ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

Digital Foot Print Using Python

Bhargava.D.B[1], Gajendra.N[2], Mahesh Babu.S[3], Manjunath.M[4] and K. Dattatreya[5]

UG Scholar, Department of ECE, Adhiyamaan College of Engineering, Hosur, TN, India[1,2,3,4]

Professor, Department of ECE, Adhiyamaan College of Engineering, Hosur, TN, India[5]

ABSTRACT:This project is based on analyzing various social sites and trapping all the information from the social sites. This technique can be used for Twitter, LinkedIn, Glass door and for different social sites.DFP is hosted as a single python process where each component is maintained separately in different sets of files. This project aims to generate a web application to detect social sites using Python and it will generate the positive and negative thoughts based on the user account from the social media. In this, searching process takes place based on their User ID and shows the positive and negative information about that particular person or an organization.

The project is mainly used to identify the digital presence of the organization or individual by performing completely passive and semi passive reconnaissance. Each and every strings becomes a unique object and allows searching without storing the data and purge data for a particular string for a particular duration. Different modules are imported along with it and searching process can be done easily. In this fast moving world people are quickly adapting the computers then popularity of computerized systems is on a high now a days with various firms' organizations. So the craze of computerized systems are going to be increased which is now necessary to make work powerful and faster.

KEYWORDS*:* Beautiful soup, Text blob, Webarticle2text,MySQLTest bench,NLTK.

I.INTRODUCTION

In print publishing era, it was challenging to trace scholars' usage of journal articles in spite of subscription records and library catalogue information. As the digitalization of the academic publishing industry and scholarly communication infrastructures, scholars are increasingly working digitally. These digital workflow encompasses diverse information acts including searching for scholarly information, saving and organizing the information, as well as the utilization of the information. Among these, the first step before the following information organizing, keeping and creating behaviour's is to find the information. Before, no traces could be tracked when scholars read printed articles; however, nowadays digital footprints are created when users visit the webpage of academic articles. By exploring these digital footprints, we have the opportunities to better understand the patterns of scholars' visits to academic articles. Although these data are not generally provided by all publishers, Peer, an Open Access publisher launched in 2012, display the usage data for each of its articles on their webpages (see Figure 1). Are scholarly search engines used more frequently in searches for scholarly articles? How many article links in tweets are clicked on when browsed?

In addition, the detailed information about the traces as opposed to only the counts of the traces are also offered, which makes it feasible to investigate the context of the traces, thus enrich the meaning of the figures. For instance, by looking at news, blogs and tweets content that link back to one article, we could better understand what the impact is to the audience, why the article is impactful, and how the impact could be potentially utilized in the future. Nevertheless, most previous altmetrics studies have been focusing on how to use digital traces of audience' behaviors to interpret the assumed impact; few studies have looked specifically into the actual behaviors and how impact is reflected by them accordingly. For instance, one article would have different impact on people who searched this article on Web of Science and those who clicked on a link to this article on

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

a Reddit post. Although these two traces both appear to be "views", they are different from each other in essence. One of them is likely to be associated with the behavior of intended scholarly information seeking, while the other is more likely to be induced by serendipity luck in social media browsing. By looking at the types and counts of traces, we could barely know the behaviors and information seeking purposes underneath the traces; however, by looking at referral data, which indicate where the audience come from, we could get a better understanding of them.

II.RELATED WORK

In this modern world, where we are almost always on the internet in our waking hours. We are searching for different things, visiting websites, using online services. All this activity can tell many things about us as a person. The keywords in our searches, the websites we visit can tell us what we like, what our interests are and many more things. This raises serious question about privacy, companies like Google and Facebook essentially can carve out our entire personality based on this data - which of course is always recorded - though privacy policies restrict them from sharing the data to a third party.

The keywords we use in our searches can alone tell many things about us. A word cloud is an illustration of the most frequent words from a corpus. The size of a word in a word cloud is determined by its frequency in the corpus. So, in a word cloud generated using our search keywords, the words which appear bigger are those about which we search the most. This is the word cloud generated using my search keywords:

There are many terms related to programming and computer science in word cloud - Algorithm, Database, System, Server, API, names of several programming languages etc. which tell us that I am either studying Computer Science or I work in that field, which is true. Also, Python is the most used word in my searches, clearly Python is my favourite programming language. Windows and Ubuntu appear next, mainly because by primary device is dual booted with these OSs. I frequently search for wallpapers and PDF files for books and research papers, so we see those two words in the word cloud as well.

If we observe carefully, the geographical locations that appear frequently in my searches are India, Uttar Pradesh, Gorakhpur and Kanpur. These are the places where I have been most of my life. IIT Kanpur is the only institute that we see in the figure that is where I studied Computer Science and Engineering.

