.github.com/articles/mapping-geojson-files-on-github/>
- Discuss in detail lessons learned, experiences, etc.

4.1 Solution

- Installed twarc[1] to retrieve 1000 tweets.
- I used twarc to build tweets wall, word cloud and to retrieve geojson.

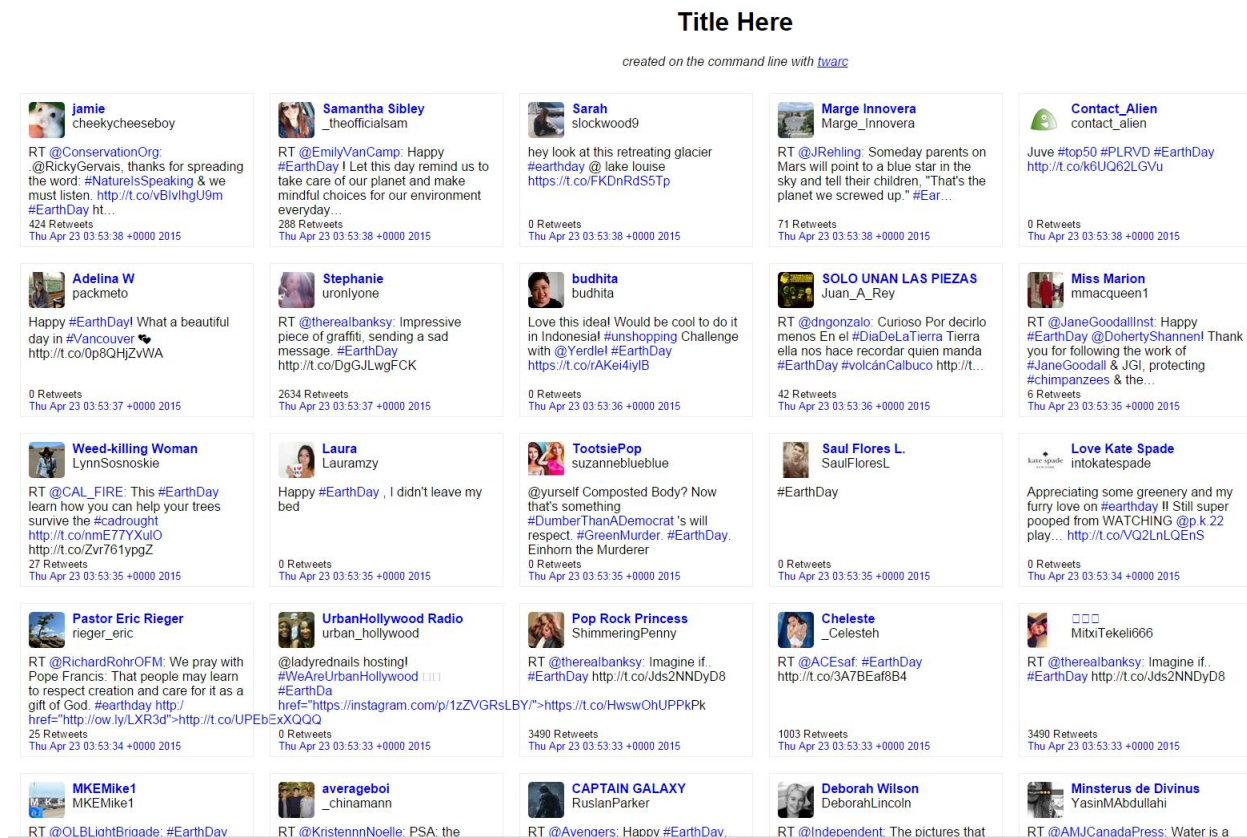


Figure 21: Tweets Wall Day1

Title Here

created on the command line with [tvarc](#)


























 Estelle cartidge Stelc81 RT @loveyogabears: Happy Earth Day to all Earthlings ♥ blessings from @loveyogabears @shaktistudios @bikramliverpool @Stelc81 #earthday ht... 2 Retweets Thu Apr 23 22:45:36 +0000 2015	 Sumicel USA sumicelusa #Samsung patents a lightweight #fitness tracking band (+photos) - http://t.co/g16thFDV0J #News #EarthDay 0 Retweets Thu Apr 23 22:45:34 +0000 2015	 Jessica Orwig orwigwrites RT @mashable: We know #EarthDay can be depressing. So here are 4 reasons to be optimistic about the planet http://t.co/7mMK1SqYEB http://t.co/... 65 Retweets Thu Apr 23 22:45:34 +0000 2015	 Bri BriThibodeaux RT @NigelBarker: It's #EarthDay & the perfect time for us to remember to look after & respect our beautiful planet #Roma #hugatree x http://t.co/... 61 Retweets Thu Apr 23 22:45:32 +0000 2015	 Consulado Ecu Queens CEcuQueens RT @ApoyaAIEcuador: Our voices are stronger together than apart. Help our brothers and sisters in the Amazons #WeDeserveJustice #EarthDay h... 12 Retweets Thu Apr 23 22:45:31 +0000 2015
 Irina Burkova IrinaBurkova RT @Jasmine_Soul: Happy #EarthDay 🌱🌱 http://t.co/OvaWf2QrmE 4 Retweets Thu Apr 23 22:45:26 +0000 2015	 Heather ourcirltime Happy #EarthDay! We made a bird feeder during the craft segment on today's http://t.co/AvDOQj87BL video activity. http://t.co/opHxQZ38n 0 Retweets Thu Apr 23 22:45:26 +0000 2015	 Hello Enterprise HelloEnterpris RT @GoodsForGoodCo: BETTER LIFE - what-EVER! http://t.co/lcBhTLE2pQ #BetterLife #EcoFriendly #homecare #EarthDay #Online #shopping http://t.co/... 3 Retweets Thu Apr 23 22:45:23 +0000 2015	 bsbafflesbrains bsbafflesbrains RT @elliosch: @firedoglake nails it. Thankfully @motherjones didn't realize their grave mistake earlier. http://t.co/TKQvW8yWNv http://t.co/... 3 Retweets Thu Apr 23 22:45:23 +0000 2015	 oops_im_amazing DestinyFugate #EarthDay 0 Retweets Thu Apr 23 22:45:21 +0000 2015
 Brian David Stockman Carlisleboy RT @WAGNERGIRLE: #APicturesWorthOneThousandWords 🌱 Obama's empty #earthday speech 🌱 http://t.co/nlNk5oXZUsUs #tcot #pcra #ccot http://t.co/... 10 Retweets Thu Apr 23 22:45:21 +0000 2015	 Sean Nelson opinionatedwino #earthday http://t.co/stfprs3jyB 0 Retweets Thu Apr 23 22:45:20 +0000 2015	 r.e.e.n.a reenahuffman13 RT @JoeyGraceffa: HAPPY #EarthDay in the spirit I've been playing @TreeStorygame try it out its so cute! #HaveFunDoGood http://t.co/FjGdzY0j2y 373 Retweets Thu Apr 23 22:45:18 +0000 2015	 Lachlan Milt MiltPozjbk RT @4BillLewis: The most expensive #Starbucks drinks ever http://t.co/q8KUXVsp9S via @nypost #socialmedia #reteeet #coffee #StarbucksProble... 133 Retweets Thu Apr 23 22:45:18 +0000 2015	 DavidPMiranda™ DavidPMiranda Happy Earth Day! So excited I got up before Dawn. Fortunately Dawn sleeps until noon. #EarthDay 0 Retweets Thu Apr 23 22:45:17 +0000 2015
 Abbey Farkas AbFark RT @radicalbytes: This #EarthDay remember personal recycling & water conservation initiatives are designed to shift blame from corporations... 237 Retweets Thu Apr 23 22:45:15 +0000 2015	 sal levett RT @WAGNERGIRLE: #APicturesWorthOneThousandWords 🌱 Obama's empty #earthday speech 🌱 http://t.co/nlNk5oXZUsUs #tcot #pcra #ccot http://t.co/... 10 Retweets Thu Apr 23 22:45:15 +0000 2015	 Jonathan Cohee joncohee RT @PaigeToccoTerra: Kudos to my daughter Emmalyn, and all who dedicate their careers to preserving wildlife habitats #EarthDay 2 Retweets Thu Apr 23 22:45:15 +0000 2015	 papadukeschaney KevinDChaneySr RT @mashable: We know #EarthDay can be depressing. So here are 4 reasons to be optimistic about the planet http://t.co/7mMK1SqYEB http://t.co/... 65 Retweets Thu Apr 23 22:45:15 +0000 2015	 Jaime Silva Jaime0 RT @FortuneMagazine: See Apple's glossy #EarthDay video http://t.co/1eXXRKjPcl http://t.co/aCmunNe6Yh 16 Retweets Thu Apr 23 22:45:14 +0000 2015
 Yordano YordanoNapa RT @NASA: Ah. #Earth. there's 0 Retweets Thu Apr 23 22:45:14 +0000 2015	 SOLFARMERS SOLFARMERS Thursdav dinner idea 4 #EarthDay by 0 Retweets Thu Apr 23 22:45:14 +0000 2015	 Abbey Farkas AbFark RT @radicalbytes: This #EarthDay 0 Retweets Thu Apr 23 22:45:14 +0000 2015	 Keith Tharp KeithTharp Our Earth Day haul at Hodoson 0 Retweets Thu Apr 23 22:45:14 +0000 2015	 Giovanna Quina Klapp gioquina #chile #volcanes esta es 1 fumarola 0 Retweets Thu Apr 23 22:45:14 +0000 2015

Figure 22: Tweets Wall Day2

Title Here

created on the command line with [twarc](#)


























 GreeneaseApp GreeneaseApp Our fabulous intern, @sevincsey, blending in some smoothies at GWU's #EarthDay http://t.co/0w17ULrReB 0 Retweets Fri Apr 24 13:40:21 +0000 2015	 Little Critterz LittleCritterz3 #Sloths are an integral part of tropical rain forest ecosystems #EarthDay #nature #environment #Earthdayeveryday 0 Retweets Fri Apr 24 13:40:21 +0000 2015	 Anjana AnjanaKandhan RT @Avengers: Happy #EarthDay, from Earth's Mightiest! http://t.co/IRBeo0kbT 1879 Retweets Fri Apr 24 13:40:21 +0000 2015	 CJ Ronneberg braintwat RT @deborahhallsted: As if on cue, snow strikes America as Pres. Obama talks global warming for #EarthDay [photos] http://t.co/bK4HYA4HL5 1 Retweets Fri Apr 24 13:40:16 +0000 2015	 Connected Universe ConnectdUnivrse Nature is a portal to deeper understanding ~ Malcom Carter, director of #ConnectedUniverse #EarthDay http://t.co/W5YNI7i9of 0 Retweets Fri Apr 24 13:40:15 +0000 2015
 OEM Off-Highway OEMOffHighway It was a busy social media week for us! Check out @intematparis coverage with #OEMatINTERMAT and #EarthDay coverage with #GreenOEM 0 Retweets Fri Apr 24 13:40:14 +0000 2015	 sevenbowie sevenbowie RT @ConservationOrg: @RickyGervais, thanks for spreading the word: #NatureIsSpeaking & we must listen. http://t.co/vB1vIhgU9m #EarthDay ht... 732 Retweets Fri Apr 24 13:40:12 +0000 2015	 Danny Ivatt Daniellvatt According to #earthday crap where I am, I need to eat less meat and be more vegan. Neither of which is conducive to being swole. #nothanks 0 Retweets Fri Apr 24 13:40:10 +0000 2015	 Ashley Ashleyash27 RT @Martha_x15: @LisaVanderpump >Important reminder on #EarthDay. <a href="http://t.co/CiObnwtkm2<...loll">http://t.co/CiObnwtkm2<...loll 10 Retweets Fri Apr 24 13:40:09 +0000 2015	 Chris cmdiehl29 RT @VictoriaJustice: How did you celebrate #EarthDay? Join me & @DoSomething by recycling clothes at @hmusal http://t.co/dFL8Ob0bAv 382 Retweets Fri Apr 24 13:40:07 +0000 2015
 gidgiegoomie gidgiegoomie RT @Martha_x15: @LisaVanderpump >Important reminder on #EarthDay. <a href="http://t.co/CiObnwtkm2<...loll">http://t.co/CiObnwtkm2<...loll 10 Retweets Fri Apr 24 13:40:04 +0000 2015	 Dolores Briseno mofodolo RT @Martha_x15: @LisaVanderpump >Important reminder on #EarthDay. <a href="http://t.co/CiObnwtkm2<...loll">http://t.co/CiObnwtkm2<...loll 10 Retweets Fri Apr 24 13:40:04 +0000 2015	 giveaway 100K coins coins_100k RT @SlashedBeauty: It's #EarthDay! Discover #ecofriendly #skincare brand, @derma_e & win their best products: http://t.co/tkQ1TyvA3J http://t.co/tkQ1TyvA3J 19 Retweets Fri Apr 24 13:40:04 +0000 2015	 PinkGreen Creative WeArePinkGreen #FeatureFontFriday is here, with #earthday as the theme. @miradax goes green with Grafolita. #typography #design #... http://t.co/6rHY8MnjAA 0 Retweets Fri Apr 24 13:39:57 +0000 2015	 Agility CSR AgilityCSR Good #EarthDay memories indeed. https://t.co/YNFdUpE1y3 0 Retweets Fri Apr 24 13:39:55 +0000 2015
 Team Lupold Thoman LupoldThoman Green Team!!! Thank you for all your hard work! Wonderful video about recycling! @NH_Elementary #earthday http://t.co/3obgMhVwPY 0 Retweets Fri Apr 24 13:39:51 +0000 2015	 Ashwini Kaikini ashkaikini RT @Nriyandhi: The earth & I have a lot in common. We're made mostly of water & every year we both get hotter! ~ @TheEllenShow @ashkai... 1 Retweets Fri Apr 24 13:39:48 +0000 2015	 Abbi Abi_Frann RT @Martha_x15: @LisaVanderpump >Important reminder on #EarthDay. <a href="http://t.co/CiObnwtkm2<...loll">http://t.co/CiObnwtkm2<...loll 10 Retweets Fri Apr 24 13:39:46 +0000 2015	 1victoria bailey gabjamom1 RT @Martha_x15: @LisaVanderpump >Important reminder on #EarthDay. <a href="http://t.co/CiObnwtkm2<...loll">http://t.co/CiObnwtkm2<...loll 10 Retweets Fri Apr 24 13:39:46 +0000 2015	 gareth lee-bell gingerkermi RT @LuxeTiffany: Happy #EarthDay https://t.co/Wmectr8IVb (Victoria Falls, Zambia) 3 Retweets Fri Apr 24 13:39:41 +0000 2015
 Mauri Solages MauriSolages RT @ColleenNGRID: Gearing up! 0 Retweets Fri Apr 24 13:39:41 +0000 2015	 Rita Rin2379 RT @findmeaifit: #Competition 0 Retweets Fri Apr 24 13:39:41 +0000 2015	 EZstorit.com ezstorit RT @PublicStorage: Inspired by 0 Retweets Fri Apr 24 13:39:41 +0000 2015	 Claire Basiloña ClaireBasilona Calling on Young People to Lead Us 0 Retweets Fri Apr 24 13:39:41 +0000 2015	 Burke County Georgia BurkeCountyGA Master Gardeners Plant Sale today at 0 Retweets Fri Apr 24 13:39:41 +0000 2015

Figure 23: Tweets Wall Day3

Title Here

created on the command line with [twarc](#)


























 aaron k. snidesnark RT @EricWolfson: Best. #ClimateChange. Tweet. EVER. (via @ScottWesterfeld) #ClimateAction #EarthDay #EarthDay2015 #EarthDayEveryDay http://... 1101 Retweets Sat Apr 25 15:14:48 +0000 2015	 Boulder Book Store boulderbooks RT @techtionary: Today #Boulder @alfalfasmarket B'Earthday #EarthDay Bash 11-4 https://t.co/yTlmVgHJm w/ @AsherBrewing @GrowingGardensB 2 Retweets Sat Apr 25 15:14:44 +0000 2015	 CHANDRA PRAKASH dv891798 RT @BrightHouseNow: April 25, Earthday Birthday, Central Florida Fairgrounds Join @1011wjrr & Bright House Networks! Lineup of bands: http://t.co/8gMoi5PpQ0 2 Retweets Sat Apr 25 15:14:44 +0000 2015	 Nita Pujiastuti nee_ta15 RT @tupperwareID: Sebelum mulai menanam #mangrove, jangan lupa berfoto bersama :) #EarthDay http://t.co/8HW81UZAhA 3 Retweets Sat Apr 25 15:14:39 +0000 2015	 CHANDRA PRAKASH dv891798 RT @BrightHouseNow: TODAY- Earthday Birthday, Central Florida Fairgrounds Join @1011wjrr & Bright House Networks! Lineup of bands: http://t.co/8gMoi5PpQ0 1 Retweets Sat Apr 25 15:14:30 +0000 2015
 रविन और Pandit ji dheerajrazdan @AgentSaffron #Nepal #NepalEarthquake God have mercy on the Nepalese population #IndiaWithNepal #EarthDay #earthquake 1 Retweets Sat Apr 25 15:14:29 +0000 2015	 Nita Pujiastuti nee_ta15 RT @tupperwareID: Ayo kita tanam 10.000 #mangrove sebagai bentuk cinta kita pada lingkungan bersama WWF & Tupperware! #EarthDay http://t.co/http://t.co/YwMD39TbwD 3 Retweets Sat Apr 25 15:14:29 +0000 2015	 haltonmastergardener HaltonMGs RT @henhere: All set up at Georgetown Market Place for #EarthDay Fair, till 4pm. @ImagineHH http://t.co/1DcdT4er05 1 Retweets Sat Apr 25 15:14:14 +0000 2015	 TheEcoVeganEvents TomasinaBarr RT @missrainbird: This #EarthDay, watch Cowspracy online for just \$1 and learn the truth about who's killing the planet! #documentary http://t.co/Dp28uvT96U 1 Retweets Sat Apr 25 15:14:08 +0000 2015	 Robert Rossachaj RRossachaj RT @the_sca: SCA CEO Jaime Matyas greets 400+ volunteers at today's #EarthDay service project w/ @AmericanExpress in #DC. http://t.co/0yhnkn... 2 Retweets Sat Apr 25 15:13:58 +0000 2015
 Oak Hills Sports OakHillsSports RT @kenmeibers: Giving back to the community. #earthday http://t.co/LpwwQ74Gyx 3 Retweets Sat Apr 25 15:13:57 +0000 2015	 Ahmed Numen AhmedNumen Sad to hear the terrible news of massive earthquake hit in Nepal. My prayers and thoughts are with the people of Nepal #EarthDay #Earthquake 0 Retweets Sat Apr 25 15:13:53 +0000 2015	 Primped and Improper primpdNimproper Pace Team Cleanup! #earthday #oakville #joshuacreek #realtor #realtorlife #realestate #century21... https://t.co/dg4De0l2Ox 0 Retweets Sat Apr 25 15:13:42 +0000 2015	 money honey g_money_honey29 RT @Martha_x15: @LisaVanderpump -Important reminder on #EarthDay. http://t.co/CiObnwtkm2<...lol 58 Retweets Sat Apr 25 15:13:40 +0000 2015	 Crystal Coast CrystalCoast_NC The #CrystalCoast #EarthDay Celebration is going on @ #FortMacon. Enjoy FREE kids activities, demos & more, 10am-2pm. http://t.co/J9vXiOnKsC 0 Retweets Sat Apr 25 15:13:38 +0000 2015
 Yenny T UNSWstudentleif RT @ThePatriot143: CLIMATE ACTIVISTS IN MELTDOWN AS 'SCEPTIC' GETS AUSTRALIAN GOVERNMENT FUNDING http://t.co/mv4xisaWuG #EarthDay 1246 Retweets Sat Apr 25 15:13:36 +0000 2015	 Estuk T_F_L RT @reason: News flash: #NorthKorea wins #EarthDay for record 45th time in a row! http://t.co/JALvCO3WT HT @Neoavataara http://t.co/MUynUG0... 1246 Retweets Sat Apr 25 15:13:24 +0000 2015	 Angelo Waglio_O RT @AlexineAudiard: #EarthDay http://t.co/gmoF0DireT 36 Retweets Sat Apr 25 15:13:24 +0000 2015	 Goran Kolasinac onikapova RT @MariaSharapova: Everyday is #EarthDay #HugATree http://t.co/SS2svT2AmA 424 Retweets Sat Apr 25 15:13:23 +0000 2015	 Ferguson-Florisant FergFlorSchools Great crowd @ our Little Creek Nature Area celebrating #EarthDay big thanks to @missioncontinue for building seating http://t.co/hPUtDsyAGJ 0 Retweets Sat Apr 25 15:13:20 +0000 2015
 Kim Rees kimba420 RT @WeeLittleHawk: @Stantec 	 li'l gal with aGlock lilglockgal Al Gore fact #EarthDay 	 mercedes.m LuisaM56_ RT @AlexineAudiard: #EarthDay 	 Warsame Abdi warsameeabdi RT @Stanford: Stanford's new 	 Barrett Henry Team bhenryteam Pace Team Cleanup! #oakville

Figure 24: Tweets Wall Day4

Title Here

created on the command line with [twarc](#)

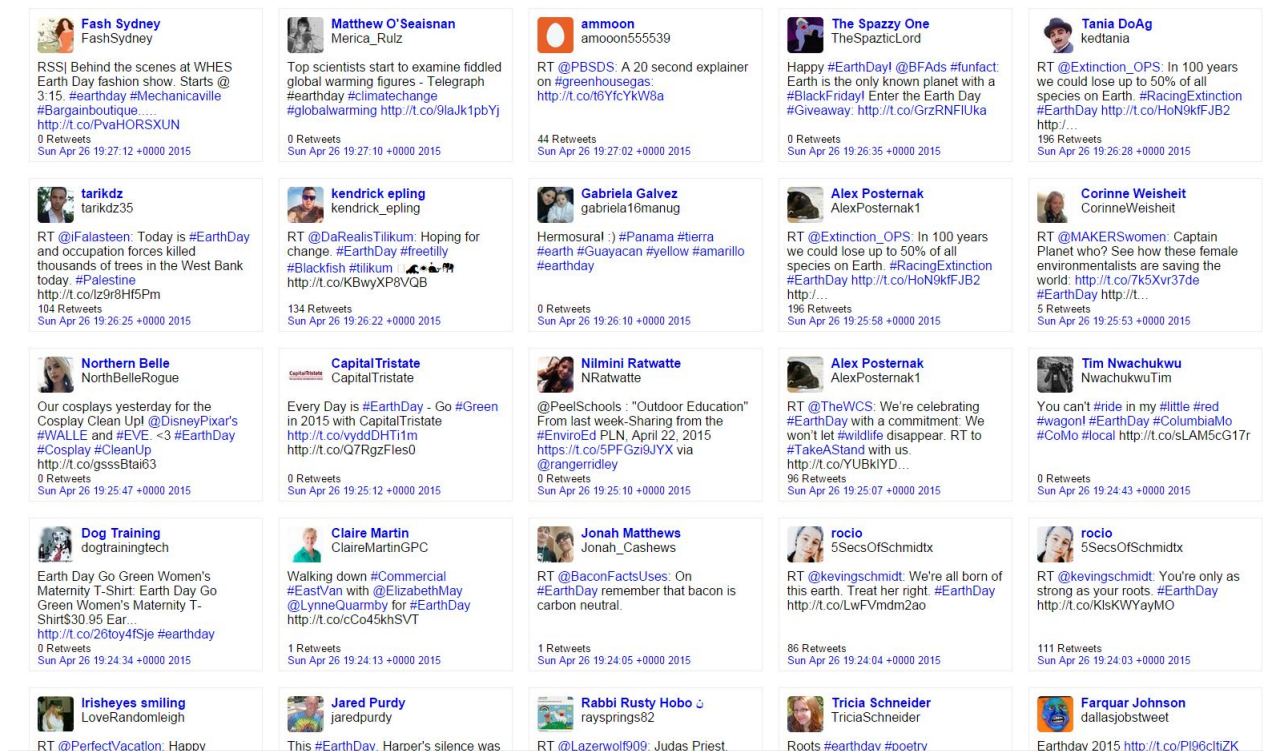


Figure 25: Tweets Wall Day5



Figure 26: Tweets Word Cloud Day1

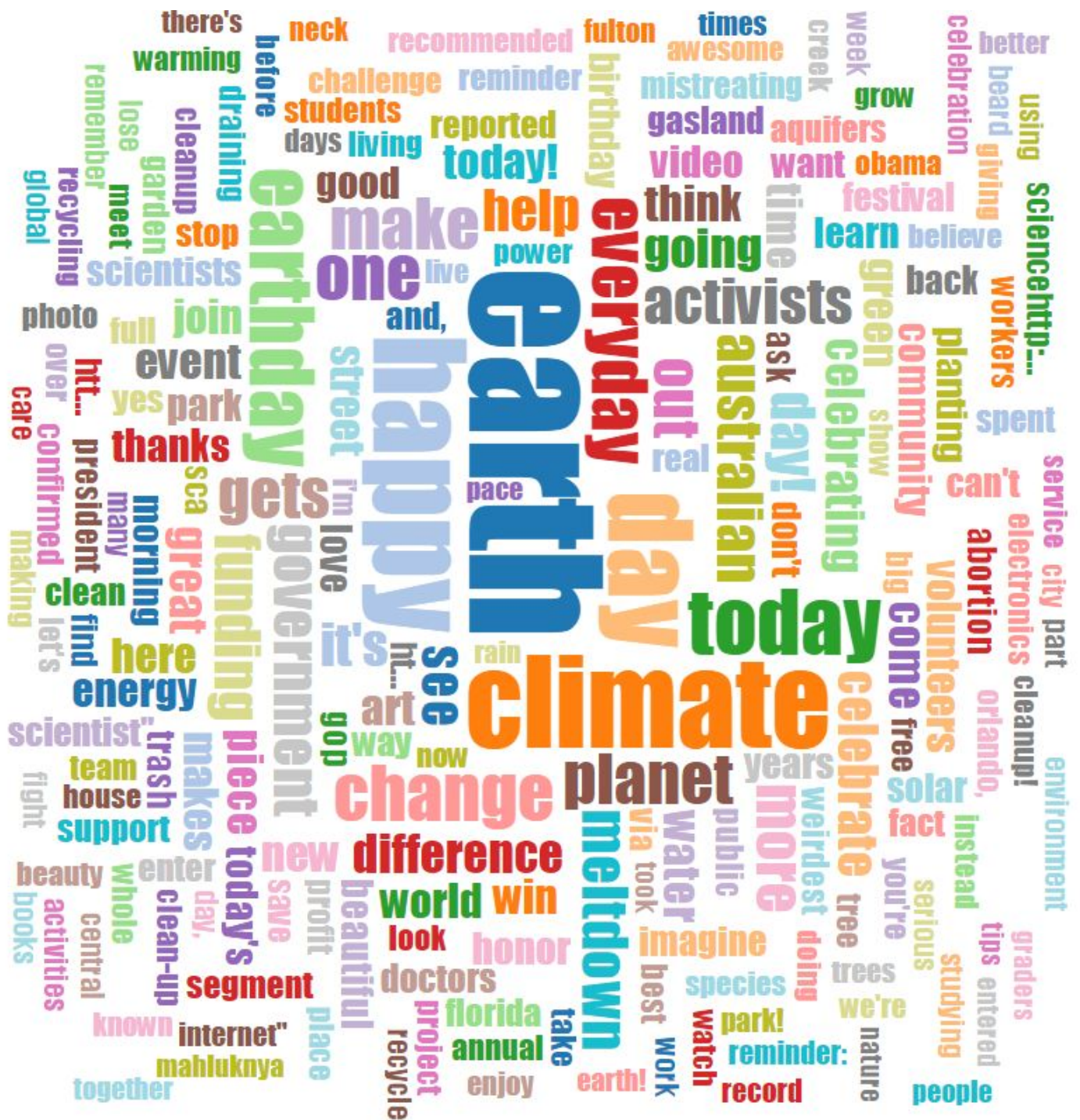


Figure 29: Tweets Word Cloud Day4

Bibliography

- [1] jusText. <https://pypi.python.org/pypi/jusText/2.1.0>.
- [2] Wordcount. <https://github.com/mlafeldt/google-python-class/blob/master/basic/solution/wordcount.py>.
Accessed: 2015-04-03.
- [3] Zipf. <https://plus.maths.org/content/mystery-zipf>.