Open University of Cyprus

FACULTY OF PURE AND APPLIED SCIENCES

Department of Social Information Systems

student: **Vasileios Basdras**
Master Dissertation

Supervisor : Klimis Ntalianis

Corporate Reputation In The Online Universe

Geneva, Switzerland

December 2018

Abstract

This dissertation is made for Social Information System Master at Open University of Cyprus. It's focus is Online Reputation of business. Starting from an individual to a multinational corporation level we are looking to find, understand and define the purpose, the need and the impact of managing the impressions made. We argue that strategies may vary regarding the type of business. We suggest that not only consumers are allowed to have an opinion about a brand and how the reputation is disseminated. The progress in communication is implying bigger social influence between the people, trends are changing fast. We are presenting different tools to find these insights and monitor the ORM, free, paid or open source. We are discussing the caveats anyone should be aware of. Finally we aim to use code in order to analyze and to see it in practice through the prism of a real business entity.

Keywords

corporate reputation ; ORM ; online reputation management ; brand ; sentiment analysis ; Insights ; social media analytics

Acknowledgments :

First and foremost, I would like to thank my parents, who helped me, see the world with my own eyes. Athina's linguistic tips are greatly appreciated. Thanks to supervisor Ntalianis Klimis and the department of Pure and Applied sciences of Open University of Cyprus. Last but not least my partner Indra who did not assist me directly in the writing of this work but in everything else.

RESUME

Vasileios Basdras (Βασίλης Μπάσδρας)

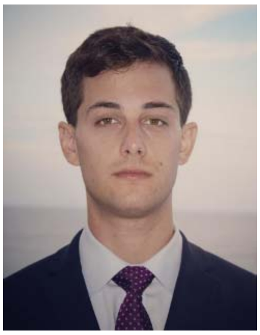
	Vasileios Basdras. Born in 1987 in Athens. Graduated from Sport Science and Physical Education B.Sc in Athens Kapodistrian University. Was admitted in Social Information Systems M.Sc. In Open University of Cyprus. Living and Working in Geneva, Switzerland. **https://www.linkedin.com/in/billybasdras/** **https://github.com/biltAtous**
Key Competences: -Programming – Python -Data analysis – Excel, Python – Math and Statistical Background – Google Analytics -Linux Administrator -WEB –HTML,CSS,SQL –CMS(Wordpress, Drupal) – SEO	Languages : Greek(native), English(full proficiency), French(Advanced)

TABLE OF CONTENTS

Use Ctrl+Click to follow the links

INTRODUCTION

Reputation in business is probably dating since the first day of business. If a business wish for recurring clients the best way is to offer worth buying services or products in the most systematic way possible. You can imagine every possible situation of which the person who sells is not doing it only for a single time (scam). From a seller in an ancient market place, all the way to the biggest modern corporations, like Amazon or Apple. So in other words positive reputation can increase the likelihood of the business being successful. Not only businesses have a reputation but also singers, politicians etc. In fact besides people with a public image, all the individuals have a reputation to care for. So if reputation is not only addressed from a business to a direct client. *What is it (reputation), why we care for it and who is addressed to?*

People we are living in highly organized structures. On our daily lives we depend on each other in many different levels. We are social creatures. We exchange information about all sorts of things. Apparently we also share vital information like where is the spring with the fresh water to drink or less vital, who is selling the best bread in neighborhood. We tent to influence each other through that process. Big recent advances in technology adds even more power in our communication needs. For some the most significant advances in terms of communication were the typography invention and the Internet. Now, more than ever, we can communicate as much and as often as we wish with other individuals wherever in the planet. Therefore we can disseminate information or information about reputation with an enormous rate. People can post about products and influence other people instantaneously and for as long as the review is visible.

A story that is trending right now in France (21 Septembre à Marseille 2018) is about a blind man that went to the supermarket with his guide dog. The supermarket policy is “No Dogs”. The blind individual was escorted outside with his dog, by the owner, denied his right to shop. [41] For owner’s and the supermarket’s bad luck, they were filmed. Soon the video became viral. The famous supermarket chain “Monoprix” made the following statements to address the situation.

"*We are sorry for this incident, as is the store manager who saw the young man that afternoon apologized. For sanitary reasons, animals are not accepted in our stores, dog guides for blind are obviously an exception.* " - Monoprix

“*Monoprix strongly condemns the facts that took place in Marseille and apologizes. Engaged for several years in the fight against discrimination, our teams are trained to welcome all audiences and have been sensitized again*” - Monoprix

(Twitter statements from Monoprix account. Translated from French).

So reputation can be seriously afflicted as well. In a gereric way we could say that one has to Built, Maintain or Recover the reputation of his/her brand.

One of the points made was that reputation can be affected. The other one was the power of technology that allowed the story to become viral. From the standpoint of the one trying to build, maintain and recover reputation, how technology can be utilized? Online reputation management (ORM) strives to offer a complete solution for companies that have to deal with information coming from different sources. But all that is relatively new. This infancy is responsible the appearance of Big Data, simply because we could not solve a problem that we never had. There are rules of thumb but not applied science yet. Together there are also a lot of social phenomena yet to explained, technological limitations and other kinds of caveats that should be acknowledged.

Chapter 1
Description of the paper

1.1 Purpose

The current document is a dissertation for the Master Degree, social information systems. It is held by the faculty of pure and applied science of Open University of Cyprus. It is looking online reputation from the standpoint of a business. A rather interdisciplinary approach could be useful for this purpose. Ideas from economics, social science, marketing and computer and network science are going to be incorporated on that note. We try to identify the dimensions of corporate reputation and brand reputation, dive deep into the current tools to deal with data regarding the brand or the corporation and actually study a real case.

1.2 Structure

In the second chapter we are setting the definitional landscape. This chapter is essential because it's providing useful definitions and important links to literature. But also it is infrastructure of the rest of the paper. To these ideas is based the thinking process of everything that follows. In chapter three we are presenting some of the tools, ways that a business can manage their reputation. Some of them are paid, some are free and some open source. Therefore a different amount of money, time and knowledge is needed to engage amongst them. In the following chapter, number four, we choose Rolex as our case study. So we briefly present some facts about the company and we develop our strategy to acquire data from Amazon, Twitter and Youtube. Some technical details are in that chapter. Because in the following chapter number five resides the analysis we did on that data of chapter four. In chapter six, we discuss about all the things to be aware of. Different technological limitations, disclaimers or caveats. But also restrictions because of our ways, social norms, capitalistic systems, we discuss the design of things. In the seventh and last chapter we are presenting our concluding thoughts, we are contemplating future work, proposals for the company. We finish traditionally with the conclusion followed by the bibliography and the appendix.

1.3 Resume of Implementation

In the making of that paper we used linux, ubuntu as operating system. Therefore the open office for our office requirements (Libre calc and Libre writer mainly) . There was a lot of terminal work and some linux scripting. Browsing was made by Mozilla Firefox. We needed to use the Youtube and Twitter API for downloading the data. Most of the work has been done with Python. To scrape, download, authorize, manipulate, save in a file, import, calculate and produce visualizations. The libraries we used are in alphabetical order: apiclient, argparse, csv, html, io, json, math, matplotlib, nltk, numpy, os, pandas, re, requests, scipy, skleanr, stopwords, textblob, time, tweepy, unidecode, urllib, urlparse, wordcloud. We also used GIMP (GNU Image Manipulation Program) and gedit as a text editor.

Chapter 2
In Search of Definitions

2.1 Self Presentation

Erving Goffman in his work, *The Presentation of Self in Everyday Life uses the imagery of the theatre in order to portray the importance of human social interaction (Goffman E., 1956) . In a way, in everyday face to face (f2f) communication we are trying to manage the impressions we create to our interlocateurs. Why would that be? In the 'pathetic dot theory' (Lawrence Lessig, 1999) there four constrains that limit our behavior and ultimately our individuality. These are: architecture, markets, law and norms, while the pathetic dot is us, helpless to avoid their forces. Focus on the last constrain, norms. People have to behave under certain ways.*

Since these norms are not written, how can we behave in an appropriate way? First of all people are grouping themselves with people with similar attributes, in social science this concept is called Homophily. There are two ways that this is facilitated, selection and social influence or socialization (Easley D. And Kleinberg J. 2010). The first one, selection, is related more with immutable characteristics, such as race or ethnicity. On the contrary social influence serves to shape the characteristics of people.

In his later work, Goffman introduces the "Face" which is a sociological concept for an individual's public self-image (Goffman E,. 1967). Face is interwoven with the concepts of pride, dignity and honor. Saving Face is an expression, often used, that signifies a desire or defines a strategy to avoid humiliation or embarrassment, to maintain dignity or preserve reputation. A faux pas is a social misstep or tactless act or remark in a social situation. When, for example, Silvio Berlusconi referred to US President Barack Obama as "suntanned". This faux pas is based on political correctness but the range of situations is very broad. Surely at some point in time, we all found ourselves in this position.

As expected the social complicatedness that has been inherited to us is exerted in our online behavior as well. In CMC (Computer Mediated Communication) the users are trying to manage the impression to facilitate the desirable relationships (Walther 2007). This communication offers different ways of deception though. It can be identity-based or message-based and of course a combinations of both (Hancock, J.T. 2007).

Even though we all use them, in almost every communicational exchange, the individualistic incentives for someone to protect his/her face, are not so implicit. This is an individualistic way of dealing with reputation. In a corporate level, reputation could affect the health of the business. But what is it really the corporate reputation?

2.2 Corporate Reputation

In this section we will focus on the work of Barnett, Michael L. and Jermier, John and Lafferty, Barbara A., Corporate Reputation: The Definitional Landscape . This paper has a focus on the definition around Corporate Reputation, therefore we will hold their meta-definition of Corporate Reputation.

The vocabulary problem is called the one that different people are talking for the same thing using different words (G. W. Furnas and T. K. Landauer and L. M. Gomez and S. T. Dumais, 1987). However the popularity of the term ”Corporate Reputation” in indexed peer-review articles has been growing (Barnett, Michael L. and Jermier, John and Lafferty, Barbara A., 2006).

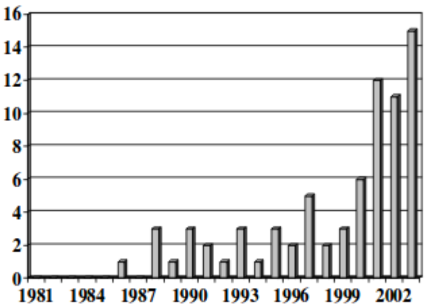


Illustration 1: Indexed peer-review articles containing ‘corporate reputation’ in title or ab-stract.

Although people can be confused between corporate ***Identity,*** corporate ***image*** and corporate ***reputation.*** In the literature review, of the mentioned ones, those are certainly distinguishable. The corporate identity can be seen as the underlying core or basic character of the firm (Melewar and Jenkins, 2002). Image is a general impression of a corporation's distinct collection of symbols, whether that observer is internal or external to the firm. Image is 'what comes to mind when one hears the name or sees the logo' of a particular firm (Gray and Balmer, 1998: 696). The transition from identity to image is a function of public relations, marketing and other organizational processes that attempt to shape the impression people have of the firm. But image can also be shaped but not controlled by an organization because of factors such as media coverage, governmental regulations and surveillance, industry dynamics and other external forces also influence impressions of the firm. Which brings us to the definition.

Corporate Reputation: Observers Collective judgments of a corporation based on assessments of the financial, social, and environmental impacts attributed to the corporation over time.

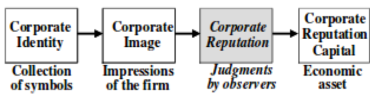


Illustration 2: Disaggregating

It is important to recognize that the identity of a firm can remain static while its image and reputation change as a result of external events. As judgments of the firm accumulate over time, reputation capital ebbs and flows. This is the economic and intangible asset quality that is often attributed to reputation.

2.3 A Different Approach To Business Reputation

Reputation Institute is a global reputation research and advisory company, headquartered in Cambridge, MA. It was founded in 1997 by Charles Fomburn. We are not affiliated with this organization in any way whatsoever and even though is a private and profit driven organization (not academic), we just gonna borrow their definition of Reputation.

They are selling an algorithm of which the premise is to provide useful insights regarding reputation of business entities, it is called RepTrak. We are not going to evaluate the correctness and the quality of the results of this intelligence system. We are going to enumerate the dimensions that this algorithm is said to be based on and we also going to present a list of the 100 best ranked (image below) companies for 2018. (https://www.reputationinstitute.com)

Following the 7 dimensions of Reputation :

1. **Products/Services**
Do you deliver on a world-class experience? High-quality Products and Services can profoundly shape reputation.

2. **Innovation**
Is your company static or dynamic? Forward-thinking and creatively-inspired companies have a reputational advantage.

3.**Workplace**
Corporate culture directly impacts recruitment, retention, and talent acquisition. Positive perceptions of a workplace can help you achieve employer of choice status.

4.**Governance**
Can your company be trusted to do the right things when no one is looking? Practicing good governance is key in earning trust in times of crisis.

5.**Citizenship**
How does your company align with social values? Being a good corporate citizen has a positive impact that helps to make the world a little better.

6.**Leadership**
Companies with executives who align brand purpose with daily business activities outperform those focused solely on financials.

7.**Performance**
Financials matter, but it is important to link your financial success with positive social impact to maintain a license to operate.

It is clear that the first point should be precedented over the rest of them. Let's think of a company that belongs in a new space. So new that all none of the company in the space has any reputation whatsoever. Then it is only the first point that stands since there are no indications/data for the rest of them to decide. In particular let's consider the medicinal cannabis market which is relatively new in the commercial world, in Europe and the United States. Someone, that is of course over 18 years old and therefore eligible to buy and smoke cannabis is presented with an array of different brands to choose from, without any clues about these companies. How he/she can decide ? Apparently the only way is to try everything and see what works the best. After he/she makes up his/her mind he might share his views with other potential customers.

The priotity of the rest of the points can be argued. But let us focus on third point, workplace, for a while. Today is more and more important to be a fair employer and to treat everybody the same regardless, age, race, sex and other attributies. As a matter of fact companies always try to include messages like these in their job announcements.

“We are an equal opportunity employer and value diversity at our company. We do not discriminate on the basis of race, religion, color, national origin, gender, sexual orientation, age, marital status, veteran status, or disability status.” - SeedStars through linkedin

In another job ad in linkedin this time from P&G, regarding the innovation point.

“P&G is a leading global consumer goods company whose winning brands are built around the model of innovation. Whatever your passion is, we want to ignite your potential to become your very best self. We hold true to our purpose, values and principles as we seek to make a difference in the world around us. You will engage in meaningful work that will touch the lives of others and have a real impact. Everything at P&G starts with understanding - understanding our consumers and our employees as we innovate to improve lives now and for generations to come.”

It seems that organizations are aware of these points. Even though both of the small quoted texts are not addressed in the end consumer they try to create an impression of the company they would like either to be or to make us think that they are. As a matter of fact these texts are addressed to a person looking for a job in linkedin but they are trying to curve their reputation regardless.

We are now left to see what is the online corporate reputation.

2.4 Online Reputation

First of all is very marketable! People try to sell services or tools that promise to their clients effective ways to manage their reputation (Online Reputation Management, ORM). We will present several visualisations of their efforts to explain Online Reputation, without endorsing any of their products or methods of operation. For now we keep our focus on ‘what-is’ Online corporate reputation, later on we will present other tools, some of them are open source.

Illustration 3: kvrwebtech.com/blog/why-orm-is-critical-to-the-success-of-a-business-todav/

Illustration 4: the-pixel.com/solutions/online-

Social media, Review Site and Forum exist in all of the depictations. The last one argues that Search Engine Results are a part of the Online Reputation. Again these are efforts to sell and since SEO(Search Engine Optimization) is a best seller, Search Engine Results can find their place in the pie. What is argued really here are the sources of data that the tool can access and evaluate in order to 'manage' the reputation. But does the search engine need to use an evaluation system for the webpages that affects the search results?

Google's PageRank is one of the most famous algorithms of the recent times. A lot of Google's success has been attributed to this very algorithm. As Stanford teaches in their students in one of their handouts (https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf), the PageRank algorithm gives each page a rating of its importance, which is a recursively defined measure whereby a page becomes important if important pages link to it. This definition is recursive because the importance of a page refers back to the importance of other pages that link to it. One way to think about PageRank is to imagine a random surfer on the web, following links from page to page. The page rank of any page is roughly the probability that the random surfer will land on a particular page. Since more links go to the important pages, the surfer is more likely to end up there.

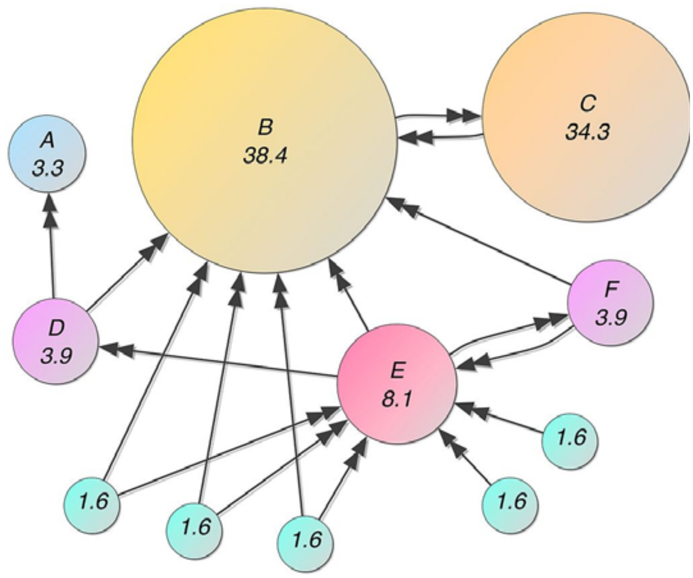
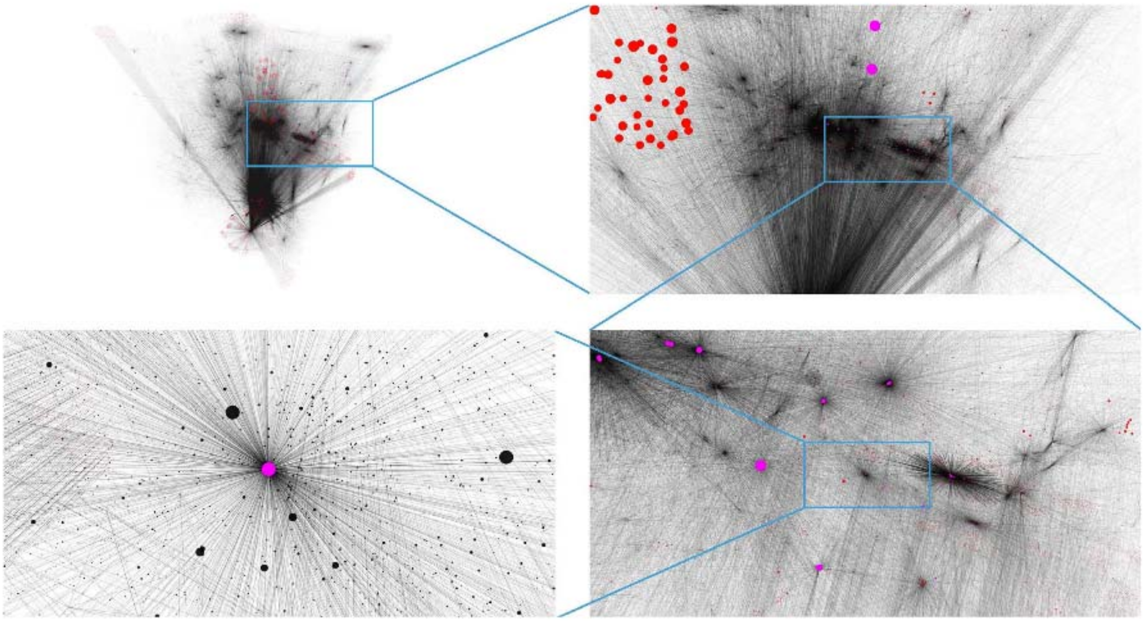


Illustration 6: Mathematical PageRanks for a simple network.

The World Wide Web is a network whose nodes are documents and the links are the uniform resource locators (URLs) that allow us to "surf" with a click from one web document to the other. With an estimated size of over one trillion documents ($N \approx 10^{12}$), the Web is the largest network humanity has ever built. It exceeds in size even the human brain ($N \approx 10^{11}$ neurons).

Even though (look illustration 7) it looks random, after a closer look we can see that some nodes (webpages) are so much bigger (linked) than others. In fact it expressed the best by a power law distribution (Al. Barabasi, 2016). There are some webpages that they act as hubs and they are linked so much more than others. For some reasearchers that creates a rich getting richer phenomenon whereby the stronger webpages become even stronger. To get a perspective people have asked to compare their average number of friends to the average of their friend's friends. Our friends have always more friends than we do! That is known as the friendship paradox. Of course it's a sampling bias. We always choose to compare with our most popular friends, these are the ones to come first in mind.



As how it has been defined a few paragraphs above, the PageRank declares the importance of a webpage. Although not viewed as such, PageRank may be thought of as a way of rating the "reputation" of web sites.(Zhang H., Goel A., Govindan R., Mason K., Van Roy B., 2004).

Finally we have to underline how people consider the offline and the online to be two different worlds. It's useful to consider them as a dipole and compare differences but they are also highly interwoven. They are many bloggers and small businesses that would not arise if it was not for the WWW. Although power and reputation in the physical or offline world can be translated in to something of the same value in the online world. We are consider for instance The Economist to be a trustworthy source of information, the fact that now we purchaise it online does not change much. Having more inbound links that improving one's PageRank could mean having to spend more? Well established companies strive to pass to the online game, spending a lot for this transition. The search engines raise not merely technical issues but also political ones. Search engines systematically exclude (in some cases by design and in some, accidentally) certain sites and certain types of sites in favor of others (Introna, L.D. and Nissenbaum, H. 2000) .

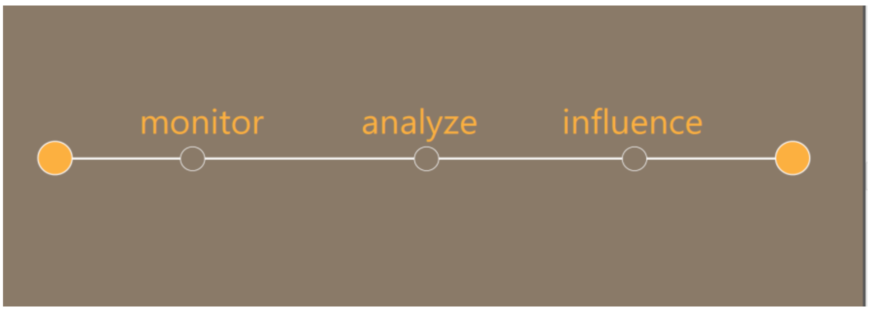
2.5 Online Reputation Management

We spend some pages on what is Corporate Reputation and online reputation. But what is management though? There are three basic aprroaches :

Building – This type of reputation management has to do with building the reputation for a business that is just getting started. It includes building a good reputation to maintaining it for your business.
Maintenance – Reputation management meant to just keep a company's good image superior in the public eye is called maintenance. This is meant for companies that are already established, and have a good reputation already.
Recovery – If your business has gotten a bad reputation for any reason, then the recovery portion of reputation management is for you.

As Warren Buffett once said:

Warren Buffett — '***It takes twenty years to build a reputation and five minutes to ruin it.***'



And it can be held in three stages.

2.6 Type Of Company Or Product

Markets are where Adam Smith's famed "invisible hand" determines the process that balance supply and demand. The 'visible hand' -'the hierarchies' is what we typically think of as firms or organaizations. Another book 'the visible hand' of Alfred Chandler traces the transition from the markets to hierarchies, from the early 1800's through the late 20th century (Arun Sunjararajan 2016).

Nowdays there are online platforms that they are P2P (peer to peer) producing Billions of dollars annually. It's also called the sharing Economy which is crowd-based capitalism. It's facinating how in our days the world's largest taxi firm, Uber, owns no cars. The world's most popular media company, Facebook, creates no content. The world's most valuable retailer, Alibaba, carries no stock. And the world's largest accommodation provider, Airbnb, owns no property (https://www.independent.co.uk/news/business/comment/hamish-mcrae/facebook-airbnb-uber-and-the-unstoppable-rise-of-the-content-non-generators-10227207.html). Inside these platforms are individuals that in a way they are micro-enterpreneurs. Therefore these companies can be considered market-hybrids. Let's consider airbnb, the recommendation system refers to the host, now the host can also write recommendations regarding the hosted ones. This creates a community that promises to find appropriate homes to travelers. This is what essentially sold by airbnb, rather that spaces that been sold by the individual hosts. As Vitalik Buterin influencial cryptocurrency writter and founder of Etherium in April 2014 : 'It is important to first understand that, in the space of tech companies and especially social networking startups, a large number of them are literally backed by almost nothing but social consencus' thereafter he explains : 'Theoretically is entirely possible for all of the employees at Snapchat, Tinder, Twitter or any other such startup to all suddenly agree to quit and start their own business, completely rebuild all of the software from scratch within months, and then immediately proceed to build a superior product. The only reason why such companies have any valuation at all is a set of two coordination problems: the problem of getting all employees to quit at the same time, and the problem of getting all of the customers to simultaneously move over onto the new network'.

That being said, there is this new way of doing business today (sharing economy) that is expected to grow significantly the years to come. Some of the above businesses we mentioned they do tranditional things translated in a modern digital sharing economy kind of way. Like for instance, Uber, airbnb and Tinder. They offer something that existed before like accomodation, friends or a ride but in a completely new way.

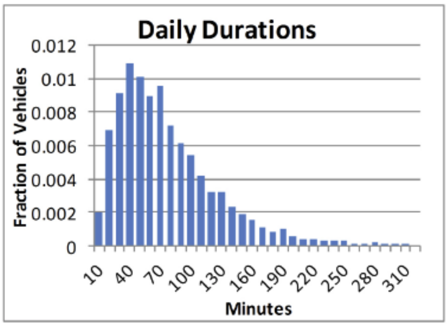


Illustration 8: vehicle usage, USA. NHTS 2009

In the image you see how much americans are using their cars per day. The rest of the cars are not utilised. In fact Americans spend about one trillion dollars annually purchasing new and used vehicles. To put it in context GDP in 2015 was 17 Trillions (Sunjararajan 2016). The Capital Impact is changing for the cars are used more. Of course you might think how that can affect the traditional business like taxis (considering Uber) or how it changes the way earn money but this is outside of the scope of this chapter.

Lisa Gansky in her book The Mesh, is proposing two dimensions that might evaluate a product before determining whether a peer-to-peer rental platform for it might emerge. How valuable is the product (cost) and how internsively is used by the owner (frequency). A product that is not used very intensively by an owner like a car is a good prospect for renting. On the other hand if the product is of minimal value the coordination costs would not made sense to be paid as the person (renter) could easily afford to buy a new one. This is a generally good idea. Following that logic renting out a luxury watch, a Rolex for instance makes a lot of sense. Since it cost a lot of money to buy a new one and it can often be used for only a few hours per week of month by their owners. As a matter of fact a company called ElevenJames was launched in 2013 in New York. Their premise was to borrow luxury watches to their clients based on a subscritption program. Today this company strives to stay alive (https://www.bloomberg.com/news/articles/2018-08-23/**what-s-happening-to-subscription-watch-club-eleven-james**). It might have been the subscription model or bad management but we believe that there is an ownership effect to these products. In other words the important thing in wearing a Rolex watch is that you can afford it. It provides a certain prestige or credit to the owner. Who is trying to manage the impressions that he creates to the rest of indivuals he deals with (see Goffman impressions previous chapter). Similarly a wedding ring cannot be borrowed. Even though is a valuable item, most of it's value is attributed to a person buying for you. Therefore the type of the product or service that a company is offering plays a decisive role.

2.7 The Pampers Paradox

Barriers to entry are factors that prevent or make it difficult for new firms to enter a market. There are many factors that manifest this phenomena. One of them is trademarks consolidated in the market. Entering a market with prestigious and established brands is extremely difficult to establish. Some branded products are so associated with product itself that in end we reffer to them by their brand names. Therfore instead of diapers we say Pampers, instead of insulated luquid containers we say Thermos, instead of praline cream we say Nutella etc. Some of these companies are introduced to economics textbooks as canonical examples of barriers to entry.

Lets consider an individual of thirty years old. This person has been using pampers as an infant. Although he does not have any memory of using the product he is consider Pampers to be an excellent product. Even though he could be potentially a father, therefore a client, he would not order online at Pampers website (thefore will not be prompted to answer a questionnaire) and would never visit another website to contribute his opinion about the product. Even if he saw an ad he would probably buy them at the local supermarket. And in that sense P&G (Procter and Gamble) who owns the brand has to compete about that shelf space in that supermarket and for this particular client the only way to be prompted to answer a survey would be on the supermarket floor.

As a matter of fact, often times people watch ads online and then purchuse the products from the official website or offline, rather the online ad. Most of these ad deals, like google adsense, they are paid per click(PPC). If the beholder does not follow the link the advertiser will not get the money. According to MIT technology review Google bought in 2017, 70% of all the credit and debit card transactions and tried to match them to people's online profiles (https://www.technologyreview.com/s/607938/google-now-tracks-your-credit-card-purchases-and-connects-them-to-its-online-profile-of-you/).

Just to set things straight, Pampers has a website to sell products and gather reviews (https://www.pampers.com/) and also runs online campaigns. Other online stores sell pampers as well. The paradoxical fact is that the person have tried the product but has no memory of it, however feels strongly about it. There is no way to prompt a user that is not bying a product and does not wish voluntairily to express his opinion. Unfortunately we do not have the data to justify how much by percentage happens online by their website, online by other seller and offline. How that changes over time? Also everyone who reads the present text should be aware that reviews and surveys are two different things. All of the above are just a happy concidence and some ideas about barriers to entry and well-established bransds. People can have opinions of any kind about products that never used of about to use. Think about Rolex watches, we all think is a great company, even though we might not appreciate the style enough to actually wear them.

2.8 Organization vs Brand vs Product

An organization represents a company itself that can be comprised by several brands and products. A brand is a name used by an organization or business person for labeling a product or group of products. A product is an article or substance that is manufactured or refined for sale. SKU (stock keeping units)

From wikipedia :

Company : “A company, abbreviated as co., is a legal entity made up of an association of people, be they natural, legal, or a mixture of both, for carrying on a commercial or industrial enterprise. Company members share a common purpose, and unite in order to focus their various talents and organize their collectively available skills or resources to achieve specific, declared goals. Companies take various forms, such as:

voluntary associations, which may include nonprofit organizations
business entities with an aim of gaining a profit
financial entities and banks

Brand: “A brand is a name, term, design, symbol, or other feature that distinguishes an organization or product from its rivals in the eyes of the customer. Brands are used in business, marketing, and advertising. Name brands are sometimes distinguished from generic or store brands.”

Product :”In marketing, a product is anything that can be offered to a market that might satisfy a want or need. In retailing, products are called merchandise. In manufacturing, products are bought as raw materials and sold as finished goods. A service is another common product type.”

The conclusion that follows:

Company →Brand →Product

The following passage (www.AssessmentDay.co.uk) illustrates the difference between brands and companies as well as the perception of indivuduals about it.

“Brand equity has become a key asset in the world of competitive business. Indeed, some brands are now worth more than companies. Large corporations themselves are widely distrusted, whereas strangely, brands have the opposite effect on people. Brands are used to humanise corporations by appropriating characteristics such as courage, honesty, friendliness and fun. An example is Dove soap, where a dove represents white, cleanliness and peace. Volkswagen like to give the impression through their advertising that they are a reliable, clever, technical product. In a sense, rather than the product itself, the image and the idea are the selling point.”

Therefore in the example above stands : Unilever→Dove→Dove soap (Company →Brand →Product). It is also pointed out that people tend to distrust organizations but they have to trust brands, in order to use the products. Remember though as *Warren Buffett* said — '*It takes twenty years to build a reputation and five minutes to ruin it.*' Even though the concept with the dove is quite clever Dove recently presented an ad that showed a black woman turning herself white. (https://www.washingtonpost.com/news/business/wp/2017/10/08/dove-ad-that-shows-a-black-woman-turning-herself-white-sparks-consumer-backlash/?noredirect=on&utm_term=.6b6f6ab93375)

A message from Dove's twitter account appeared thereafter : "An image we recently posted on Facebook missed the mark in representing women of color thoughtfully. We deeply regret the offense it caused." (Oct 7, 2017) But some things are hard to be unsaid. It makes you think how something so obvious did not occur to all the people involved. Starting from the girls on the picture, the photographer, the marketer who conceptualized and the supervisor who approved.

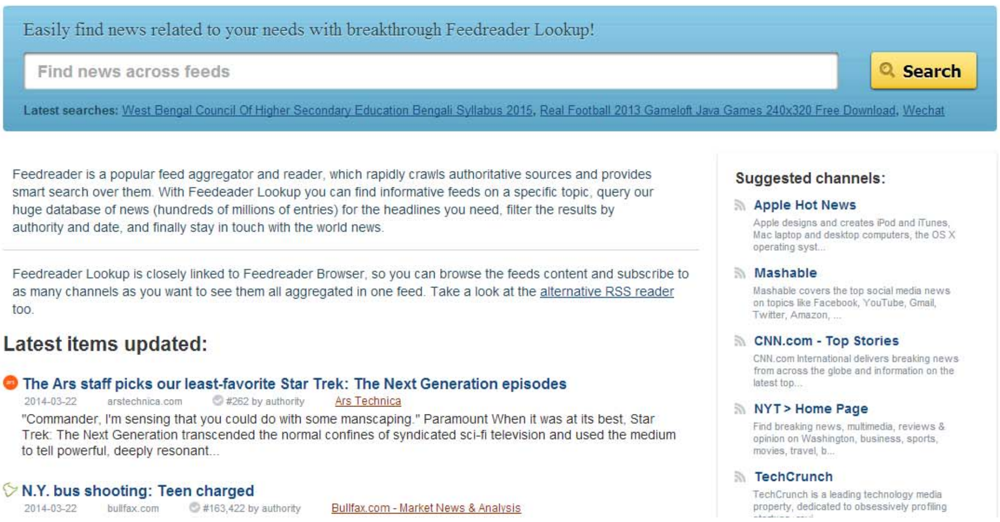
Chapter 3
Tool revision

3.1 Revision Of Some Online Tools For Online Reputation Management

There are several tools online at the time. The following is a small selection we choose.

Feed Reader

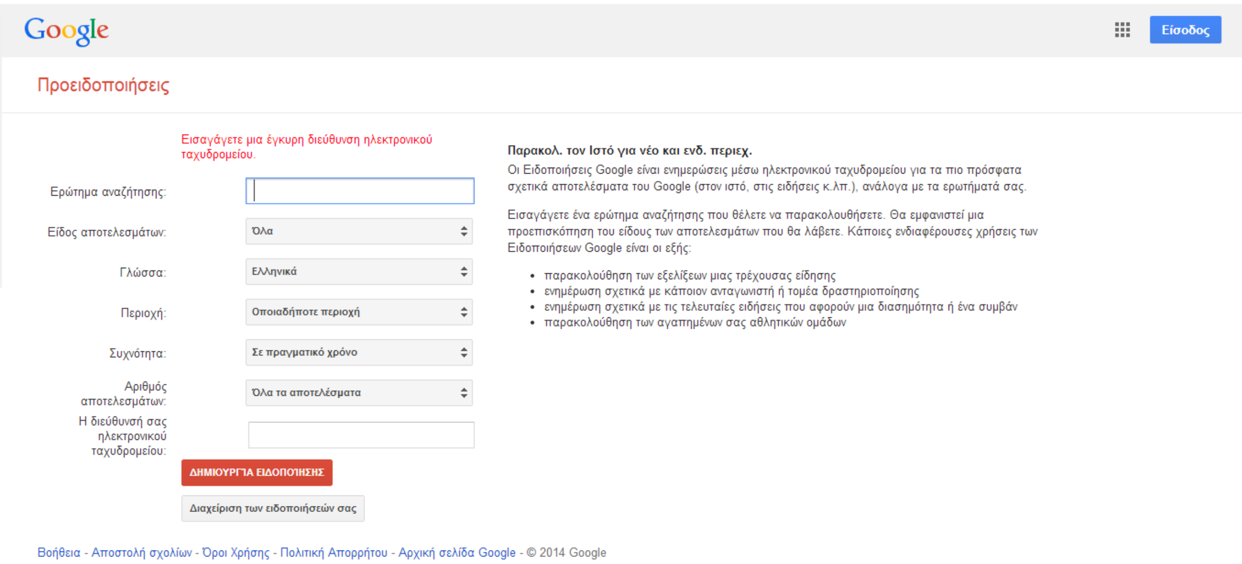
A relatively simple platform that has both free and paid options.



Google Alerts

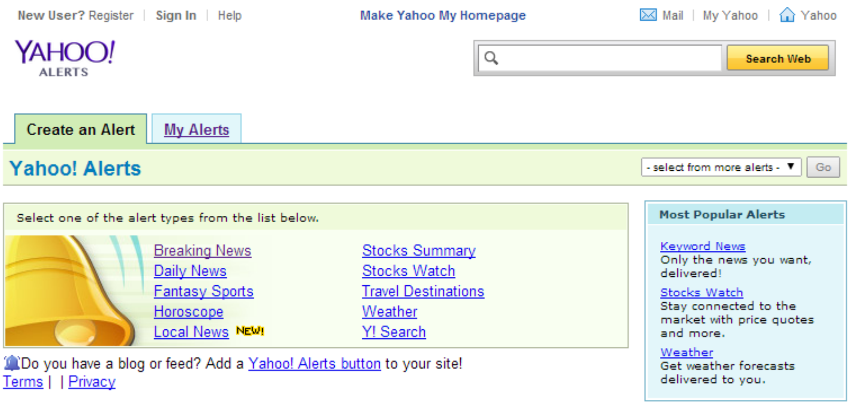
Google Alerts is a free google services that follows specific keywords and notifies the user on these instances. Results may point to Web, news, blogs, videos or groups. The filtering that is allowed consist of the following 5 options:

1. Kind of results All/News/Blog/Video/Talks/Books
2. Language of Results
3. Area
4. Frequency of updates in email RealTime/Once a Day/Once a week
5. The number of results Only the best/All of the results



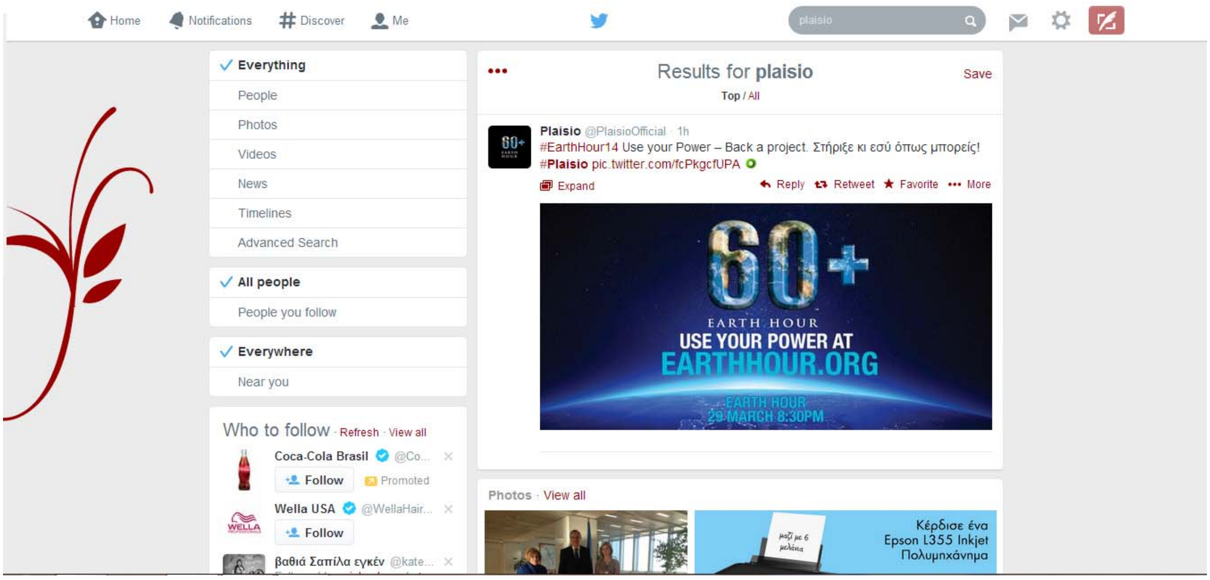
Yahoo Alerts

Another free tools, similar to the Google Alerts. It can be seen as the competitive service of Yahoo.



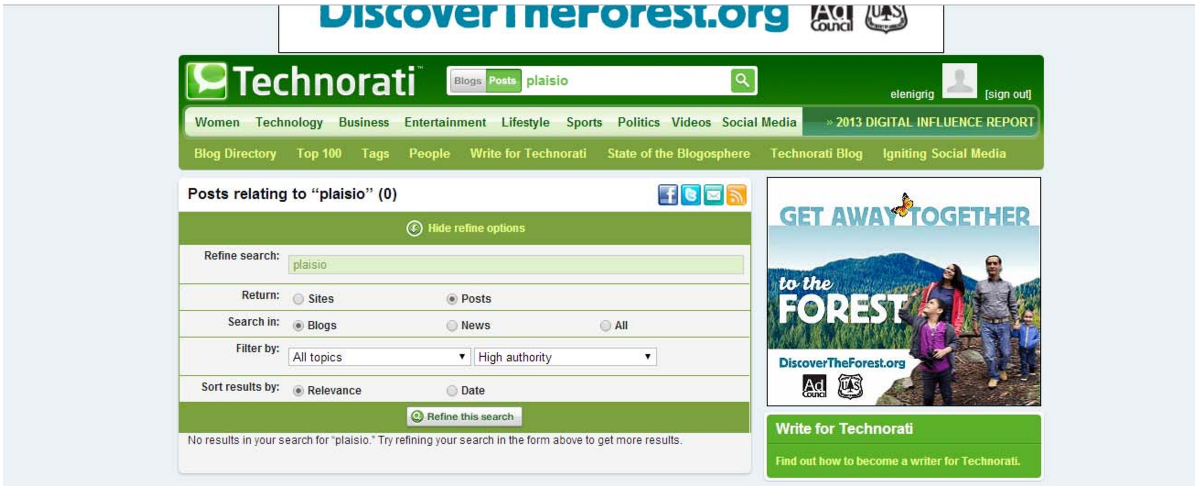
Twitter Search

Known as Summize, acquired by Twitter in 2008, now as Twitter Search facilitates the search in Twitter for any referances of a particular keyword/phrase. There are few parameters to choose from, such as Location, Sentiment, Date, Links, specific user etc.



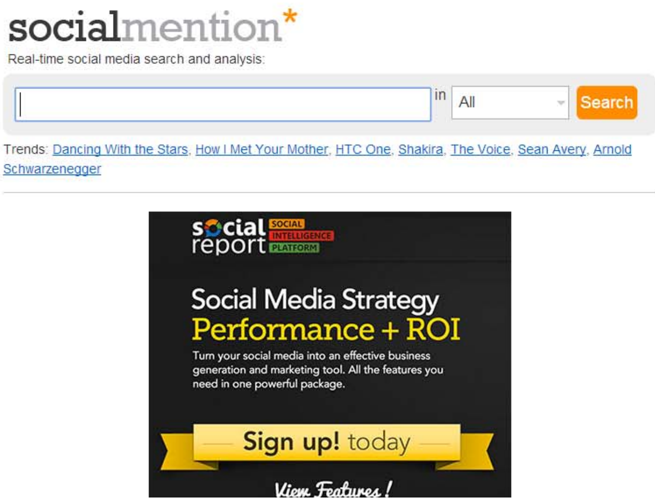
Technorati

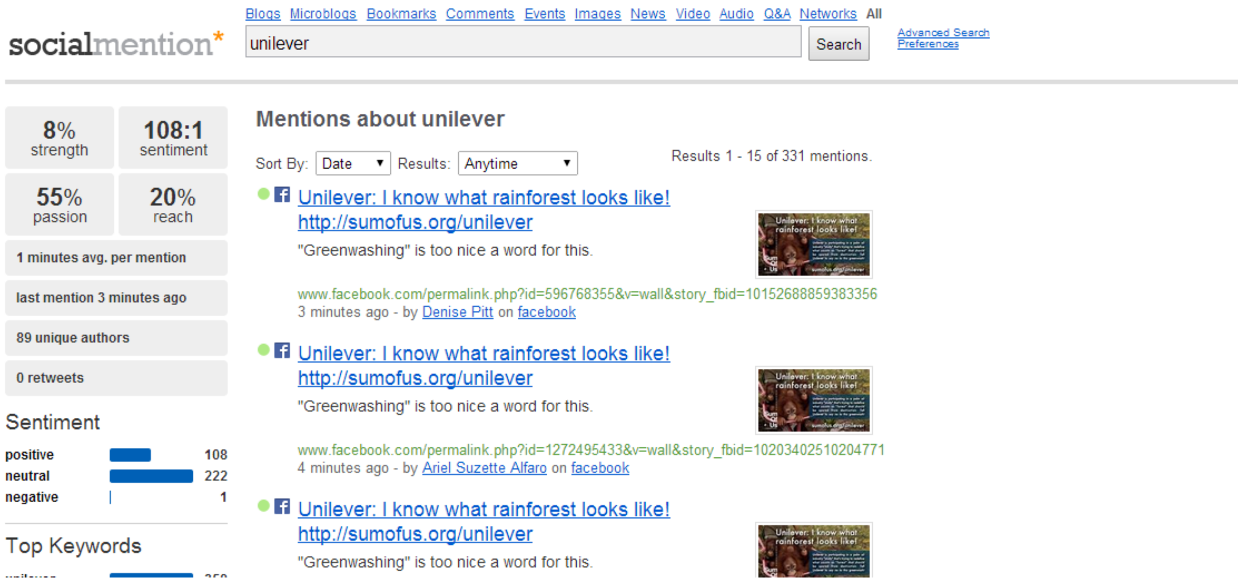
Technorati launched its ad network in 2008, and at one time was one of the largest ad networks reaching more than 100 million unique visitors per month. The name Technorati was a portmanteau of the words technology and literati, which invokes the notion of technological intelligence or intellectualism. Technorati uses real-time market insights to optimize digital advertising interactions across its publisher network with the use of technology designed to help publishers get discovered by advertisers and earn more for their content.



Social Mention

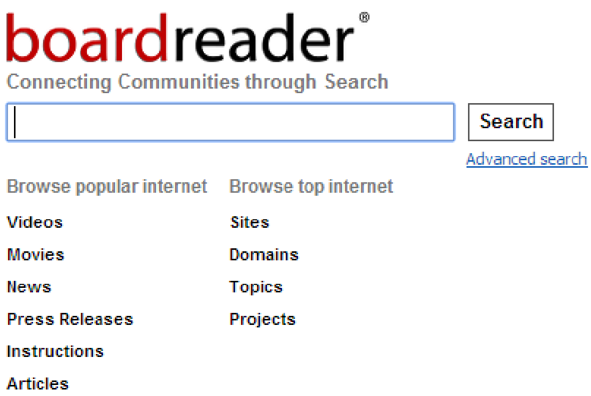
This interesting tool besides that is free to use focuses on sources exclusively from social media and even provides a sentiment analysis automatically.





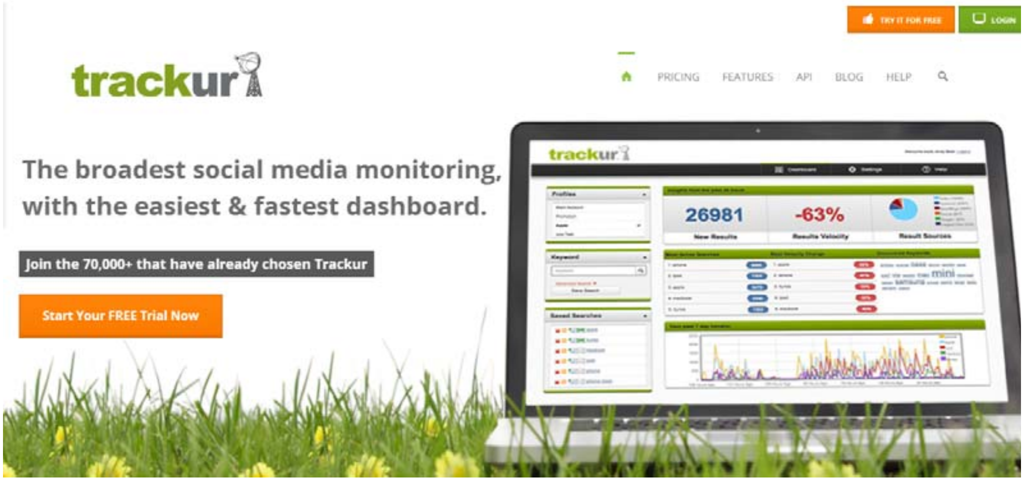
BoardReader

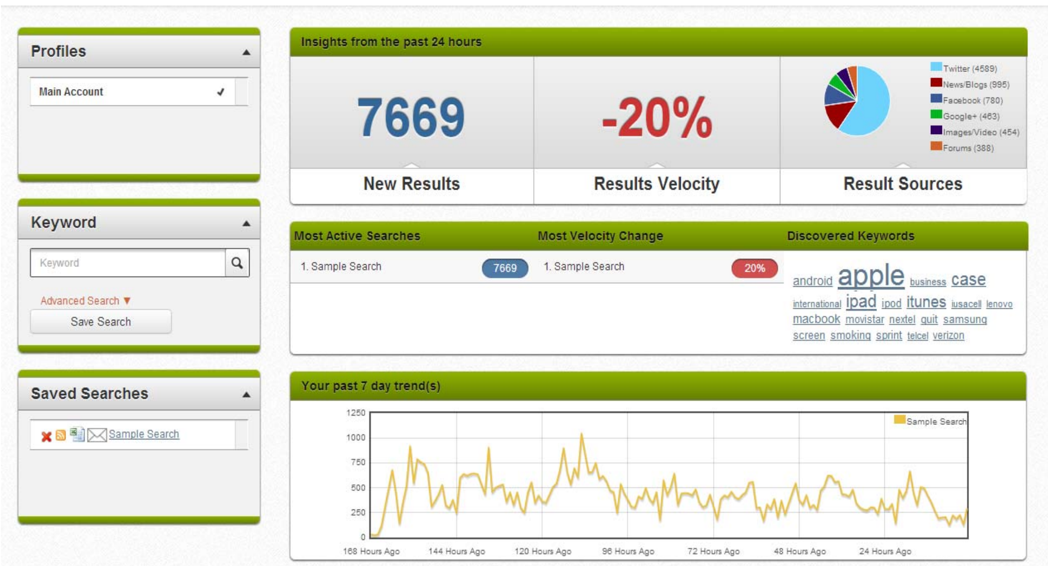
Board Reader is a tool that follows discussions – comments on forums and messengers. A simple platform that offers options between Videos/ Movies/ News/ Press Releases/ Instructions/ Articles. You might choose to browse between Sites/ Domains/ Topics/ Projects

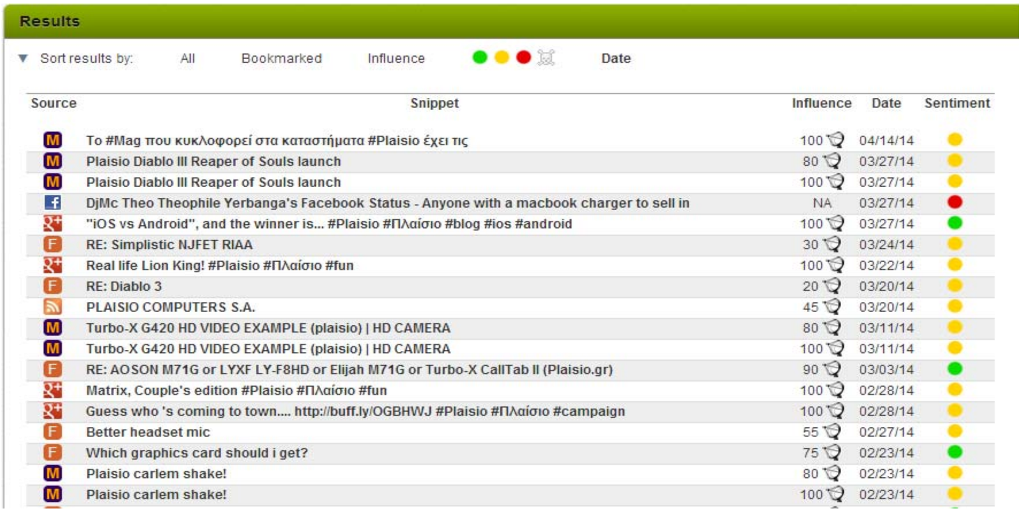


Trackur

Trackur has no free options. It pulls referals regarding the keywords one looks from a variety of sources.







Results

▼ Sort results by: All Bookmarked Influence Date

Source	Snippet	Influence	Date	Sentiment
M	Το #Mag που κυκλοφορεί στα καταστήματα #Plaisio έχει τις	100	04/14/14	
M	Plaisio Diablo III Reaper of Souls launch	80	03/27/14	
M	Plaisio Diablo III Reaper of Souls launch	100	03/27/14	
f	DjMc Theo Theophile Yerbanga's Facebook Status - Anyone with a macbook charger to sell in	NA	03/27/14	
g+	"iOS vs Android", and the winner is... #Plaisio #Πλαίσιο #blog #ios #android	100	03/27/14	
F	RE: Simplistic NJFET RIAA	30	03/24/14	
g+	Real life Lion King! #Plaisio #Πλαίσιο #fun	100	03/22/14	
F	RE: Diablo 3	20	03/20/14	
	PLAISIO COMPUTERS S.A.	45	03/20/14	
M	Turbo-X G420 HD VIDEO EXAMPLE (plaisio) | HD CAMERA	80	03/11/14	
M	Turbo-X G420 HD VIDEO EXAMPLE (plaisio) | HD CAMERA	100	03/11/14	
F	RE: AOSON M71G or LYXF LY-F8HD or Elijah M71G or Turbo-X CallTab II (Plaisio.gr)	90	03/03/14	
g+	Matrix, Couple's edition #Plaisio #Πλαίσιο #fun	100	02/28/14	
g+	Guess who 's coming to town.... http://buff.ly/OGBHWJ #Plaisio #Πλαίσιο #campaign	100	02/28/14	
F	Better headset mic	55	02/27/14	
F	Which graphics card should i get?	75	02/23/14	
M	Plaisio carlem shake!	80	02/23/14	
M	Plaisio carlem shake!	100	02/23/14	

Of course there is always the RepTrak tool we borrowed the definition from, in the previous chapter.

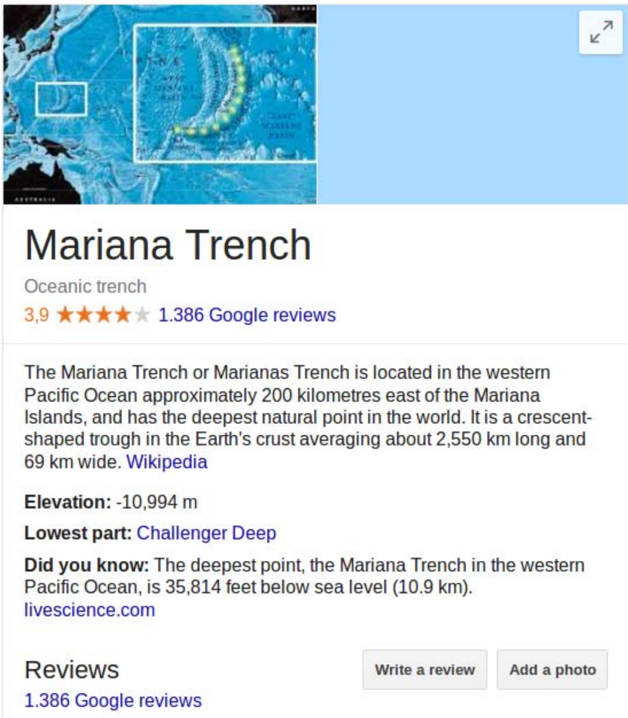
3.2 Strong Online Indicators

The tools we mentioned in the previous section are to track online management reputation. Most of them are collecting bits and pieces from around the web and present them in a report. In this section we are going to present tools provide indications about our organization.

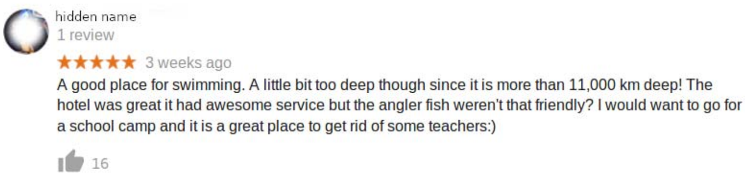
Google Reviews

Google reviews for places require a physical location that is being validated with a post letter containing a pin. That does not mean the business could not be just online based like an e-shop but rather that should have a registered location.

Public places can also be in the google places reviews. That creates confusion in some cases with a salient example the one of Mariana Trench. It is one of the ocean's deepest point(illustration 19) with depth almost 11Km.



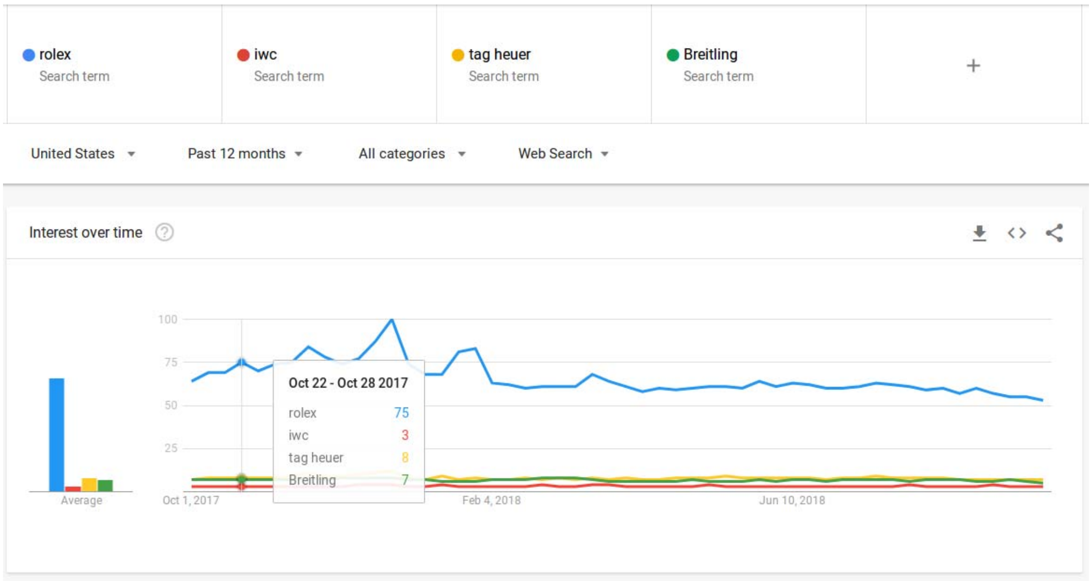
Of course people are reviewing but ...(illustration 20).



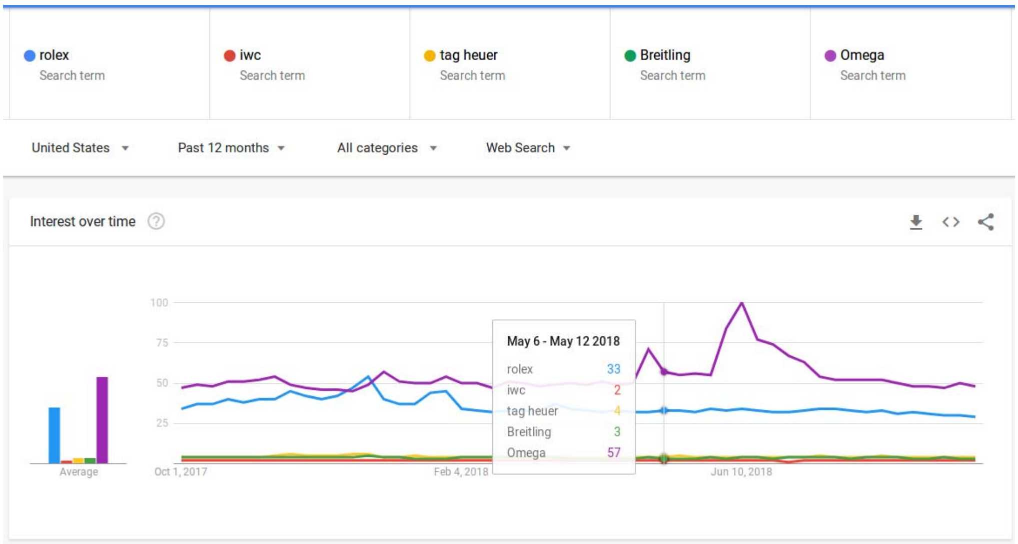
We consider that this tool has become important because simply it's featured on the right column on the page next to the search results.

GOOGLE TRENDS

In this tool we could search different terms and see their interest in time and over regions and subregions. The number of terms is limited to five. Certainly here someone is not limited to use company names but any kind of terms.



Therefore one should be careful on how to input the terms and what should be the expectations for the output. You can notice the predominance of Rolex over the rest of the terms (75 over 3, 8, 7). See what happens when we add the term Omega for the same period and region(illustation 21).



Now Omega is over Rolex with the index of the first 57 over 33 and 4, 3 and 2 the rest of the companies. While Omega is not only a watch company the rest of the companies have unique names.

As we mentioned earliear Google Page Rank is one (indicator) of it's own.

3.3 Social Media Analytics

A big part in today's business, in terms of insights, is social media analytics. While Social Media often provides insights and reports on our personal pages, they also provide API's (Application Programming Interfaces). API's is a way to get data from the back-end. After a small validation as developer these API's give us access to data that they are available to the public. With some knowledge of programming, a developer can access download and analyze social media data. Some of the cases that are often used is Twitter, Youtube, Facebook, Instagram and Pinterest.

Even though this is not an one button solution like the tools we presented earlier, this is one is for free. Also it is more complecated but complexity means more customisation and more options. Sometimes complexity derives from bad design, we do not consider that case. In the following chapters we will plan and execute a strategy to receive, manipulate and analyze social media data using this approach. The review of ReLap 2014, that follows, is also based in the collection of Data from Twitter.

The programming languages that often used for the authentication, download, manipulation and analysis (sentiment analysis as well) of the data are usually are Python and R but it can be others. It is considered that since we can communicate with some ease the process of social infuence is happening much more. As a consequence the trends are changing fast. It also means more data to look insights for. Some reasearchers and businesses believe that for that reason a portion of Marketing will be shifting to Insights. Therefore these methods are not to be neglected. Following is the overview of RepLab which is a study that uses input data from twitter as well.

3.4 Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management

For avoiding any confusions around the work and the organization presented we are quoting the abstract of the

"*Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management"* (Enrique Amigó and Jorge Carrillo de Albornoz and Irina Chugur and Adolfo Corujo and Julio Gonzalo and Edgar Meij and Damiano Spina, 2014)

' This paper describes the organization and results of RepLab 2014, the third competitive evaluation campaign for Online Reputation Management systems. This year the focus lied on two new tasks: reputation dimensions classification and author profiling, which complement the aspects of reputation analysis studied in the previous campaigns. The participants were asked ***(1) to classify tweets applying a standard typology of reputation dimensions*** *and* ***(2) categorise Twitter profiles by type of author as well as rank them according to their influence****. New data collections were provided for the development and evaluation of systems*

that participated in this benchmarking activity. '

RepLab is a competitive evaluation exercise supported by the EU project LiMo-SINe. It aims at encouraging research on Online Reputation Management and providing a framework for collaboration between academia and practitioners in the form of a "living lab": a series of evaluation campaigns in which task design and evaluation are jointly carried out by researchers and the target user community (in our case, reputation management experts). Similar to the previous campaigns [1,2], RepLab 2014 was organized as a CLEF lab. Previous RepLab editions focused on problems such as entity resolution (re-solving name ambiguity), topic detection (what are the issues discussed about the entity?), polarity for reputation (which statements and opinions have negative/positive implications for the reputation of the entity?) and alert detection (which are the issues that might harm the reputation of the entity?). Although online monitoring pervades all online media (news, social media, blogosphere, etc.), RepLab has always been focused on Twitter content, as it is the key media for early detection of potential reputational issues. In 2014, RepLab focused on two additional aspects of reputation analysis– reputation dimensions classification and author profiling that complement the tasks tackled in the previous campaigns. As we will see below, reputation dimensions contribute to a better understanding of the topic of a tweet or group of tweets, whilst author profiling provides important information for priority ranking of tweets, as certain characteristics of the author can make a tweet (or a group of tweets) an alert, requiring special attention of reputation experts.

Vis-a-vis with the dimensions we are presenting the following table (illustration 10) describing the dimensions

Table 1: RepTrak dimensions. Definitions and examples of tweets.

Dimension	Definition and Example
Performance	Reflects long term business success and financial soundness of the company. `Goldman Profit Rises but Revenue Falls: Goldman Sachs reported a second-quarter profit of $1.05 billion,... http://dlvr.it/bmVY4`
Products & Services	Information about the company's products and services, as well as about consumer satisfaction. `BMW To Launch M3 and M5 In Matte Colors: Red, Blue, White but no black...`
Leadership	Related to the leading position of the company. `Goldman Sachs estimates the gross margin on ACI software to be 95% O_o`
Citizenship	The company's acknowledgement of the social and environmental responsibility, including ethical aspects of business: integrity, transparency and accountability. `Find out more about Santander Universities scholarships, grants, awards and SME Internship Programme bit.ly/1mM12OX`
Governance	Related to the relationship between the company and the public authorities. `Judge orders Barclays to reveal names of 208 staff linked to Libor probe via @Telegraph soc.li/mJVPh1R`
Workplace	Related to the working environment and the company's ability to attract, form and keep talented and highly qualified people. `Goldman Sachs exec quits via open letter in The New York Times, brands bank working environment ''toxic and destructive'' ow.ly/9EaLc`
Innovation	The innovativeness shown by the company, nurturing novel ideas and incorporating them into products. `Eddy Merckx Cycles announced a partnership with Lexus to develop their ETT Hme trial bike. More info at...http://fb.me/1VAeS3zJP`

Author Categorisation.

The task was to classify Twitter profiles by type of author: Company (i.e., corporate accounts of the company itself), Professional (in the economic domain of the company), Celebrity, Employee, Stockholder, Investor, Journalist, Sportsman, Public Institution, and Non-Governmental Organisation (NGO). The system's output was expected to be a list of profile identifiers with the assigned categories, one per profile.

Author Ranking.

Using as input the same set of Twitter profiles as in the task above, systems had to find out which authors had more reputational influence 1440 (who the influencers or opinion makers are) and which profiles are less influential or have no influence at all. For a given domain (e.g., automotive or banking), the systems' output was a ranking of profiles according to their probability of being an opinion maker with respect to the concrete domain, optionally including the corresponding weights. Note that, because the number of opinion makers is expected to be low, we modelled the task as a search problem (hence the system output is a ranked list) rather than as a classification problem.

Data Collection

This data collection is based on the RepLab 2013 corpus 12 and contains over 48,000 manually labelled English and Spanish tweets related to 31 entities from the automotive and banking domains. The training set was composed of 15,562 Twitter posts and 32,446 tweets were reserved for the test set. Both data sets were manually labelled by annotators trained and supervised by experts in Online Reputation Management from the online division of a leading Public Relations consultancy Llorente & Cuenca.

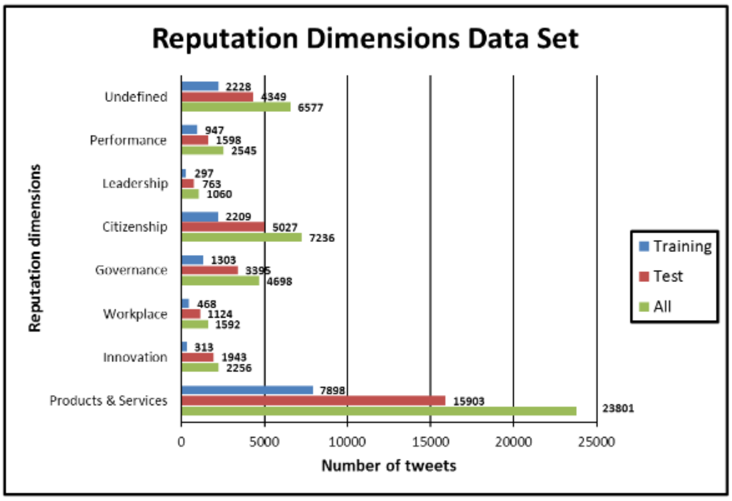
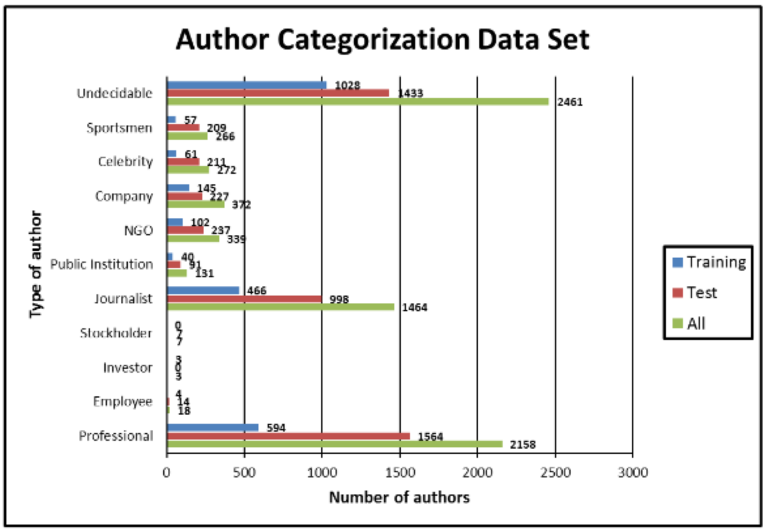


Illustration 25: Distribution of classes in the Reputation Dimensions data



In RepLab 2014 they developed an evaluation methodology and test collections for two different reputation management problems: (1) classification of tweets according to the reputation dimensions, and (2) identification and categorisation of opinion makers. Once more, the manual annotations were provided by reputation experts from Llorente & Cuenca (48,000 tweets and 7,000 author profiles annotated).

Being the first shared evaluation on these tasks, participants explored a wide range of approaches in each of them. The classification of tweets according to their reputation dimensions seems to be feasible, although it is not yet clear which are the best signals and techniques to optimally solve it. Author categorisation, on the other hand, proved to be challenging in this initial approximation.

Chapter 4
Case Study, Description, Design

4.1 Rolex

As RepTrak 2018 is suggesting Rolex should hold the first place for best reputation score for the year 2018. Our query in google trends between brands Rolex, IWC, TagHeuer and Breitling suggests that Rolex is a term used much more in a certain period. Without a doubt Rolex is a company who have been striving to convience consumers about their company and probably have done a great job. So for those reasons, more importantly because there are going to be great resources, we are choosing Rolex as reference point for our case study.

Rolex SA (/ˈroʊlɛks/) is a Swiss luxury watchmaker. The company and its subsidiary Montres Tudor SA design, manufacture, distribute and service wristwatches sold under the Rolex and Tudor brands. Founded by Hans Wilsdorf and Alfred Davis in London, England in 1905 as Wilsdorf and Davis, Rolex moved its base of operations to Geneva, Switzerland in 1919. *Forbes* ranked Rolex 64th on its 2016 list of the world's most powerful global brands. Rolex is the largest single high end watch brand. (Wikipedia)

Rolex is often sponsor elite or expensive sports and has found success with it. The fact that Rolex sells high end expensive watches might automatically means they are looking for people who can afford them. Probably by sponsorship these events brings them closer to potential customers but also help them to maintain a good reputation too.

Rolex watches are frequently counterfeited, and these are often illegally sold on the street and online. Counterfeit Rolex watches vary in quality: some use the cheapest of movements, while others use automatic movements, and some use an ETA movement. However, the majority of these counterfeit watches are easily identifiable by jewellers and other experts.

The strategy and implementation will be straight forward. We will follow the social media analysis way. We will obtain data from 3 different sources. Twitter, Youtube and Amazon. After making sure we have structured data the way we expected, in JSON and CSV, we will follow analysis. The objective will be to assess some information input using data we can freely collect from web. In our case, it's about How rather than making the best analysis on Rolex ever performed. Also other limitations that we will discuss later on, on chapter 6.

4.2 Amazon

Here can be applied the idea we presented in earlier chapters that simply a procudure cannot work for all types of businesses. Who would buy a Rolex watches of mutliple thousands at Amazon? Never the less based on these expectations the results following to an analysis(next chapter) were fairly adequate.

The process we followed was. Going to search Rolex in amazon.com and then choose from the filters only Rolex. What we left with is a 7 results pages all with different Rolex listings. We downloaded the html page source of these result pages and we concatanate them in to a single txt file. We found and replaced the part we were interested to harvest, and we break an extra line after it. Of course that was no other than data-asin="...". ASIN stands for *Amazon Standard Identification Number*.

Combining python and regular expression we end up with a txt file that is a list for all the ASIN of the 308 watches that are listed in amazon.com.

```
#! /usr/bin/python
# -*- coding: utf-8 -*-

import re
amazon = open("amazon.txt","r")
#mind that we make should for each line on the file we load there is no more than 1 time the matching regex
#the re.search if find an instance in an element of the loop moves
#to the next line regardless if they are other instances on the same line

for whatever in amazon:
    asin=re.search(r'asin="(.*?)"', whatever, re.IGNORECASE)
    #print type(whatever)
    if asin:
        print asin.group(1)
amazon.close()
```

*code in the appendix

after with another script that we did not create but we customized it for our purposes. It has multiple parts but basically is a function that takes the array of asin's from the first script and scrapes the data we need and save them in a dataframe. A review in amazon is comprised by the following fields : rating, title, author, date, body, helpful.

Original script found at : https://www.knoyd.com/blog/amazon-review-scraper , also in appendix.

4.3 Twitter

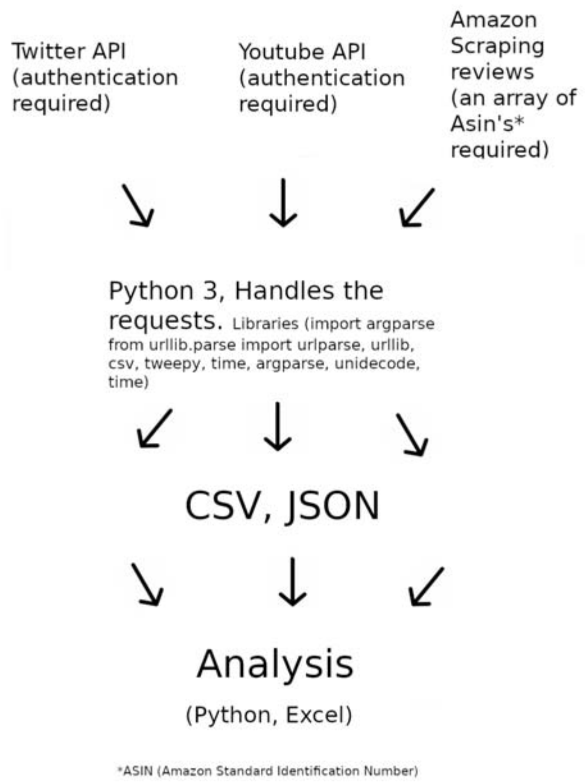
After having the appropriate credentials from twitter API, we used a script to download 3200 tweets containing the “Rolex” keyword. The script we again customized for our needs, we did not create. It uses the python library tweepy to handle authorization and the requests. The output is always a csv file. The limitation of our tweets to 180 to a quarter of an hour made that process last for several hours. The fields we end up with are created at(date_stamp), retwc(number of retweets, int), hashtag(str), followers(int), friends, tweet_id(int), cords(coordinates) , source, if_link(boolean, checks if there is link in text), username, followers, authorloc(author location), lang(language of tweet). In the picture you can see the function that does most of the work and you can find the full code in the appendix.

```python
def tw_search(api):
    counter = 0
    # Open/Create a file to append data
    csvFile = open('result.csv','w')
    #Use csv Writer
    csvWriter = csv.writer(csvFile)
    csvWriter.writerow(["created", "text", "retwc", "hashtag", "followers", "friends","tweet_id","cords","source","if_link","username","followers","authorloc","lang"])

    for tweet in tweepy.Cursor(api.search,
                               q = qw,
                               g = ge,
                               lang = l,
                               result_type = t,
                               count = c).items():

        #TWEET INFO
        created = tweet.created_at   #tweet created
        text    = tweet.text         #tweet text
        tweet_id = tweet.id          #tweet ID# (not author ID#)
        cords   = tweet.coordinates  #geographic co-ordinates
        retwc   = tweet.retweet_count #re-tweet count
        lang   = tweet.lang
        try:
            hashtag = tweet.entities[u'hashtags'][0][u'text'] #hashtags used
        except:
            hashtag = "None"
        try:
            rawurl = tweet.entities[u'urls'][0][u'url'] #URLs used
            urls = url_fix(rawurl)
        except:
            urls    = "None"
        #AUTHOR INFO
        username  = tweet.author.name            #author/user name
        usersince = tweet.author.created_at      #author/user profile creation date
        followers = tweet.author.followers_count #number of author/user followers (inlink)
        friends   = tweet.author.friends_count   #number of author/user friends (outlink)
        authorid  = tweet.author.id              #author/user ID#
        authorloc = tweet.author.location        #author/user location
        #TECHNOLOGY INFO
        geoenable = tweet.author.geo_enabled     #is author/user account geo enabled?
        source    = tweet.source                 #platform source for tweet
                # Dongho 03/28/16
        if_link=False
        if 'https' in text:
            if_link=True
        time.sleep(1)
        csvWriter.writerow([created, str(text).encode("utf-8"), retwc, hashtag, followers, friends,tweet_id,cords,source,if_link,username,followers,authorloc,lang])
        counter = counter +1
        print ('...'+'tweets collected:'+str(counter))

        if (counter == c):
            break
    csvFile.close()
```

4.4 Youtube

To access youtube API, it is required to have credentials. So again we created the keys to access and we used a python script. Unfortunately, youtube’s limits are even less and we have 50 information about video with “Rolex” in the title. This script resides in the appendix as well.

4.5 Schema and research expectations



To sum up the chapter we going to define our expectations of the project and the implementation we used to arrive there. We would like to know 1) what is the reputation around rolex watches? 2) How this reputation is being formed and disseminated ? 3) Which platform users perform most comments? 4) Who are the biggest contributors or most influencial users? 5) Which are the most used words collectivelly that users typed? 6) How often they post throughout the day and what is the volume per hour and minute? 7) What is the sentiment of their text?

In order to find clues about these questions we:

Amazon

1) Used python code to strip out all of the product id's from amazon.com
2) Used python code to scrape all of the rolex product reviews
3) Cleaned and preprocess the data for analysis
4) We performed review length, average rating, timeseries analysis, wordcloud

Twitter

1) Used python code to pass the api authorization, download tweets
2) Cleaned and preprocess the data for analysis
3) We performed correlation between the fields, wordcloud, tweet volume by hour and minute, biggest contributors, if the tweet contains a link, biggest contribution by area, sources (devices) that tweets made, sentiment analysis

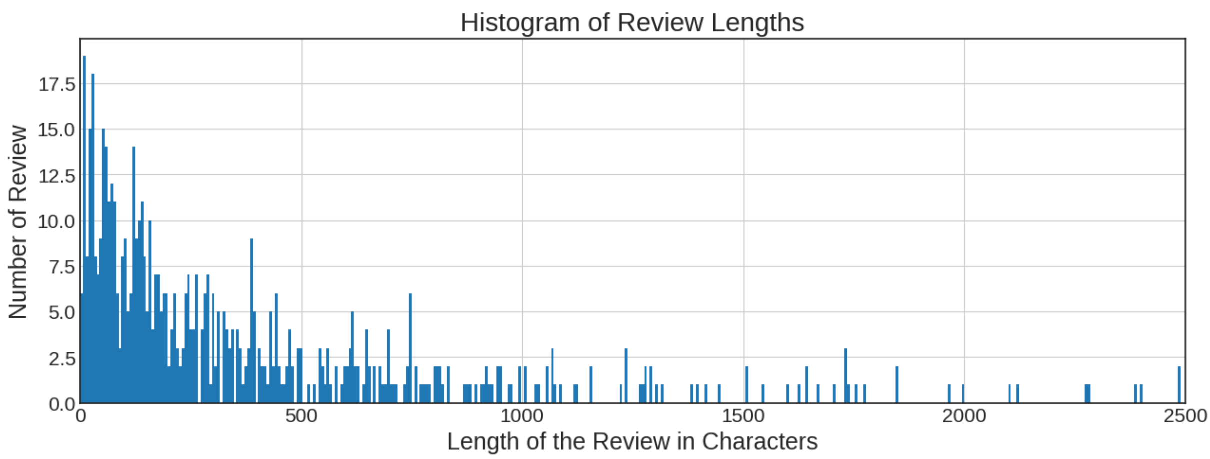
Youtube

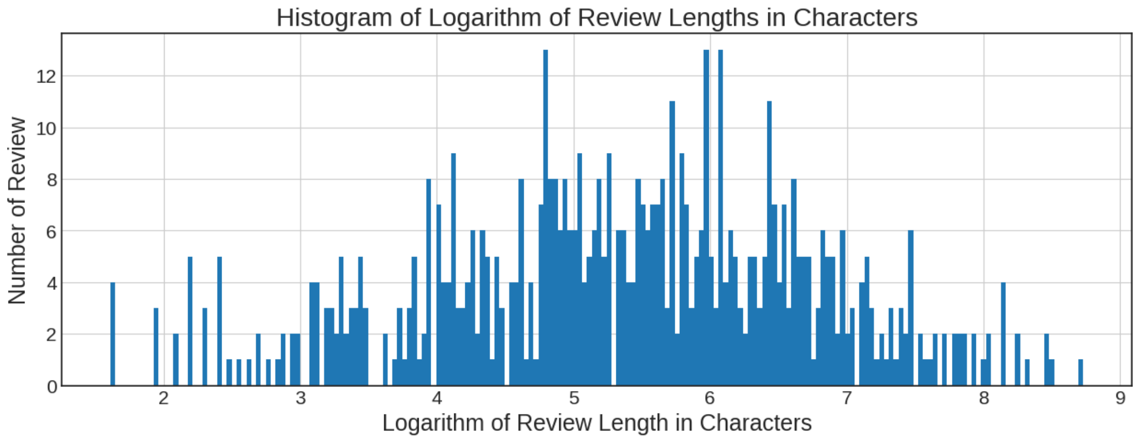
1) Used python code to pass the api authorization, download the youtube structured data
2) Cleaned and preprocess the data for analysis
3) Performed wordclouds and correlations between the fields

Chapter 5
Analysis –Result Presentation, Evaluation

5.1 Amazon

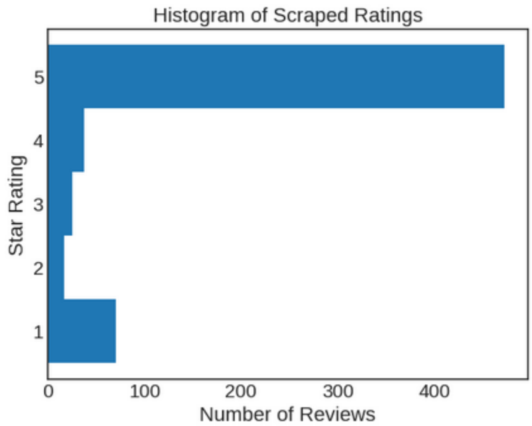
Starting with amazon, we managed to scrape 767 reviews. We cleaned up a little bit the dataframe. Cutting off stopwords etc. We start with a Histogram of Reviews' Lengths visualization.



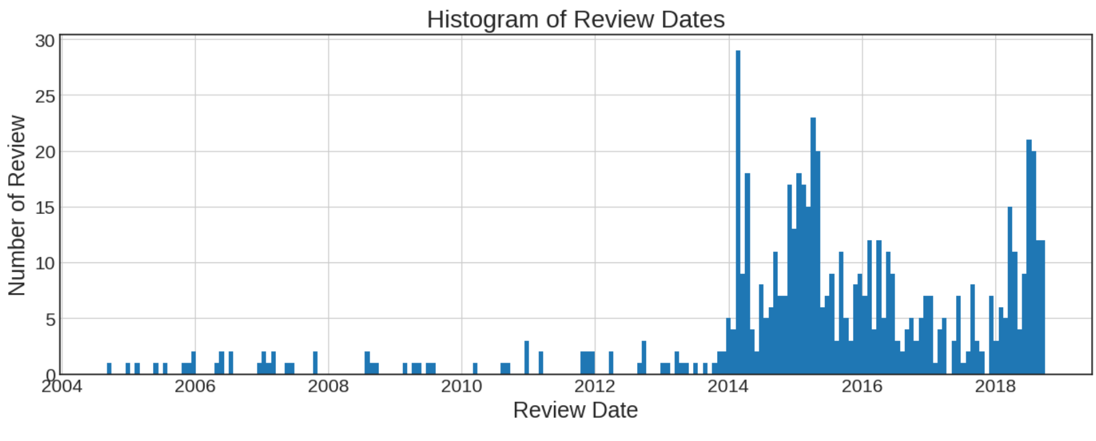


And with a log function.

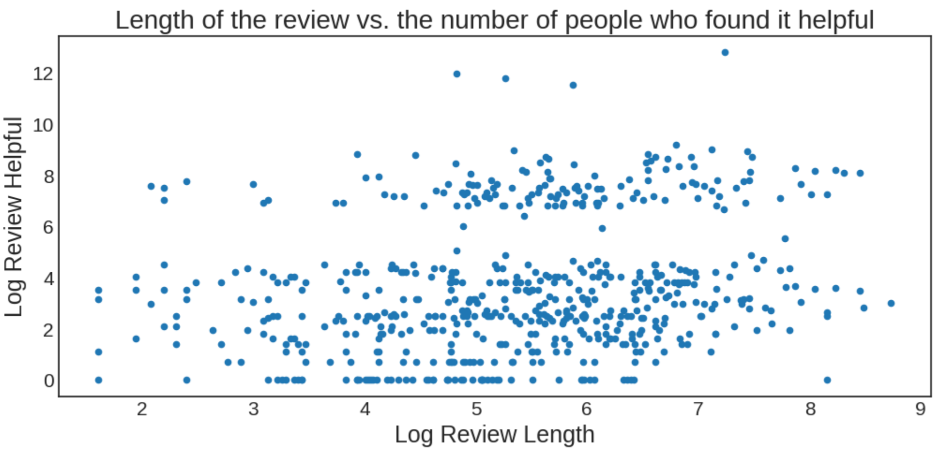
The next one is a amazon review style histogram. It averages these 767 reviews we had. Once we tried some single product reviews that were somelike popular, the histogram was matching exactly amazon's histogram for the product. Here we don't have one to compare one from amazon. A nice way to think about it is like a Unified Rolex Product in amazon's listings. Mind that there were a lot of listings without review that are not part of this average. This average is between the listings that had a review.



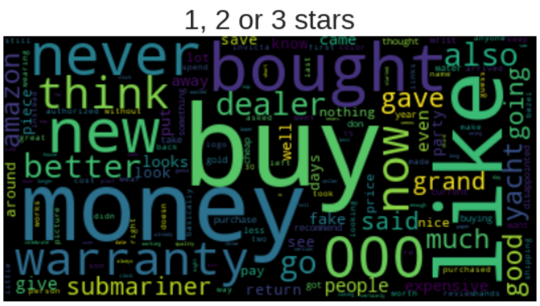
A review_date histogram.

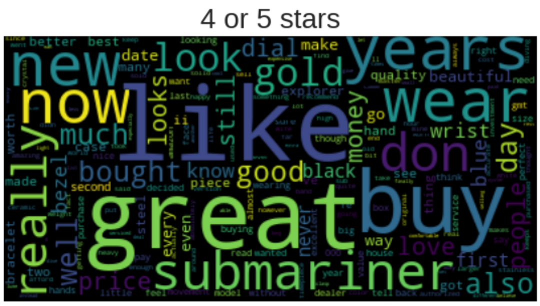


Length of the review vs. the number of people who found it helpful.



We extend our stopwords additionally apending ‘Rolex’, ‘Rolex watches’,’watches’, etc. We do not want to see them in our wordcloud and therefore we exclude them.





5.2 Twitter

We obtained 3200 tweets. That is regarding 07 and 08 day of October. We uploaded the dataframe in to the memory together with all the libraries needed and we ask for a correlation table between the variables.

	retwc	followers	friends	tweet_id	if_link
retwc	1.000000	-0.006004	-0.018493	-0.043247	-0.000249
followers	-0.006004	1.000000	0.126278	0.033488	0.023590

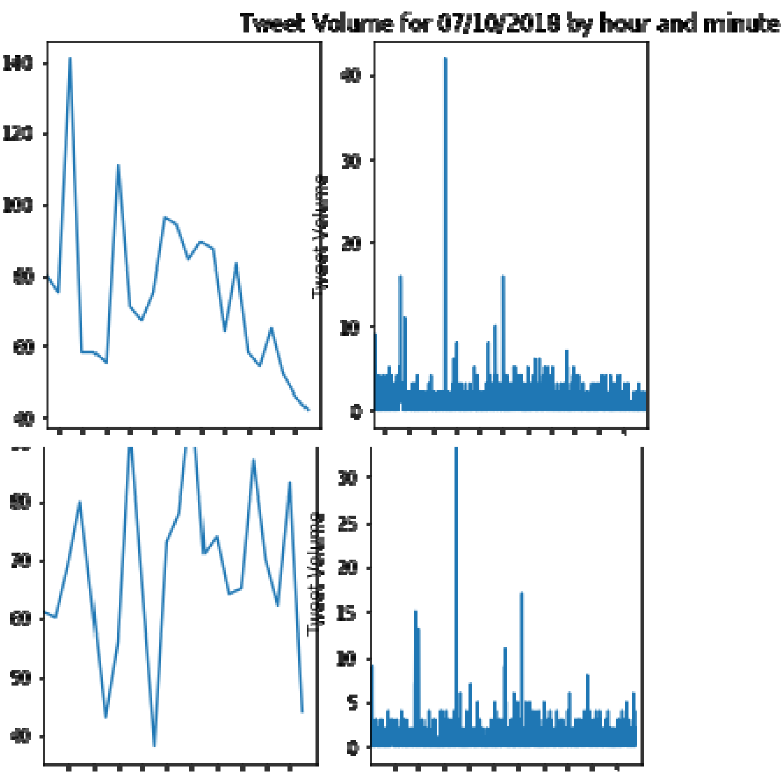
friends   -0.018493   0.126278 1.000000 -0.041453 0.065467
tweet_id   -0.043247   0.033488 -0.041453 1.000000 0.037880
if_link   -0.000249   0.023590 0.065467 0.037880 1.000000
followers.1 -0.006004   1.000000 0.126278 0.033488 0.023590

We do not see any worthmentioning correlations. Friends with followers they have a small positive correlation which is not surprising.

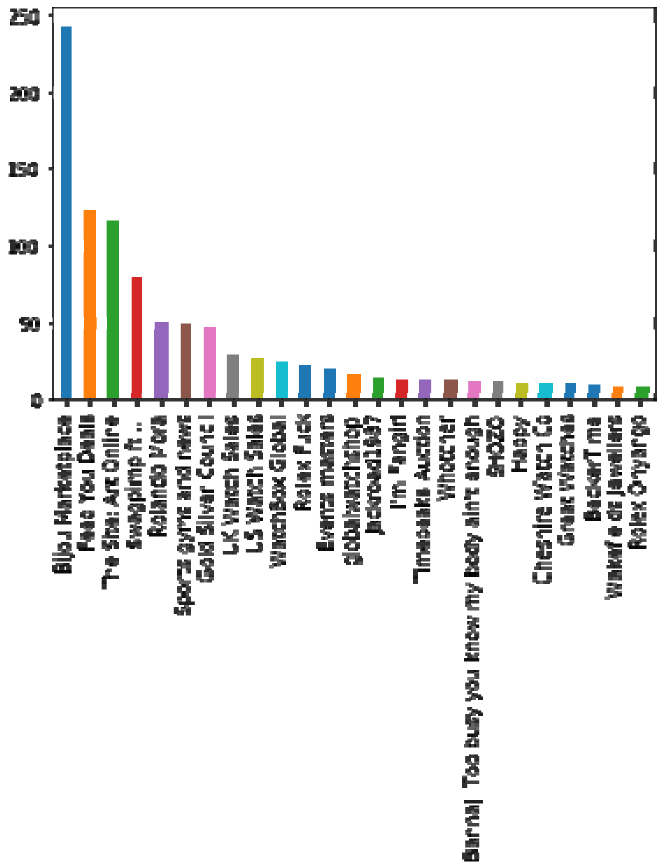
Let’s see a wordcloud, after we removed stopwords etc..

Keyword was “Rolex”.

So lets move on to a timeseries analysis for each of the days.



There are a few minutes throughout the day that have around 40 tweets containing “Rolex”, in both days.
We try to locate the most influencial users. So there is a bar chart with the 25 most popular, in descending order.



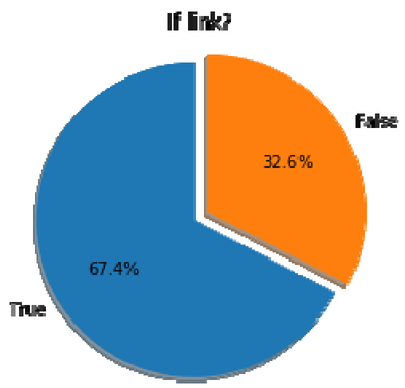
And a table to complete the visualization.

Bijou Marketplace	243
Feed You Deals	123
The Site: Art Online	116
$wagpimp ft ...	79
Rolando Mora	49
Sports gyms and news	48
Gold Silver Council	46
UK Watch Sales	28
US Watch Sales	26

WatchBox Global	24
Rolex Fuck	21
Events masters	19
globalwatchshop	15
Jackroad1987	13
I'm Fangirl	12
Timepeaks Auction	12
Whotcher	12
Banna‖ Too busy you know my body ain't enough	11
SHOZO	11
Happy	10
Cheshire Watch Co	10
Great Watches	10
BeckerTime	8
Wakefields Jewellers	7
Rolex Onyango	7

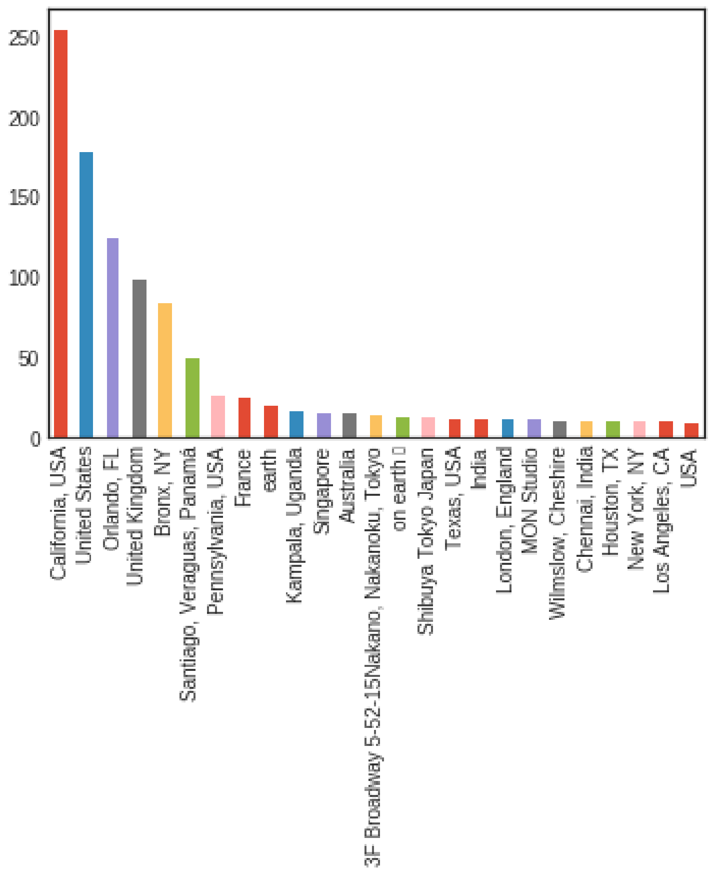
In other words that means that Bijou Marketplace made a 243 tweets containing keywork "Rolex" only in 7 and 8 of October, 2018.

All of 3200 tweets are made in English.
There is another field that derives from our implementantion in the tw_search() function. It returns a Boolean value if the text of the tweet contains a link.



Fact that makes us believe that the main purpose of the tweet is to direct the bystanders to another place of the web, probably for adverstising purposes.

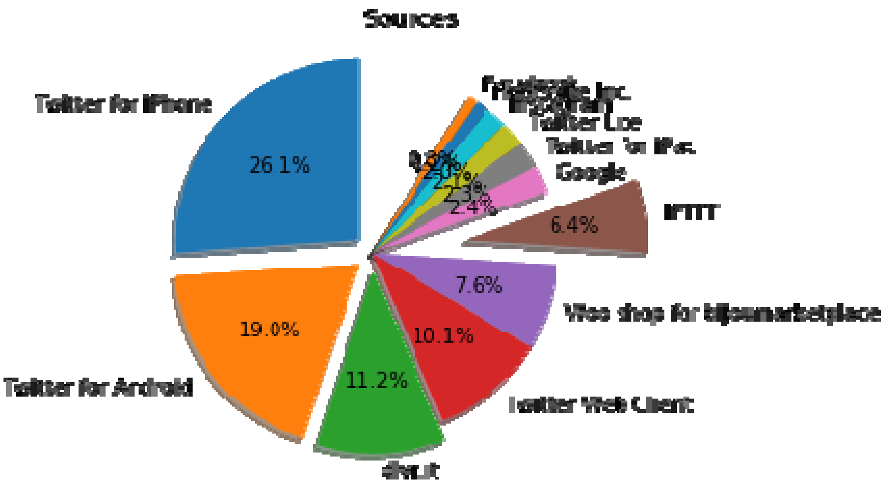
The countries that the authors are located. Note that some might overlap, example USA with California, USA. We did not address that cause we think it adds to the visualization.



Very interesting graph is the following one that basically depicts the sources that these tweets arrived. We cut all the ones that counts for >0.009 of the total. Not collectively but indivudially. That means the following graph is about 92.1% of the information available. The reserved space is depicted but without the information on the sources. We left with 12 sources.

Twitter for iPhone	834
Twitter for Android	607
dlvr.it	360
Twitter Web Client	323
Woo shop for bijoumarketplace	243
IFTTT	206
Google	77
Twitter for iPad	75
Twitter Lite	68
Instagram	64
Hootsuite Inc.	37
Facebook	27

..and 7.9% other.



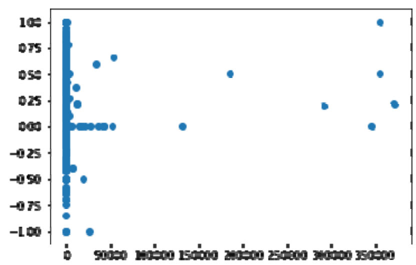
So it's obvious that more than 45.1% of the users post with their phones(Iphone and Twitter). With Iphone users to be 28.6 over 20.8 for android users. Is it because Iphone users are more or they are just more verbose? Conclusion that does not follow. Also interesting source is IFTTT. A piece that we pulled even more from the pie and it's brown. So IFTTT it's an application that is basically an action and a trigger and allows user to tweet automatically or do others automated options within Twitter (and other applications).

5.3 Twitter Sentiment Analysis

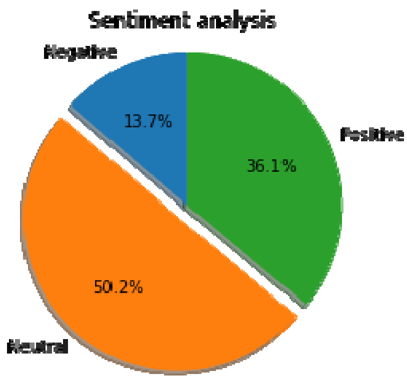
Now we add two more columns in our dataframe. For polarity and subjectivity. We will try to manipulate the data and make some sentiment analysis. A new correlation table with just the new two columns in regards two the rest of the fields.

	polarity	subjectivity
retwc	0.052344	0.023351
followers	0.023739	0.015885
friends	-0.009149	-0.028197
tweet_id	0.084995	0.101574
if_link	0.048843	-0.035365
followers.1	0.023739	0.015885
polarity	1.000000	0.384942
subjectivity	0.384942	1.000000

Subjectivity with polarity are **0.384942** positively correlated. A scatter plot between polarity and number of retweets.



Then a pie regarding the polarity, negative, neutral, positive.



It might look disappointing but it is not. Neutral could be expected to be more. The half that is not neutral is mostly Positive fits the reputation of the company.

These are some tweets that score <-0.5 in polarity. The characters between the backslashes are emojis or symbols.

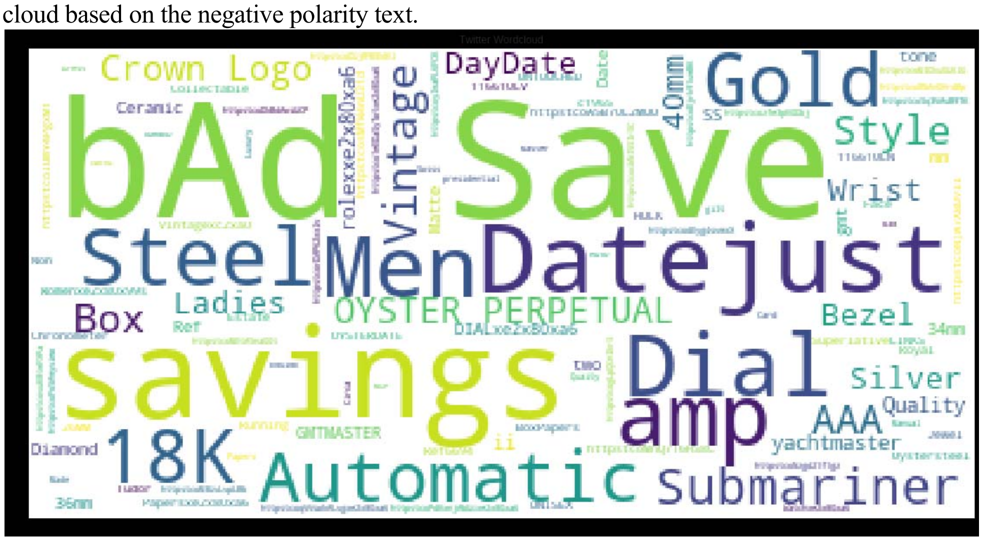
b'I want a Rolex so bad. I need to get a replica to hold me over until I can afford one.'
b'@boksunghao THIS FUCKING SENDS THE WAY HE WEARS HIS ROLEX OVER THE GLOVES'
b"RT @nacthebackwood: @on__BD @ioNKarma he's mad me and lil rolex are funnier than him"

b'I want a Rolex so bad \xf0\x9f\xa4\xa6\xf0\x9f\x8f\xbd\xe2\x80\x8d\xe2\x99\x82\xef\xb8\x8f'

b'Yes your coffee is cold.... you\xe2\x80\x99ve been sitting in here for 3 hours. No it\xe2\x80\x99s not my fault you didn\xe2\x80\x99t drink it, no I\xe2\x80\xa6 https://t.co/9fM6bEpyCH'
b'TELL A BAD BIHH GIRL LETS GO HANG...\nJU KNOW ME ROLEX GOLD CHAIN...'

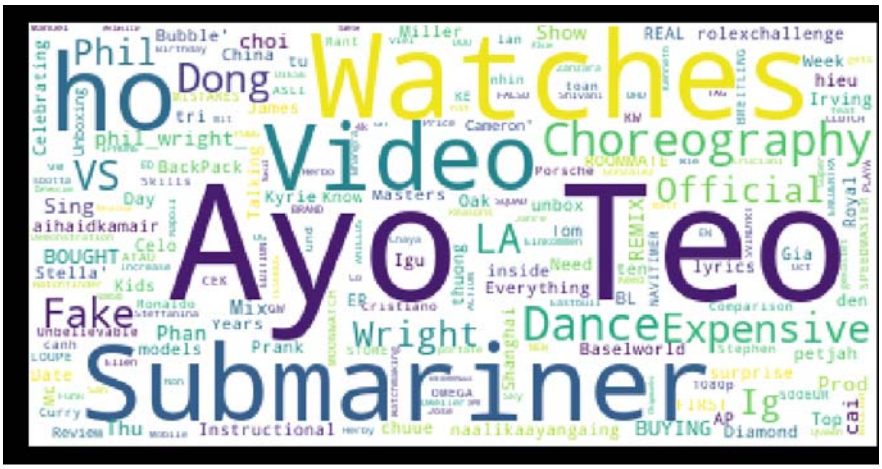
'RT Ain\xe2\x80\x99t nun retarded bout Gucci but this gold Rolex!'

There is a word cloud based on the negative polarity text.



5.4 Youtube

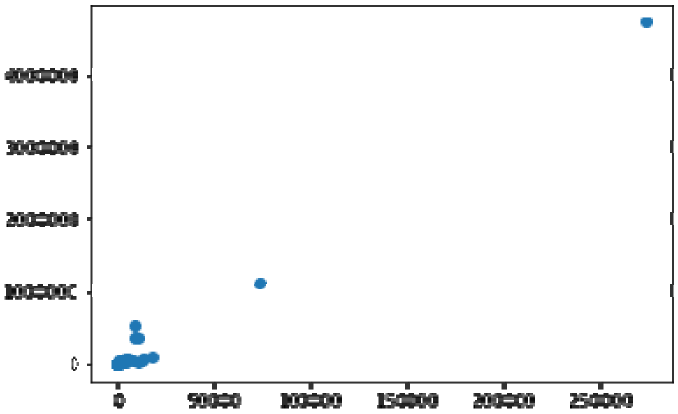
Here the sample is definetely poor. We only have information about 50 videos regarding Rolex. Unfortunately there are not many corners to cut when it comes to download data from Youtube API unless you own the data (a.k.a. originally uploaded the videos). Also the information that comes structured together with a video are poor in comparison to amazon or twitter. Nevertheless let's see a wordcloud made out of Youtube titles.



What is interesting is the correlation table between the fields.

	viewCount	likeCount	dislikeCount	commentCount	favoriteCount
viewCount	1.000000	0.995165	0.997707	0.988973	NaN
likeCount	0.995165	1.000000	0.992276	0.994758	NaN
dislikeCount	0.997707	0.992276	1.000000	0.987267	NaN
commentCount	0.988973	0.994758	0.987267	1.000000	NaN
favoriteCount	NaN	NaN	NaN	NaN	NaN

basically everything is highly correlated with a positive way. In other words as viewCount goes up , likeCount , dislikeCount, commentCount go up as well.



We make a linear regression model. Based on Ordinary Least Squares.

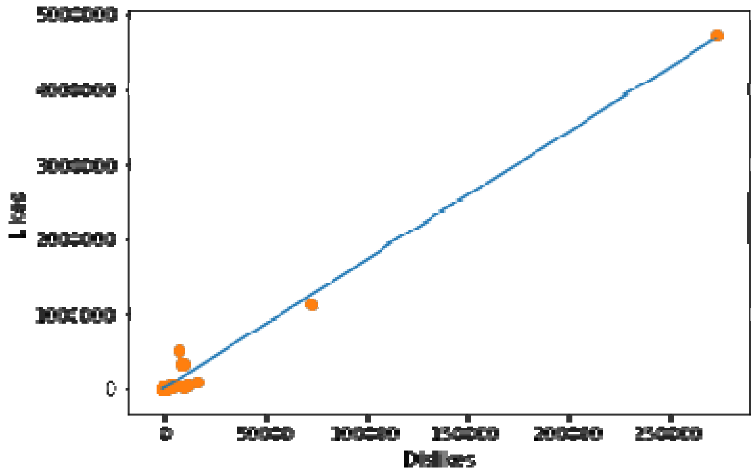
OLS Regression Results

==

Dep. Variable:	likeCount	R-squared:	0.985
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	2879.
Date:	Thu, 11 Oct 2018	Prob (F-statistic):	1.94e-42
Time:	20:41:14	Log-Likelihood:	-600.94

```
No. Observations:              47   AIC:                         1206.
Df Residuals:                  45   BIC:                         1210.
Df Model:                       1
Covariance Type:        nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2985.9986   1.33e+04     -0.225      0.823   -2.97e+04    2.37e+04
dislikeCount  17.1463      0.320     53.660      0.000      16.503      17.790
==============================================================================
Omnibus:                       29.408   Durbin-Watson:                   2.089
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              103.396
Skew:                           1.440   Prob(JB):                     3.53e-23
Kurtosis:                       9.671   Cond. No.                     4.28e+04
==============================================================================
```

and then if we plot the linear model together with the scatter plot(LikevsDislike).



There is almost perfect correlation. With the number of dislikes significantly smaller. They grow together, with the same rate but that does not mean for every person that likes one person dislikes. Roughly 100/5 (1000000/50000) likes over dislikes.

The 10 most viewed titles in our list. First song has a staggering half a Billion views!

```
                                              title      videoId  viewCount  \
0                  Ayo & Teo - Rolex (Official Video)  lwk5OUII9Vc  534297370
7   Ayo & Teo - Rolex | Prod. BL$$D & BackPack Mil...  rCx4NFNuYgI  129930147
2        Ayo & Teo - Rolex - Dance Instructional Video  A6ki_XuCMzA   23906261
37  "ROLEX" - Ayo & Teo Dance Choreography | Matt ...  gh_LKIn-bzo   20139527
```

1	Ayo & Teo - "Rolex" | Phil Wright Choreography...	Zcb6E3dje38	17911347
35	Ayo & Teo - "Rolex" | Phil Wright Choreography...	sl75jOlATuc	14226724
13	I BOUGHT MY ROOMMATE A $20,000 ROLEX! (surprise)	h9xSpyq0nf8	11772775
41	"ROLEX" - Bhangra Funk Dance | Ayo & Teo | Shi...	apJAt0BZt4k	8127965
31	Keed - ROLEX (REMIX) feat. OG Eastbull & Super...	NNN2mGWDoRM	5865564
4	Rolex	PYIJhau0Rz4	4942560

Chapter 6

Caveats And Disclaimers

6.1 CMC is lacking

Getting a review from an individual that stands in front of you can be very different from a written review. People we are constrantly training to get clues, ques or signs from communicating with each other. As seen earlier in the text, socialization is a process that creates this sense of homophily. Similarly, we learn to broadcast and receive verbal and non-verbal signs that add to the communication. This can include all the facial signs, like grimaches, body language and prosody.

All of these are missing from CMC(Commuter Mediated Communication). It has been argued that even the medium we choose to use ad something to the communication even the message itself. A cliche story we all know, is the break up story. The message could be ‘it’s not you, it’s me :(’ but if done by SMS will be considered pitiful and unkind. Even though the SMS would look like the most convienent way since the person has intention to discontinue the relationship.

As in f2f communication people are trying to manage the impressions, similarily in CMC the users exploit the technological aspects in order to enhance the messages they construct to manage impressions and facilitate desired relationships. It has been examined how CMC users managed message composing time, editing behaviors, personal language, sentence complexity, and relational tone in their initial messages to different presumed targets, and the cognitive awareness related to these processes (Walther, J.B., 2007).

It is a challenge that technology faces to try to minimize the gap between the social requirements and the technical feasibility (Ackerman, M. 2000). It could be that is why emoticons have survived the CMC evolution. We are presenting a table describing the affordances of each possible way in CMC (Brennan, S. E., and Lockridge, C.B. 2006).

Affordances of Media	*Face-to-face*	*Video conference*	*Telephone*	*Instant messaging or chat*	*E-mail*
(1) *Physical co-presence*: Participants share a physical environment, including a view of what each is doing and looking at.	++	??	−−	−−	−−
(2) *Visibility*: One participant sees another, but not necessarily what the other is doing or looking at.	++	+	−−	−−	−−
(3) *Audibility*: One participant can hear another.	++	++	++	−−	−−
(4) *Cotemporality*: Messages are received without delay (close to the time that they are produced and directed at addressees), permitting fine-grained interactivity.	++	??	++	??	−−
(5) *Simultaneity*: Participants can send and receive messages at the same time, allowing communication in parallel.	++	??	++	??	−−
(6) *Sequentiality*: Participants take turns in an orderly fashion in a single conversation at a time; one turn's relevance to another is signaled by adjacency.	++	++	++	−−	−−
(7) *Reviewability*: Messages do not fade over time.	−−	−−	−−	++	++
(8) *Revisability*: Messages can be revised before being sent.	−−	−−	−−	++	++

[a]Present in a particular medium: ++; present to a limited extent: +; present in some systems: ??; absent: −−. Physical co-presence (1), the hallmark of face-to-face communication, nearly always includes affordances (2) through (5).
Adapted from Clark & Brennan, 1991.

6.2 Sentiment Analysis

Like if that wasn't enough, reviews online are a lot and it would be great to be able to infer in an automated way. Recent progress in Artificial Intelligence can allow this procedures to become reality. Most specifically big data and machine learning come a long way and we now use them in many different services. Data is tamed and understood using computer and mathematical models. These models, like metaphors in literature, are explanatory simplifications. They are useful for understanding, but they have their limits. A model might spot a correlation and draw a statistical inference that is unfair or discriminatory, based on online searches, affecting the products, bank loans and health insurance a person is offered, privacy advocates warn. Applied statistics focus on estimation and have been established after years of using them. We are using power calculations to find the sample size or we are using confidence intervals to understand how much we can trust our estimations. It's important to understand that Big Data are based on prediction. We are using a very big array of data that we split in two parts, a very big one and a very small one. Training the machine learning code in the big part and testing it in the small part. Of course this is an over simplification of what machine learning is, but some of these methods are considered rules of thumb rather than applied science.

Most specifically a way to analyze text is sentiment analysis. It refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and health care materials for applications that range from marketing to customer service to clinical medicine.

In sentiment analysis the results can range from -1 to +1. While zero can be considered neutral, -1 tents to be negative and +1 positive with all the subdivisions. For the purpose of this assignment we will be using python's sentiment analysis. Certainly there are many libraries. But we are choosing the one on the NLTK (https://en.wikipedia.org/wiki/Natural_Language_Toolkit). It was developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania. They are based on **naive Bayes classifiers** that they are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Same technology has been used in spam filter in our mail boxes but even though they promise 99.5% accuracy we still receive spam e-mails. And worst sometimes we miss mail that is not spam. Therefore this is an heuristic rather than an algorithm (Lessig, 1999). Other factors should be considered, like how energy consuming are these procedures. Being content-based also translates to all of our incoming emails being scanned. We assume that this code is benign.

6.3 Structured Fraudulence

It is in a person's, who owns an online business, best interest to have positive reviews, recommendations and testimonials. Nobody more than personal ethics can prohibit to an individual to create an account and write a dithyrambic testimonial. Likewise someone can create a prestigious review site to outbound link to the original business. Here is a very recent example of it. There is a rehabilitation center in Malibu, USA called Cliffside Malibu (cliffsidemalibu.com). There is evidence that people who own the business also own two review websites for rehabilitation services. Unfortunately these results appear in google's first page for the query 'Cliffside Malibu'. The first one is thefix.com and the second one is rehabreviews.com. Someone who is considering to use their services and leave there a loved one wants to make a double check on the facilities, the staff and the business reputation in general. You can easily understand how this is misleading, scummy and unfair.

6.4 Reviews And The Rating Game

In March 2013 , Uber drivers started a protest. The reason was that they had been dropped from the platform because of low rating. It would be useful to separate the 'bad' or dangerous drivers but could that be just bad draw of customers? Josh Dziera in his 2015 article argued 'The proliferation of online feedback systems has simply turned us customers into really bad bosses'. In social science there are much reasearch in the topics on infromation cascades, conformism and network effects. The citations could be endless, from Milgram to very recent work. Even phenomena like suicide and obesity can be transmitted through the network (Christakis, N. A., & Fowler, J. H. 2009).

The "anchoring effect" names our tendency to be influenced by irrelevant numbers. Shown higher/lower numbers, experimental subjects gave higher/lower responses. As an example, most people, when asked whether Gandhi was more than 114 years old when he died, will provide a much larger estimate of his age at death than others who were asked whether Gandhi was more or less than 35 years old. Experiments show that our behavior is influenced, much more than we know or want, by the environment of the moment. (Kahneman, Daniel. 2011). There is strong evidence that people have a propensity to be biased by the ratings they have already seen, and rate a highly rated restaurand (in Yelp) based simply on the fact that the establishment has a higher rating to begin with (M.Luca 2011).

That kind of bias should be taken seriously. Could that be that a african-american man or a woman paid less? In the last work has also been found that a one-star increase in Yelp rating leads to a 5-9 percent increase in revenue. In tripadvisor (shown below) you can actually choose which of the five stars review you wish to see.

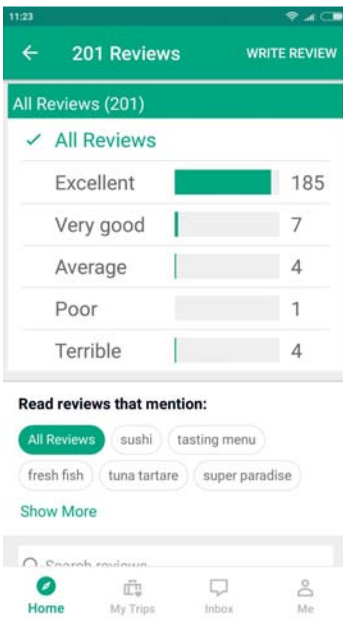


Illustration 30: TripAdvisor

The review or rating system that one chooses to use also plays an integral part in the process. For instance the five star system that TripAdvisor uses give a chance to the user to choose the 3rd star which is the median and can be considered to be average or neutral. While a four star review does not give this oppurtinity, it's either good or bad and it can only vary in these directions. There are a lot to be said about the rating systems that they are out of the scope of the present dissertation. Some complecated systems seem to make sence by unlocking previleges by earning points. For instance a new user in stackoverflow (Q&A's regarding software and technology) cannot rate other people responses, he/she has to first propose a solution that will be voted positively. The user can contributing, if they have positive effect privilages will keep unlocking.

For someone who is rating, if the rating is not obligatory, the incentives to submit an opinion could be either the user is not happy or the user is satisfied with the service. It can be agrued that one of the two cases offers more incentive to be proactive. It can also be the case that people are kind with their reviews, even though they didn't find the service satisfactory, cause they do not want to affect negatively the revenue of the business. Lastly at has been stated as social loafing or the free ride problem (Hardin, R. 2003). People want to harness the power of reviews to define their choices but after they are done they are being lazy to contribute. We are aware of this since a long time and there are efforts to motivate people, usually with emails, but with other ways as well. (Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K. Li, X., Cosley, D., Grankowski, D., Terveen, L. Rashid, A.M., Resnick, P. and Kraut, R. 2005).

Nearly 95% of Airbnb properties boast an average user-generated rating of either 4.5 or 5 stars (the maximum); virtually none have less than a 3.5 star rating. With a juxtaposition with the ratings of approximately half a million hotels worldwide that have been collected on TripAdvisor, where there is a much lower average rating of 3.8 stars, and more variance across reviews . Moreover, there is only weak correlation in the ratings of individual cross-listed properties across the two platforms. It could be the case that TripAdvisor users prefer higher-priced accommodations, while Airbnb users are more price-conscious. Yet, when comparing properties within each price segment, their relative preferences are the same. (Zervas, Georgios and Proserpio, Davide and Byers, John, 2015). Another idea, not backed up with data, is that reciprocity affects the ratings. In airbnb the host can vote for the hosted and not only the other way around, while in tripavisor it;s not the case. In the dawn of facebook likes, they seemed redundant but soon people started using them for a plethora of reasons (Veikko Eranti and Markku Lonkila 2015).

For the immunization of online reputation reporting systems against unfair ratings and discriminatory behavior two mechanisms have been proposed. The controlled anonymity and the cluster filtering (C.Dellarocas, 2000).

6.5 API Limitations

As mentioned earlier Twitter, Youtube and virtually every social network that is providing an API to the world is limiting the information over a period of time that a user is able to request. First of all they demand an authorization for pulling the data. So making sure that a specific person is being restricted to these limits. In twitter we were allowed to pull 180 tweets per 15 min, up to 3200tweets, for a keyword (aka "Rolex"). Which is exactly the amount of tweets we collected. Youtube allows again only authorized users, 50 videos per keyword. For videos that are publicly available and does not belong to the user. For the case that the account requesting with the account of the uploader is the same then different limits apply. For amazon we did not use the API they are providing. Instead we scrape their product review webpages. We used a workaround to find the ASIN we were interested for, gave the list to the script to scrape the webpages of the corresponding ASIN's. Amazon's server however has a limit of request that a user is able to do. So we set a small delay for that purpose. Amazon's server like any other server likes to know from what browser and device we are accessing the server. We have the power to ovewrite this information, we declared we are : *'User-Agent': 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:62.0) Gecko/20100101 Firefox/62.0'}.*

6.6 Segmentation

We are not using segmantation methods in our analysis. Segmentation is very useful in strategizing. Knowing how products are doing to specific crowds is very important for the marketing creative process. However we are not dealing with future strategies of the company and therefore we have not include segmentation in our paper. We might want to group the information with filters like age, race, religion, gender, family size, ethnicity, income, and education. That would be only demographic segmentation, it could also be behavioral, psychographic or geographic.

Chapter 7
Concluding Thoughts

7.1 Ideas For The Future

A competitive intelligence analysis would be necessary to carry out an analysis about Rolex. We need to create a relative index with the competitive companies to understand where this company stands. Our results where rather positive but we do not know how positive are the results or the other companies. The volume of the data, the hours that people post, the individual days that the sentiment spikes, geolocation, all sort of comparisons can be made.

One thing that can bring to the scope is having data from a bigger range. Twitter is very good since there are so many people using it and so many try to create software to analyse it but there are relative data everywhere. This paper from the analysis is missing segmentation. It would be very important if we actually had to strategize for the company to make this segmentation, between age, sex, location etc.

So an index for each product category amongst the companies that withold the majority of the market share for that product. Then individual a baseline value that we can always compare to. Of course the utlimate goal would be to have an algorithm that can download, handle and analyse the data all at once. Create a sort of index between rival companies. Setting and providing the KPI's and creating all the useful reports inside the company. Then the same software to deal with the reputation and the data outside the company. Combing that with the customer relationship management software. Creating in a way a power tool that deals with everything altogether. For instance you might use outside events to make an inventory forecast. That way these sorts of data we can collect from outside the company could affect decisions about the utilisation of the assets, the supply chain rather that future marketing strategies which is the more obvious and already widely used.

7.2 Conclusion

Reputation is the collective judgements of an entity that people are aware of. By being aware of the entity automatically we have an opinion. Even though it can be only based on the visual impact of the logo or the words on their slogan.

On the other hand as entity it can be also a wide range of things, like individuals, places, political parties, countries, businesses, e-shops etc. As individuals we have self-refective conciusness something that philophers from Pluto to Descartes we acknowledge our existance and we try to shape the way other people think of us. We avoid to say things that are embarrassing for us. Goffman has been describing and exploring how people try to manage these impressions on an individual level. A political party or a corporation is only an extent of that. A company can be viewed as one single brain that is comprised by single units (individuals), parts of the brain, for different functions. Difference is that business usually can be treated in a more quantifiable way, as with bad reputation damage can be estimated numerically, or with a good campaign revenue can be increased. Business is about making profit, reputation can interfere to that process and that is why it is so important in a corporate level. People seem to be opposed big corporations but like the individual brands that comprise them. It has to be that way in order for these brands to be profitable. We might not understand as consumers that this is the way to make the corporation profitable as well.

Our social nature, needs us to socialize a lot and group ourselves in to homogenic groups. This socialization process that makes us more unified, also spreads and dessiminates ideas and information. That information may include the reputation of a brand. This is the reason that everybody's opinion should matter. Of course companies take mostly input from their clients, firstly because they are more available, they visit their website to purchaise item for example, and secondly because they are actually more important, since they are the ones producing the revenue. Nevertheless a potential client or a person who can influence potential clients can be quite important. For instance an activist. As a for profit company who would not care to make him (the activist) their client, as the acquiring client cost would be unsustainable, but they definitely do not need a campaign against the business. As we have seen reputation can be hurt easily.

On the other hand leveraging social media can be an invaluable asset. Either commencing a campaign that can become viral and create revenue, or foresee a trend that is coming and design a product to accomodate it. This is where social media analytics come in to play. We need to be able to identify these insights and we need to be able to inference the data our users produce. But that demands a know-how in terms of programming and some good analysts. Which is good because more complexity more options, but also time consuming and complicated. This need has let a lot of people to offer ORM (Online Reputation Management) solutions. Especially having to deal with data that are coming from a lot of different sources.

This social nature we spoke of, is also fairly complex. There are a lot of ways we are communicating facial expressions, prosody, body language and all sorts of nuances. Information that our technology cannot always afford. As a concequence this part of the information is lost through the transmission. The way we evolved socially could be different. Proof is that in different areas and groups different social norms are applied. There is a technological gap between the face to face communication and CMC (Computer Mediated Communication). We also discussed additional technological limitations that are not related to our social nature. So the system we created through the web and mobile communication is created because of the way we are but is not translating everything exactly, information are often missed or distorted (noise). Technology also gives the power to be heard to people that did not have it before, that for some advocates exposes the sillyness of the human nature and is mostly responsible for the post-truth era.

We have argued that the type of the company or the product can play an integral role to the way the business is interpreted in the online world or the new era. Just because a pleathora of businesses can be benefited by a new advancement does not mean it can be applicable in all of the cases. Interesting are the cases of sharing economy, where a lot of writers think it can change the way we produce and are getting paid.

We chose and briefly examined the case of Rolex. We indeed found that is a company with a good reputation. But that can be only relative. It is important to point out that all the data we collected are publicly available and if the scripts are working well and are optimised for a watch product analysis then they should be applied to all of the competitive companies. That way a competitive analysis can be very useful in terms of relativity and future strategy decisions. If we were actually Rolex we could have more data to analyse through the official twitter account but also real data about the company's balance sheet that could be probably correlated with events from the aggregated data.

Markets have been definitely the moving power. In this paper we have not discussed a lot the marketing aspect of it all. Online marketing has evolved amazingly, companies make auctions to buy keyword to advertise their products and services. Companies make millions selling data that is harvested from their clients, like Facebook and Linkedin and data is undeniably a very good source of income for these companies. We are also expecting that besides markets and the giants of the web, other individual companies, individuals, academic institutions will provide the way to improve, implement and understand the way things are being done online today. These initiatives, good architecture and design, good practices based on laws that are made to provide freedom yet safety to personal data and are recently revised. We expected them to make a good change.

BIBLIOGRAPHY

- Goffman, E. (1956). *The presentation of self in everyday life*. Edinburgh, University of Edinburgh. [1]

- Lawrence Lessig, (1999). Code 2.0 [2]

- Easley D. And Kleinberg J. (2010) Networks, crowds and markets: Reasoning about a highly connected world. [3]

- Goffman, Erving (1967): On Face-Work. An Analysis of Ritual Elements in Social Interaction. In: Ders.: Interaction Ritual. New York: Doubleday. 5-45. [4]

- Walther, J.B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language and cognition. Computers in Human Behavior 23: 2538-2557 [5]

- Hancock, J.T. (2007). Digital deception: Why, when and how people lie online. In: Joinson, A., McKenna, K., Postmes, T. and Reips, U.-D. (eds.) *The Oxford Handbook of Internet Psychology*. Oxford: Oxford University Press. [6]

- Barnett, Michael L. and Jermier, John and Lafferty, Barbara A., Corporate Reputation: The Definitional Landscape. Corporate Reputation Review, Vol. 9, No. 1, 2006. Available at SSRN: https://ssrn.com/abstract=868492 [7]

- G. W. Furnas and T. K. Landauer and L. M. Gomez and S. T. Dumais: The vocabulary problem in human-system communication, Communications of the ACM Volume 30 Issue 11, Nov. 1987 [8]

- Melewar, T.C. and Jenkins, E. (2002) 'Defining the corporate identity construct', Corporate Reputation Review, 5(1), 76-90 [9]

- Gray , E.R. and Balmer, J.M.T. (1998) 'Managing corporate image and corporate reputation', Long Range Planning, 315, 695– 702 . [10]

- https://www.technologyreview.com/s/607938/google-now-tracks-your-credit-card-purchases-and-connects-them-to-its-online-profile-of-you/ , May 25, 2017 [11]

- https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf [12]

- AL **Barabási** (2016), http://networksciencebook.com/ [13]

- Zhang H., Goel A., Govindan R., Mason K., Van Roy B. (2004) Making Eigenvector-Based Reputation Systems Robust to Collusion. In: Leonardi S. (eds) Algorithms and Models for the Web-Graph. WAW 2004. Lecture Notes in Computer Science, vol 3243. Springer, Berlin, Heidelberg [14]

- Walther, J.B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language and cognition. Computers in Human Behavior 23: 2538-2557 [15]

- Ackerman, M. (2000). The intellectual challenge of computer-supported cooperative work: the gap between social requirements and technical feasibility. *Human-Computer Interaction* 15: 179-203. [16]

- https://s3.amazonaws.com/academia.edu.documents/34393761/2_The_New_York_Times_on_The_Age_of_Big_Data.pdf [17]

- https://en.wikipedia.org/wiki/Sentiment_analysis [18]

- https://www.theverge.com/2017/11/3/16601668/rehab-center-review-sites-top-rated-bias-rehabreviews-thefix [19]

- Introna, L.D. and Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. The Information Society 16(3): 169-185. [20]

- Brennan, S. E., and Lockridge, C.B. (2006). Computer-mediated communication: A cognitive science approach." *ELL2, Encyclopedia of Language and Linguistics*: 775-780 [21]

- Arun Sunjararajan 2016, The sharing Economy , The end of employment and the rise of crowd-based capitalism MIT press [22]

- https://www.independent.co.uk/news/business/comment/hamish-mcrae/facebook-airbnb-uber-and-the-unstoppable-rise-of-the-content-non-generators-10227207.html [23]

- Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks* and how they shape our lives. [24]

-Kahneman, Daniel. (2011) Thinking, Fast and Slow. New York: Farrar, Straus and Giroux,. [25]

-M Luca (2011) - Reviews , Reputation, and Revenue: The Case of Yelp .com [26]

- Hardin, R. (2003) The Free Rider Problem. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2003 Edition) [27]

- Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K. Li, X., Cosley, D., Grankowski, D., Terveen, L. Rashid, A.M., Resnick, P. and Kraut, R. (2005). Using social psychology to motivate contributions to online communities. Journal of Computer-Mediated Communication 10(4). [28]

-Zervas, Georgios and Proserpio, Davide and Byers, John, (2015) A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average. [29]

- C Dellarocas (2000) - Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior [30]

-Veikko Eranti and Markku Lonkila 2015 - The social significance of the Facebook Like button. [31]

- . Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management [32]

-https://www.reputationinstitute.com [33]

- www.AssessmentDay.co.uk [34]

- https://www.washingtonpost.com/news/business/wp/2017/10/08/dove-ad-that-shows-a-black-woman-turning-herself-white-sparks-consumer-backlash/?noredirect=on&utm_term=.6b6f6ab93375 [35]

-https://en.wikipedia.org/wiki/Rolex [36]

- https://www.microsoft.com/en-us/research/wp-content/uploads/2016/12/2012-01-0489-SAE-published.pdf [37]

-Lisa Gansky 2010. The Mesh: Why the Future of Business Is Sharing [38]

- https://www.bloomberg.com/news/articles/2018-08-23/what-s-happening-to-subscription-watch-club-eleven-james [39]

- https://www.knoyd.com/blog/amazon-review-scraper [40]

-https://www.koreus.com/video/malvoyant-chien-guide-monoprix.html [41]

APPENDIX

Code that takes as input a file with html result pages from **amazon** and return a list of the asins

```
#! /usr/bin/python
# -*- coding: utf-8 -*-

import re
amazon = open("amazon.txt","r")
#mind that we make should for each line on the file we load there is no more than 1 time the matching regex
#the re.search if find an instance in an element of the loop moves
#to the next line regardless if they are other instances on the same line

for whatever in amazon:
    asin=re.search(r'asin="(.*?)"', whatever, re.IGNORECASE)
    #print type(whatever)
    if asin:
```

```
        print asin.group(1)
amazon.close()
```

Takes input list of asins and scrapes **amazon** reviews, stores in dataframes, produces insightful visualizations

```
#! /usr/bin/python3

import os
import matplotlib as mpl
import matplotlib.image as mpimg
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import requests
from lxml import html
from sklearn.feature_extraction.text import CountVectorizer
from stop_words import safe_get_stop_words
from wordcloud import WordCloud


# Scrape and Analyse Amazon Reviews

def scrape_reviews(asins):
    ratings_dict = {}
    reviews_list = []
    reviews_df = pd.DataFrame()

    headers = {
        'User-Agent': 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:62.0) Gecko/20100101 Firefox/62.0'}
    XPATH_REVIEWS = '//div[@data-hook="review"]'
    XPATH_REVIEW_RATING = './/i[@data-hook="review-star-rating"]//text()'
    XPATH_REVIEW_HEADER = './/a[@data-hook="review-title"]//text()'
    XPATH_REVIEW_AUTHOR = './/a[@data-hook="review-author"]//text()'
    XPATH_REVIEW_DATE = './/span[@data-hook="review-date"]//text()'
    XPATH_REVIEW_BODY = './/span[@data-hook="review-body"]//text()'
```

```
 XPATH_REVIEW_HELPFUL = './/span[@data-hook="helpful-vote-statement"]//text()'

# p_num = 0
 for asin in asins:
    p_num=0
    while True:
        print ('Scraping review page nr. {}'.format(p_num))
        amazon_url = 'https://www.amazon.com/product-reviews/' + asin + '?pageNumber=' + str(p_num) + '&sortBy=recent'
        # Add some recent user agent to prevent amazon from blocking the request
        # Find some chrome user agent strings  here https://udger.com/resources/ua-list/browser-detail?browser=Chrome
        page = requests.get(amazon_url, headers=headers)
        page_response = page.text.encode('utf-8')

        parser = html.fromstring(page_response)

        reviews = parser.xpath(XPATH_REVIEWS)
        if not len(reviews) > 0:
            break
        # Parsing individual reviews
        for review in reviews:
            raw_review_author = review.xpath(XPATH_REVIEW_AUTHOR)
            raw_review_rating = review.xpath(XPATH_REVIEW_RATING)
            raw_review_header = review.xpath(XPATH_REVIEW_HEADER)
            raw_review_date = review.xpath(XPATH_REVIEW_DATE)
            raw_review_body = review.xpath(XPATH_REVIEW_BODY)
            raw_review_helpful = review.xpath(XPATH_REVIEW_HELPFUL)

            review_dict = {
                'review_text': raw_review_body,
                'review_posted_date': raw_review_date,
                'review_header': raw_review_header,
                'review_rating': raw_review_rating,
                'review_helpful': raw_review_helpful,
                'review_author': raw_review_author
```

```
            }
            reviews_df = reviews_df.append(review_dict, ignore_index=True)
        p_num += 1
        if p_num > 7:
            break
    return reviews_df
```

```
## Scrape or load reviews
filename = 'echo_dot_reviews.pickle'
asins =
['B07C6LPX12','B01MR0ROVO','B06XT59GNB','B013UT2BK4','B07F24DP3G','B078N4392Q','B00BLIDTV2','B07DL9R61
M','B00COQKLR0','B01KHJP418','B074JGP3PN','B07G5NYDLR','B008GR4QHC','B075R38MXV','B01F7T52LA','B015ST
OINC','B076D1M9ZZ','B073498Y9X','B07HXHWRNW','B01LXLO5P8','B01KOTY0MK','B06XNVBHCZ','B00WZN5HIA','
B001VDE984','B079B9RKGB','B075DK2QN4','B01K9DTCGK','B077NVNXRS','B0035WSYS0','B075QPW81Z','B013L550
V4','B072YWZXM8','B079QTG9KT','B00PIUSLZ8','B075DJXM7W','B07BKRFYC1','B00BCYM8ZS','B015B5YNDS','B008
84L67G','B0744ZSSYN','B073DDMCC3','B01MZ04K04','B07G8SPNYN','B07CJY37R1','B071HSCFDD','B00BCYJ8K6','B0
0K78HP2Q','B009TBCYIW','B00B286PEY','B00CM1IKV6','B001TDITJG','B017KQ35L2','B001115D9K','B008GXOOWS','
B00F9SYOW8','B013L5597E','B00C5O7XXG','B075DH1C9X','B07DPHF1H5','B07BBWQH63','B0725F1PMY','B00UB503
HY','B0134XKHCK','B07G8PJ5V8','B006WV17EY','B01KDOSVC6','B00EKLQA4A','B01KJSRPXW','B07H41WN85','B00
BCYJMZC','B075DJVBXD','B00ML5AGRY','B00C5O86GO','B0041GJ3YI','B0759ML77C','B07DPJM99X','B00XZ69V92','
B01MUBF73X','B0066SFUPK','B00THF1SZ0','B00C5OJC8U','B01K85LP0U','B07DVJ4SKR','B01IT0OXV4','B07H4684CV'
,'B00VAUU3GU','B07FVGHY49','B077CZWYYG','B00BCYKTYK','B007GYKB7A','B073XX9M55','B001MSOW6M','B01J
P9BHLM','B00RDK34VW','B06X8ZL7H8','B00LW9E1R6','B075X1HMB4','B01LNANLD8','B01LYA4IXD','B07FB5XRYG'
,'B01MRV9V1J','B016YGAJJU','B019L8UDWG','B003KV5UJM','B00KC9U2SE','B01MY2ZBN5','B075QPG21G','B0042SN
ZL2','B00VTWOW9S','B076PVH6FS','B079XXZ1GB','B00K78KCI0','B00IISUN6C','B078WXJF1J','B00P9VMS5U','B074T9
T5VL','B07DQRN58W','B01KJSY9SQ','B00K78FZQ4','B016N56F54','B0174D17D4','B00PWPO1OO','B07DLC5W2D','B07
GL1WQS4','B01B3KQZ22','B077XJP2Y6','B07F8ZLJPX','B01KOUYJDO','B00ZDA1NJE','B004GYXO9U','B00K7865HC','
B078G5G35G','B01KDMWMG4','B01KIGRD8C','B07859ND2V','B07G5N6915','B019S6UQPA','B009AQADWK','B07DV3
K6R4','B004OMIJ0I','B00K786MAW','B01KTWINF2','B07F1H9J5K','B00COQKM4C','B07BGHKWJF','B00BD5I96S','B07
DKYJKPB','B01KHIVNPU','B076PTTPFF','B00AT85HOW','B076NCVQNH','B075G3H9PY','B0069AICIM','B00A9C3ZTM'
,'B01NASAVIT','B00A8DD49S','B00BLID5BG','B0131LSWRM','B00K78OIVW','B073V8Z7R1','B07F2PDFTW','B07F3CP4
3P','B00K40TH9G','B01KJSLJ1Q','B075DHLR5X','B074XHM8GW','B007HBM97C','B07DWDYK8S','B076KYFB7R','B01E
IIQ79W','B00K78CBT8','B01KIG39OY','B07CHRCRBC','B01KJPQ64E','B079J5GP45','B07GHBPK59','B07GTWSD74','B07
9DCML2M','B07GCSHHB4','B01IRBRF06','B007FKF7BA','B01KDSMD4Y','B07FXVSLXX','B003KV2GTO','B01KAXZI3
0','B00BPFJQD6','B07FSS8ZMG','B00ZB2T0L2','B07HGH5VD3','B01KN40Y1C','B01JMN1QZI','B00C5O84H0','B07H13B3
38','B0773FVR8N','B07GDVG8C4','B07CK3VF7B','B00C5OMHPK','B07H7VB9YK','B00TECQ0JU','B07FPVJD36','B07G8
SBTVH','B00O3GU4Z8','B017KQ3EN6','B078HMN7YZ','B01KDT1RWC','B00K78I0E8','B002ZUGCB4','B00ZB2Y83C','B0
7D5GZVXM','B014DQ8KRG','B07DLDT3P5','B00K78A3WA','B017KQ3BWA','B01AGOQRL0','B017KQ37SS','B019YGJY
RK','B075VFZRKK','B074HWSRBP','B00BPFJNHK','B079P7F6XS','B000SSKZT0','B07F2R4CPT','B00C5OUWTS','B00BP
FJNEI','B07H3BN7GS','B01KTYSRQ0','B01K2NLXBE','B079G7GWR7','B01KIGL1MQ','B07361SGWL','B07GL328K1','B0
0DKS2KK2','B07HKY6ZBH','B00375DTC6','B07GHXCXKT','B07GH6JPX1','B07GDVBD4H','B00BLIFNR0','B07GC56FR
X','B07GTTBKS6','B07GX4TQVG','B0079GCK9W','B00K78NZV6','B07C6LPX12','B07G8HZFKN','B007D4TGF6','B07CL
```

```
BBHSH','B07H8PQSJ9','B07GDT5BR1','B07GDR145P','B07GTTFCZ3','B00K78KO4W','B07DLD5VH5','B07G8QW2HZ','B
01KEW8UIW','B00BPFJGRC','B00K78FBMC','B07GDT5W67','B01KJKURVC','B07CHZDX6C','B07G7G9FG1','B07G8QR
945','B00APWNYGU','B01KHKS4JQ','B01KIFVBX6','B01N30HUQC','B07HBMFJDF','B00K78AKCS','B07CL5LZ9M','B00
KC9NOOI','B0761JVHB2','B07DLC654K','B075ZF92S1','B077XP1BPL','B00K787SOG','B00K787LOS','B079SM5Z96','B07
F2N8ZM4','B07GL2V6PY','B01KDNWG0U','B0781962P8','B07H9HZ3BD','B07448MTPP','B01KOUAM26','B07FYXK4P4','
B00XNPH7SM','B07HFKNQ9D','B01MS4SBP8','B076JGCT6N','B075QP97CK','B07GGG32Z9','B07G9N2D53','B0762BTS
M7','B01K88A3MI','B00K78DU6Q','B07922R4KQ','B01MYMPYNI','B07GFRKGK9','B00KC9UBVC','B000KL5D24','B077
NVNXRS','B07H4ZNT59','B00BPFJUM8','B07DLF5V4S','B07CJZKMSR','B07GFM13N5','B07HZN3CSG','B07GGCB3ZF']

reviews_df = pd.DataFrame()
if os.path.isfile(filename):
    print ('Loading reviews from disk')
    reviews_df = pd.read_pickle(filename)
    print ('Loaded {} reviews'.format(len(reviews_df)))
else:
    print ('Scraping reviews')
    reviews_df = scrape_reviews(asins)
    print ('Scraped {} reviews'.format(len(reviews_df)))
    print ('Saving reviews to disk')
    reviews_df.to_pickle(filename)


## Data cleaning and preprocessing


for col in reviews_df.columns:
    reviews_df[col] = reviews_df[col].apply(lambda x: '\n'.join(x))

reviews_df['review_helpful'] = (reviews_df['review_helpful']
                                .str.replace('One', '1')
                                .str.replace(r'[^0-9]', ''))
reviews_df['review_helpful'].loc[reviews_df['review_helpful'] == ''] = '0'
reviews_df['review_helpful'] = reviews_df['review_helpful'].astype(int)
```

```
reviews_df['review_posted_date'] = pd.to_datetime(reviews_df['review_posted_date']
                                   .str.strip('on'))

reviews_df['review_rating'] = reviews_df['review_rating'].str.strip('out of 5 stars').astype(float)
reviews_df.loc[reviews_df['review_rating'] == 0, 'review_rating'] = 5

reviews_df['review_length'] = reviews_df['review_text'].apply(lambda x: len(x))

reviews_df.drop_duplicates(inplace=True)



## Histogram of Reviews' Lengths



mpl.rcParams.update(mpl.rcParamsDefault)
plt.style.use('seaborn-white')
plt.figure(figsize=(15, 5))
mpl.rc('xtick', labelsize=15)
mpl.rc('ytick', labelsize=15)

axes = reviews_df['review_length'].hist(bins=1000)
axes.set_xlim(0, 2500)
plt.title('Histogram of Review Lengths', fontsize=20)
plt.xlabel('Length of the Review in Characters', fontsize=18)
plt.ylabel('Number of Review', fontsize=18)
plt.show()



## Histogram of Logarithm of Reviews' Lengths in Characters

plt.style.use('seaborn-white')
```

```
plt.figure(figsize=(15, 5))
mpl.rc('xtick', labelsize=15)
mpl.rc('ytick', labelsize=15)
axes = np.log(reviews_df['review_length'].loc[reviews_df['review_length'] > 0]).hist(bins=200)
plt.title('Histogram of Logarithm of Review Lengths in Characters', fontsize=20)
plt.xlabel('Logarithm of Review Length in Characters', fontsize=18)
plt.ylabel('Number of Review', fontsize=18)
plt.show()


## Mode of Review Lengths


reviews_df['review_length'].mode()


## Histogram of Ratings


plt.style.use('seaborn-white')

fig = plt.figure()
fig.set_size_inches(15, 5)
a = fig.add_subplot(1, 2, 1)
#img = mpimg.imread('amz_hist_ratings.png')
#lum_img = img
#imgplot = plt.imshow(lum_img)
plt.title('Histogram of Ratings on the Amazon Page', fontsize=16)

a.axis('off')
a = fig.add_subplot(1, 2, 2)

mpl.rc('xtick', labelsize=15)
mpl.rc('ytick', labelsize=15)
```

```
a.hist(reviews_df['review_rating'], bins=np.arange(1, 7) - 0.5, orientation="horizontal")

plt.title('Histogram of Scraped Ratings', fontsize=16)
plt.ylabel('Star Rating', fontsize=16)
plt.xlabel('Number of Reviews', fontsize=16)

plt.show()



## Histogram of Review Dates



plt.style.use('seaborn-white')
plt.figure(figsize=(15, 5))
mpl.rc('xtick', labelsize=15)
mpl.rc('ytick', labelsize=15)
axes = reviews_df['review_posted_date'].hist(bins=200)
plt.title('Histogram of Review Dates', fontsize=20)
plt.xlabel('Review Date', fontsize=18)
plt.ylabel('Number of Review', fontsize=18)
plt.show()



## Length of the review vs. the number of people who found it helpful



reviews_df['log_review_helpful'] = None
reviews_df['log_review_helpful'].loc[reviews_df['review_helpful'] > 0] = np.log(
    reviews_df['review_helpful'].loc[reviews_df['review_helpful'] > 0])
reviews_df['log_review_helpful'] = reviews_df['log_review_helpful'].astype(float)

reviews_df['log_review_length'] = None
```

```
reviews_df['log_review_length'].loc[reviews_df['review_length'] > 0] = np.log(
    reviews_df['review_length'].loc[reviews_df['review_length'] > 0])
reviews_df['log_review_length'] = reviews_df['log_review_length'].astype(float)

plt.style.use('seaborn-white')
fig = plt.figure()
fig.set_size_inches(12, 5)
mpl.rc('xtick', labelsize=15)
mpl.rc('ytick', labelsize=15)
a = fig.add_subplot(1, 1, 1)
a = reviews_df.dropna(subset=['log_review_length', 'log_review_helpful']).plot.scatter('log_review_length',
                                                                'log_review_helpful', ax=a)

plt.title('Length of the review vs. the number of people who found it helpful', fontsize=20)
plt.xlabel('Log Review Length', fontsize=18)
plt.ylabel('Log Review Helpful', fontsize=18)
plt.show()


## Word clouds


def make_word_cloud(df, ngram_min, ngram_max):
    plt.style.use('seaborn-white')

    fig = plt.figure()
    fig.set_size_inches(15, 15)

    for ind, row in df.iterrows():
        data = row['review_text']
        min_df = int(max([min([9, 0.04 * len(data)]), 2]))
        num_words = 200
        ngram_range = (ngram_min, ngram_max)
```

```
        count_vectorizer = CountVectorizer(min_df=min_df,
                                           lowercase=True,
                                           stop_words=stop_words,
                                           ngram_range=ngram_range)
        counts = count_vectorizer.fit_transform(data)
        counts = counts.toarray().sum(axis=0)
        count_weighting = dict(zip(count_vectorizer.get_feature_names(), counts))
        count_weighting_df = pd.DataFrame.from_dict(count_weighting, orient='index')
        count_weighting_df = count_weighting_df.reset_index(drop=False)
        count_weighting_df.columns = ['word', 'count']

        count_weighting_df = count_weighting_df.sort_values(['count'], ascending=False)
        count_weighting_df = count_weighting_df.set_index('word')

        word_cloud_freq = count_weighting_df['count'].head(num_words).to_dict()
        wordcloud = WordCloud(collocations=False).generate_from_frequencies(word_cloud_freq)
        a = fig.add_subplot(1, 2, ind + 1)
        plt.title(str(row['rating_text']), fontsize=20)
        plt.imshow(wordcloud, cmap=plt.cm.bone, interpolation='bilinear')
        plt.axis("off")
    plt.show()

stop_words = safe_get_stop_words('en')
stop_words.extend(['rolex',
                   'Rolex watches',
                   'echo',
                   'dot',
                   'can',
                   'just',
                   'work',
                   'music',
                   'one',
                   'get',
                   'will',
```

```
        'time',
        'speaker',
        'use',
        'app',
        'google',
        'home',
        ])

reviews_df['rating_text'] = None
reviews_df['rating_text'].loc[reviews_df['review_rating'] < 4] = '1, 2 or 3 stars'
reviews_df['rating_text'].loc[reviews_df['review_rating'] >= 4] = '4 or 5 stars'
reviews_by_rating = reviews_df.groupby('rating_text')['review_text'].apply(list)
reviews_by_rating = reviews_by_rating.reset_index(drop=False)
```

```
make_word_cloud(reviews_by_rating, 1, 1)

make_word_cloud(reviews_by_rating, 2, 2)

make_word_cloud(reviews_by_rating, 3, 3)
```

Twitter :

```
#! /usr/bin/python3
# -*- coding: utf-8 -*-

import argparse
```

```
from urllib.parse import urlparse
import urllib
import csv
import tweepy
import time


# URL CLEANUP
def url_fix(s, charset='utf-8'):
    if isinstance(s, unicode):
        s = s.encode(charset, 'ignore')
    scheme, netloc, path, qs, anchor = urlparse.urlsplit(s)
    path = urllib.quote(path, '/%')
    qs = urllib.quote_plus(qs, ':&=')
    return urlparse.urlunsplit((scheme, netloc, path, qs, anchor))

# COMMAND PARSER
def tw_parser():
    global qw, ge, l, t, c, d

# USE EXAMPLES:
# =-=-=-=-=-=-=
# % twsearch <search term>            --- searches term
# % twsearch <search term> -g sf      --- searches term in SF geographic box <DEFAULT = none>
# % twsearch <search term> -l en      --- searches term with lang=en (English) <DEFAULT = en>
# % twsearch <search term> -t {m,r,p} --- searches term of type: mixed, recent, or popular <DEFAULT = recent>
# % twsearch <search term> -c 12      --- searches term and returns 12 tweets (count=12) <DEFAULT = 1>
# % twsearch <search term> -o {ca, tx, id, co, rtc}   --- searches term and sets output options <DEFAULT = ca, tx>

# Parse the command
    parser = argparse.ArgumentParser(description='Twitter Search')
    parser.add_argument(action='store', dest='query', help='Search term string')
    parser.add_argument('-g', action='store', dest='loca', help='Location (lo, nyl, nym, nyu, dc, sf, nb')
    parser.add_argument('-l', action='store', dest='l', help='Language (en = English, fr = French, etc...)')
```

```
parser.add_argument('-t', action='store', dest='t', help='Search type: mixed, recent, or popular')
parser.add_argument('-c', action='store', dest='c', help='Tweet count (must be <50)')
args = parser.parse_args()

qw = args.query     # Actual query word(s)
ge = ''

# Location
loca = args.loca
if (not(loca in ('lo', 'nyl', 'nym', 'nyu', 'dc', 'sf', 'nb')) and (loca)):
    print ("WARNING: Location must be one of these: lo, nyl, nym, nyu, dc, sf, nb")
    exit()
if loca:
    ge = locords[loca]

# Language
l = args.l
if (not l):
    l = "en"
if (not(l in ('en','fr','es','po','ko', 'ar'))):
    print ("WARNING: Languages currently supported are: en (English), fr (French), es (Spanish), po (Portuguese), ko (Korean), ar (Arabic)")
    exit()

# Tweet type
t = args.t
if (not t):
    t = "recent"
if (not(t in ('mixed','recent','popular'))):
    print ("WARNING: Search type must be one of: (m)ixed, (r)ecent, or (p)opular")
    exit()

# Tweet count
if args.c:
```

```
        c = int(args.c)
        if (c > cmax):
            print ("Resetting count to ",cmax," (maximum allowed)")
            c = cmax
        if (not (c) or (c < 1)):
            c = 1
    if not(args.c):
        c = 1

    print ("Query: %s, Location: %s, Language: %s, Search type: %s, Count: %s" %(qw,ge,l,t,c))

# AUTHENTICATION (OAuth)
def tw_oauth(authfile):
    with open(authfile, "r") as f:
        ak = f.readlines()
    f.close()
    auth1 = tweepy.auth.OAuthHandler(ak[0].replace("\n",""), ak[1].replace("\n",""))
    auth1.set_access_token(ak[2].replace("\n",""), ak[3].replace("\n",""))
    return tweepy.API(auth1)

def tw_search_json(query, cnt=5):
    authfile = './auth.k'
    api = tw_oauth(authfile)
    results = {}
    meta = {
        'username': 'text',
        'usersince': 'date',
        'followers': 'numeric',
        'friends': 'numeric',
        'authorid': 'text',
        'authorloc': 'geo',
        'geoenable': 'boolean',
        'source': 'text'
```

```
    }
    data = []
    for tweet in tweepy.Cursor(api.search, q=query, count=cnt).items():
        dTwt = {}
        dTwt['username'] = tweet.author.name
        dTwt['usersince'] = tweet.author.created_at      #author/user profile creation date
        dTwt['followers'] = tweet.author.followers_count #number of author/user followers (inlink)
        dTwt['friends']   = tweet.author.friends_count   #number of author/user friends (outlink)
        dTwt['authorid']  = tweet.author.id              #author/user ID#
        dTwt['authorloc'] = tweet.author.location        #author/user location
        dTwt['geoenable'] = tweet.author.geo_enabled     #is author/user account geo enabled?
        dTwt['source']    = tweet.source                 #platform source for tweet
        data.append(dTwt)
    results['meta'] = meta
    results['data'] = data
    return results


# TWEEPY SEARCH FUNCTION
def tw_search(api):
    counter = 0
    # Open/Create a file to append data
    csvFile = open('result.csv','w')
    #Use csv Writer
    csvWriter = csv.writer(csvFile)
    csvWriter.writerow(["created", "text", "retwc", "hashtag", "followers",
"friends","tweet_id","cords","source","if_link","username","followers","authorloc","lang"])

    for tweet in tweepy.Cursor(api.search,
                    q = qw,
                    g = ge,
                    lang = l,
                    result_type = t,
                    count = c).items():
```

```
#TWEET INFO
created = tweet.created_at   #tweet created
text    = tweet.text         #tweet text
tweet_id = tweet.id          #tweet ID# (not author ID#)
cords   = tweet.coordinates  #geographic co-ordinates
retwc   = tweet.retweet_count #re-tweet count
lang   = tweet.lang

try:
   hashtag = tweet.entities[u'hashtags'][0][u'text'] #hashtags used
except:
   hashtag = "None"
try:
   rawurl = tweet.entities[u'urls'][0][u'url'] #URLs used
   urls = url_fix(rawurl)
except:
   urls    = "None"
#AUTHOR INFO
username  = tweet.author.name            #author/user name
usersince = tweet.author.created_at      #author/user profile creation date
followers = tweet.author.followers_count #number of author/user followers (inlink)
friends   = tweet.author.friends_count   #number of author/user friends (outlink)
authorid  = tweet.author.id              #author/user ID#
authorloc = tweet.author.location        #author/user location
#TECHNOLOGY INFO
geoenable = tweet.author.geo_enabled     #is author/user account geo enabled?
source    = tweet.source                 #platform source for tweet
                  # Dongho 03/28/16

if_link=False
if 'https' in text:
   if_link=True
time.sleep(1)
```

```
        csvWriter.writerow([created, str(text).encode("utf-8"), retwc, hashtag, followers, friends,tweet_id,cords,source,if_link,username,followers,authorloc,lang])
        counter = counter +1
        print ('...'+'tweets collected:'+str(counter))



        if (counter == c):
            break

    csvFile.close()

# MAIN ROUTINE
def main():

    global api, cmax, locords

    # Geo-coordinates of five metropolitan areas
    # London, NYC (lower, middle, upper), Wash DC, San Francisco, New Brunswick (NJ)
    locords =  {'lo': '0, 51.503, 20km',
            'nyl': '-74, 40.73, 2mi',
            'nym': '-74, 40.74, 2mi',
            'nyu': '-73.96, 40.78, 2mi',
            'dc': '-77.04, 38.91, 2mi',
            'sf': '-122.45, 37.74, 5km',
            'nb': '-74.45, 40.49, 2mi'}
    # Maximum allowed tweet count (note: Twitter sets this to ~180 per 15 minutes)
    cmax = 3200
    # OAuth key file
    authfile = './auth.k'

    tw_parser()
```

```
    api = tw_oauth(authfile)
    tw_search(api)
```

Youtube:

```
#! /usr/bin/python
# -*- coding: utf-8 -*-

from apiclient.discovery import build
#from apiclient.errors import HttpError
#from oauth2client.tools import argparser # removed by Dongho
import argparse
import csv
import unidecode
import time

# Set DEVELOPER_KEY to the API key value from the APIs & auth > Registered apps
# tab of
#   https://cloud.google.com/console
# Please ensure that you have enabled the YouTube Data API for your project.
DEVELOPER_KEY = "AIzaSyAZOsUnOsZHjq8suzD1ExekEh88Ljv5Pzk"
YOUTUBE_API_SERVICE_NAME = "youtube"
YOUTUBE_API_VERSION = "v3"



def youtube_search(options):
    youtube = build(YOUTUBE_API_SERVICE_NAME, YOUTUBE_API_VERSION, developerKey=DEVELOPER_KEY)
    # Call the search.list method to retrieve results matching the specified
    # query term.
    search_response = youtube.search().list(q=options.q, part="id,snippet", maxResults=options.max_results).execute()

    videos = []
```

```
channels = []
playlists = []

# create a CSV output for video list
csvFile = open('video_result.csv','w')
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["title","videoId","viewCount","likeCount","dislikeCount","commentCount","favoriteCount"])

# Add each result to the appropriate list, and then display the lists of
# matching videos, channels, and playlists.
for search_result in search_response.get("items", []):
  if search_result["id"]["kind"] == "youtube#video":
    #videos.append("%s (%s)" % (search_result["snippet"]["title"],search_result["id"]["videoId"]))
    title = search_result["snippet"]["title"]
    title = unidecode.unidecode(title)  # Dongho 08/10/16
    videoId = search_result["id"]["videoId"]
    video_response = youtube.videos().list(id=videoId,part="statistics").execute()


    for video_result in video_response.get("items",[]):
      viewCount = video_result["statistics"]["viewCount"]
      if 'likeCount' not in video_result["statistics"]:
        likeCount = 0
      else:
        likeCount = video_result["statistics"]["likeCount"]
      if 'dislikeCount' not in video_result["statistics"]:
        dislikeCount = 0
      else:
        dislikeCount = video_result["statistics"]["dislikeCount"]
      if 'commentCount' not in video_result["statistics"]:
        commentCount = 0
      else:
        commentCount = video_result["statistics"]["commentCount"]
      if 'favoriteCount' not in video_result["statistics"]:
```

```
            favoriteCount = 0
          else:
            favoriteCount = video_result["statistics"]["favoriteCount"]




        csvWriter.writerow([title,videoId,viewCount,likeCount,dislikeCount,commentCount,favoriteCount])

    csvFile.close()

if __name__ == "__main__":
    parser = argparse.ArgumentParser(description='Search on YouTube')
    parser.add_argument("--q", help="Search term", default="Google")
    parser.add_argument("--max-results", help="Max results", default=50)
    args = parser.parse_args()
    #try:
    youtube_search(args)
    #except HttpError, e:
    #    print ("An HTTP error %d occurred:\n%s" % (e.resp.status, e.content))
```

ASIN'S

B07C6LPX12 B013L550V4 B0725F1PMY B073XX9M55 B0174D17D4 B076NCVQNH B007FKF7BA B07DLDT3P5 B07GX4TQVG B0761JVHB2 B077NVNXRS

B01MR0ROVO B072YWZXM8 B00UB503HY B001MSOW6M B00PWPO1OO B075G3H9PY B01KDSMD4Y B00K78A3WA B0079GCK9W B07DLC654K B07H4ZNT59

B06XT59GNB B079QTG9KT B0134XKHCK B01JP9BHLM B07DLC5W2D B0069AICIM B07FXVSLXX B017KQ3BWA B00K78NZV6 B075ZF92S1 B00BPFJUM8

B013UT2BK4 B00PIUSLZ8 B07G8PJ5V8 B00RDK34VW B07GL1WQS4 B00A9C3ZTM B003KV2GTO B01AGOQRL0 B07C6LPX12 B077XP1BPL B07DLF5V4S

B07F24DP3G B075DJXM7W B006WV17EY B06X8ZL7H8 B01B3KQZ22 B01NASAVIT B01KAXZI30 B017KQ37SS B07G8HZFKN B00K787SOG B07CJZKMSR

B078N4392Q B07BKRFYC1 B01KDOSVC6 B00LW9E1R6 B077XJP2Y6 B00A8DD49S B00BPFJQD6 B019YGJYRK B007D4TGF6 B00K787LOS B07GFM13N5

B00BLIDTV2 B00BCYM8ZS B00EKLQA4A B075X1HMB4 B07F8ZLJPX B00BLID5BG B07FSS8ZMG B075VFZRKK B07CLBBHSH B079SM5Z96 B07HZN3CSG

B07DL9R61M B015B5YNDS B01KJSRPXW B01LNANLD8 B01KOUYJDO B0131LSWRM B00ZB2T0L2 B074HWSRBP B07H8PQSJ9 B07F2N8ZM4 B07GGCB3ZF

B00COQKLR0 B00884L67G B07H41WN85 B01LYA4IXD B00ZDA1NJE B00K78OIVW B07HGH5VD3 B00BPFJNHK B07GDT5BR1 B07GL2V6PY

B01KHJP418 B0744ZSSYN B00BCYJMZC B07FB5XRYG B004GYXO9U B073V8Z7R1 B01KN40Y1C B079P7F6XS B07GDR145P B01KDNWG0U

B074JGP3PN B073DDMCC3 B075DJVBXD B01MRV9V1J B00K7865HC B07F2PDFTW B01JMN1QZI B000SSKZT0 B07GTTFCZ3 B0781962P8

B07G5NYDLR B01MZ04K04 B00ML5AGRY B016YGAJJU B078G5G35G B07F3CP43P B00C5O84H0 B07F2R4CPT B00K78KO4W B07H9HZ3BD

B008GR4QHC B07G8SPNYN B00C5O86GO B019L8UDWG B01KDMWMG4 B00K40TH9G B07H13B338 B00C5OUWTS B07DLD5VH5 B07448MTPP

B075R38MXV B07CJY37R1 B0041GJ3YI B003KV5UJM B01KIGRD8C B01KJSLJ1Q B0773FVR8N B00BPFJNEI B07G8QW2HZ B01KOUAM26

B01F7T52LA B071HSCFDD B0759ML77C B00KC9U2SE B07859ND2V B075DHLR5X B07GDVG8C4 B07H3BN7GS B01KEW8UIW B07FYXK4P4

B015STOINC B00BCYJ8K6 B07DPJM99X B01MY2ZBN5 B07G5N6915 B074XHM8GW B07CK3VF7B B01KTYSRQ0 B00BPFJGRC B00XNPH7SM

B076D1M9ZZ B00K78HP2Q B00XZ69V92 B075QPG21G B019S6UQPA B007HBM97C B00C5OMHPK B01K2NLXBE B00K78FBMC B07HFKNQ9D

B073498Y9X B009TBCYIW B01MUBF73X B0042SNZL2 B009AQADWK B07DWDYK8S B07H7VB9YK B079G7GWR7 B07GDT5W67 B01MS4SBP8

B07HXHWRNW B00B286PEY B0066SFUPK B00VTWOW9S B07DV3K6R4 B076KYFB7R B00TECQ0JU B01KIGL1MQ B01KJKURVC B076JGCT6N

B01LXLO5P8 B00CM1IKV6 B00THF1SZ0 B076PVH6FS B004OMIJ0I B01EIIQ79W B07FPVJD36 B07361SGWL B07CHZDX6C B075QP97CK

B01KOTY0MK B001TDITJG B00C5OJC8U B079XXZ1GB B00K786MAW B00K78CBT8 B07G8SBTVH B07GL328K1 B07G7G9FG1 B07GGG32Z9

B06XNVBHCZ B017KQ35L2 B01K85LP0U B00K78KCI0 B01KTWINF2 B01KIG39OY B00O3GU4Z8 B00DKS2KK2 B07G8QR945 B07G9N2D53

B00WZN5HIA B001115D9K B07DVJ4SKR B00IISUN6C B07F1H9J5K B07CHRCRBC B017KQ3EN6 B07HKY6ZBH B00APWNYGU B0762BTSM7

B001VDE984 B008GXOOWS B01IT0OXV4 B078WXJF1J B00COQKM4C B01KJPQ64E B078HMN7YZ B00375DTC6 B01KHKS4JQ B01K88A3MI

B079B9RKGB B00F9SYOW8 B07H4684CV B00P9VMS5U B07BGHKWJF B079J5GP45 B01KDT1RWC B07GHXCXKT B01KIFVBX6 B00K78DU6Q

B075DK2QN4 B013L5597E B00VAUU3GU B074T9T5VL B00BD5I96S B07GHBPK59 B00K78I0E8 B07GH6JPX1 B01N30HUQC B07922R4KQ

B01K9DTCGK B00C5O7XXG B07FVGHY49 B07DQRN58W B07DKYJKPB B07GTWSD74 B002ZUGCB4 B07GDVBD4H B07HBMFJDF B01MYMPYNI

B077NVNXRS B075DH1C9X B077CZWYYG B01KJSY9SQ B01KHIVNPU B079DCML2M B00ZB2Y83C B00BLIFNR0 B00K78AKCS B07GFRKGK9

B0035WSYS0 B07DPHF1H5 B00BCYKTYK B00K78FZQ4 B076PTTPFF B07GCSHHB4 B07D5GZVXM B07GC56FRX B07CL5LZ9M B00KC9UBVC

B075QPW81Z B07BBWQH63 B007GYKB7A B016N56F54 B00AT85HOW B01IRBRF06 B014DQ8KRG