Flash and Wonder Woman are the only superheroes that appear in the word cloud. Not a very traditional choice (considering Batman, Ironman etc.) but yes, The Flash and Wonder Woman are my favourites. Though, I like Batman very much as well.

The word movie also appears considerably big which tells that I watch movies frequently, which is true. I watch one or two movies every week. If we talk about celebrities Anne Hathaway is the only one who appears in my wordcloud and she is indeed the one I like the most. I don't have a liking for any particular male actor so none appears in the wordcloud.

If we look even closer, Sony and Lenovo are the electronics companies that appear in my wordcloud and that's because my phone is an Xperia and my laptop is a Lenovo Idea pad.

III. EXISTING SYSTEM

The existing method was based on detecting the digital presence of a company or anorganization Digital footprints can be done in many ways depending on the situation. In an online environment a footprint may be stored in an online data base as a "hit". This footprint may track the user IP Address, when it was created. In an offline environment, a footprint may be stored in files using database. This method is mainly used to detect the digital presence of an individual account or an organization website

IV.PROPOSED SYSTEM

Digital footprint or digital shadow refers to one's unique set of traceable digital activities, actions, contributions and communications that are manifested on the Internet or on digital devices.One's unique set of digital

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

activities ,actions and communications that leave a data trace on the internet or other computer devices can be identified and data can be fetched.This method is mainly used to detect the digital presence of a individual account or a organization website and track all the positive and negative comments from the particular website.As persistent attacks on internet are growing exponentially, an organization should be well-aware of the type of information being disclosed on the internet. A sensitive information which is leaked over the internet may be used by the anonymous users to compromise the assets of the organization, cause reputational damage to the organization etc.

ADVANTAGES OF PROPOSED SYSTEM

- Fraud or legal issues can more easily be detected
- Personalization (like suggested products or related advertising) serves to add value to our use of the Internet
- Companies can more easily offer incentives based on interests and needs, sometimes resulting in cost savings
- This technique is more useful in detecting the digital presence of an individual or an organization
- Efficiency of digital presence is very high.

V.RESULT

Specifically, the first place is the number of unique visitors from bookmark or typed URL, which is about 1.44 million. The number of visitors from Google Search is about 837 thousand, being ranked the second. Other top links are found to come from news reports and blogs (including Io9, Science Daily, and the Annals of Improbable Research, etc.), general search engines (Bing, Yahoo Search and Wikipedia, etc.), academic search engines (National Center for Biotechnolog Information, Web of Science, and DOI.ORG, etc.), as well as social network (Facebook, Twitter and Reddit, etc.).

All the referrals listed in Table 1 were harvested from peerj.com directly and are presented as their original forms. Some referral sources may have different URLs, for example, there are two URLs for Facebook, which are m.facebook.com and www.facebook.com. Some websites have mobile version with the same contents. i.e., m.huffpost.com and huffingtonpost.com. Most referrals in Table 1 are from the United States. Besides those Services associated with Google, others include general and scholarly search engines, social networking services, news outlets, and blogs. Yahoo search and Bing are the two search engines following Google. Scholarly search engines include NCBI (National Centre for Biotechnology Information) and Web of Science. Housing a series of databases relevant to biotechnology and biomedicine, NCBI focuses on similar disciplines to peer's publication scope.

VI.CONCLUSION

The subject for this thesis came from Clue tail Ltd. The company needed a research to be done about sentiment analysis; more precisely, about the possibilities in document level opinion-mining and how such system could be implemented in Python. Studying the subject led to an understanding about the problem and the available solutions. It also helped to understand how sentiment analysis is a high-level application, coming from an actively studied research field that incorporates techniques from data mining, machine learning and natural language processing. Finding the document-level opinion is often done using discovered linguistic features. This can be achieved by using existing or manually assembled corpora, which in the latter case can be a time consuming process. The suggested solution comes from combining supervised and unsupervised learning methods. Compared to its counterpart, unsupervised learning may not have as many learning techniques available; however, using it can lead to good results with less work. That can be a key factor if the research resources are limited. The opinion mining process consists of two key elements. The first one is choosing the set of features that represent the opinion information. The second one is the document classification based on the information discovered from the chosen features. Classification can be done by using machine learning methods, for example, Bayes networks, Neural Networks or Support Vector Machines.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

REFERENCES

1. Bornmann, L. (2014). Do altimetry's point to the broader impact of research? An overview of benefits and disadvantages of altimetries. Journal of Informatics, 8(4), 895-903. doi:10.1016/j.joi.2014.09.005
2.Cheung, M. K. (2013). Altimetries: Too soon for use in assessment. Nature, 494(7436), 176176. Christozov, D., & Toleva-Stoimenova, S. (2013). Knowledge diffusion via social networks: The 21st century challenge. International Journal of Digital Literacy and Digital Competence (IJDLDC), 4(2), 1-12.
3.Costas, R., Zahedi, Z., & Wouters, P. (2014). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology. Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: