Search Engine Results Comparison for Result Filtering

Frank o'Neill

December 21, 2016

MSc Data Analytics
School of computing National College of Ireland

Supervisor: Michael Bradford

Contents

List of Figures

List of Tables

Abstract

Search Engines are a highly complex mixture of technologies and business motives. Behind this complexity incentives to engineer the results for business motives or otherwise are sometimes made. One claim is that Google results are directed towards user search history or identity. Search results bubble towards user preferences. While the claim of another search engine DuckDuckGo is that user identity is not used to engineer search results. In this report two search engine results are compared DuckDuckGo and Google in a two way Google Signed In / Out configuration.
A new Google account is opened and a profile is built for this Signed In account with three specific search terms over a six week period. An automated web browser is then used to gather data for three configurations of - Google Signed In, Google Signed Out and Duckduckgo. The first thirty URL returns of each Search Term for each Search Engine configuration are then categorised using the Web Service Similarsites.com. The returned categories are then evaluated for filtering. Findings suggest Google Search has a difference between a Signed In and Signed Out Search. Divergence is stronger depending on the particular topic - a filter bubble indication
. The exercise has the limitation of reliance on similarsites.com integrity, a small time frame for analysis.

1 Introduction

1.1 Engine as medium

Search Engines are now an everyday part of human communications and activity with contemporary culture and consumption mediated to a large extent through the internet. Questions of fairness, inclusivity and representation among others can arise just as inother media outlets. While other major industries with high impact and social responsibility have a degree of regulation, search engines are so far are unregulated. Telecommunications Act in the US limits the reach of television companies to 35 percent. The German State Treaty for Broadcasting supposes that a market-dominant position exists when its 30 or 35 percent. (Machill et al., 2008) Perhaps one of the most critical aspects of search engines is its use in journalism. Claims of independence for journalism or even social media discourse may not be as neutral as it appears.Search engine testing then is in the public interest from an ethical and business point of view. Therecommendation by the American Federal Trade Commission thatsearch engine companies disclose paid link policies and preferred placement schemes was significant for Internet users.(Rogers, 2014) A counter argument is that of the Google Dilemma a search engine that does not select and rank would be useless. There is no search without bias.(Diaz, 2005)

1.2 Overview

This paper reports on the findings of a comparison between two Search engines. The Google engine in two modes (signed in and signed out) and the DuckDuckgo search engine. The DuckDuckgo search engine claim is that previous searches are not recorded or stored. On initial comprehension this implies an independent and unbiased search for each search event. Again independence from market and commercial forces might be the perception of the DuckDuckgo engine. The attempt of this exercise is to examinethe actuality of a difference between the search engines using the concrete search results, and determine a possible measure ofdifference and its interpretation. Its a comparative assessmentwithout reference to absolutes or standard metrics. The comparison has a bases on categories returned per search termonly. Again a more comprehensive comparison might require interpretation of website contents or other complex dimensions. But this would spread the focus to internal search engine complexities which are difficult to comprehend and make for a less manageable analysis. A major claim of the DuckDuckGo search engine is that it does not record previous search history. Search data

"is arguably the most personal data people are entering into anything. You're typing in your problems, yourdesires. It's not the same as things you post publicly on a social network."(Arthur, 2013) While DuckDuckGo and Google engines are compared, results for Google Signed In and Google Signed Out are also compared and evaluated. This Google comparison is to examine the claim of a filter bubble. Google Signed Out may also imply a privacy element as in DuckDuckGo. Definition: Filter Bubble internet search engines and their algorithms are creating a situation where users increasingly aregetting information that confirms their prior beliefs.(Holone, 2016)

Another Definition of filter bubble: "personalized" Internet filters whose very purpose is to narrow,rather than expand, the world that we see. (Albanese, 2013)

The paper introduces some of the important elements of Search Engines and some background information on features used in the project. Literature Review examines connected projects and theirapproach. An approach that seems to capture the essence and comprehension of search engines is presented the business model. This is included as it greatly aids in interpretation for the results derived in this exercise, and captures a bigger picture. This is followed by methodology, implementation, results and discussion.

1.3 Research Questions

1) Search results are a filter bubble. What is the evidence?
2) Can search engine analysis be simplified?

1.4 Background

1.4.1 Engine types

Engines consist of three broad types. Crawler - based, Human edited Directories and Meta- Search. Crawler based engine has three fold components of Crawler, Indexer and engine software. At the basic level Crawler scans web pages including links and indexes them for quick reference by search. This takes place on a scheduled basis to keep index in date. The software interprets the query, searches the index and returns the result that is subject software engineering. Directory type - search engines listings are created manually. Websites are submitted for inclusion and then reviewed by human mediators for allocation or indexing to a search base. Meta search - engines integrate search term results from multiple search engines. A Meta engine has its own algorithms that filter and rank the third party search engine results.(Zhang, 2016)

1.4.2 Similarsites

Websites are commonly correlated for security, market and competitor research, risk management, compliance, filtering and ethical reasons. The correlation may be based on characteristics such as similar structure, same server or IP address, owner, content, category. These databases of correlation can be publicly accessed and one of these is the category database similarsites.com. Category database entries may be through manual judgment Knowledge Engineering (KE) manually defining a set of rules how to classify documents under given categories. (Sebastiani, 2002) A weakness in this method is its overt subjectivity. In recent years categorisation has become a machine learning process, but again not totally immune from subjectivity which have latent bias in algorithm design or choices. Text classification in itself is a major specialised area that can employ a suite of different machine learning algorithms. SVMs, neural networks, probabilistic classifiers etc to name but a few. Text categorization is the task of assigning a Boolean value to each pair dj, ci DC, where D is a domain of documents and C = c1, . . . , c—C— is a set of predefined categories. (Sebastiani, 2002) Categorisation then is not an exact science but nevertheless can be measurement of effectiveness as opposed to efficiency which is done through the standard metrics of precision and recall. Categorisation can allow for the target audience to be more focused and within a context of meaning. This is particularly the case in a sematic search engine but this is an ongoing developments in keyword engines also. Categories and put software in the position of being able to access the contents of the web automatically and sensibly on a content-related basis. (Machill et al., 2008) Other uses of categorisation include content classification, security, malicious URL detection, filtering, screening.

1.4.3 Terms

With the new account, a Google Signed In profile was built around three terms. Three terms were used to make the project a manageable exercise. Suit,tooth implant and abortion.The last term abortion was choosen due to its high trending or controversy in contempoary debate. Tooth implant was chosen due its high costs and anecdotal evidence of competion for the business. While suit is a more mundane item notwithstanding the high profile nature of the online clothing market.

2 Literature Review

Search engine analysis given its position in contemporary societal structures has been approachedfrom many different angles. The very nature of evolving Search Engine dynamics can make an analysis have a short time span of validity. Much of theliterature becomes very technical involvingtime series, meta search, transaction logs, word frequency, natural language analysis and more. Again with analysis from a business view point, advertising, e-document, value chain, content, social analysis to mention but a few. It encompasses a large number body of research with many different angles and approaches. The different methods are outlined below gives a picture of the field thats in play. 1. Relevance, 2. Ranking, 3. User satisfaction, 4. Size/coverageof the Web, 5. Dynamics ofSearch results, 6. Few relevant/knownitem, 7. Specific Topic/Domain, 8. Automatic. (BibAli and Beg, 2011) However there are some standard attempts at search Engines analysis. Indexing evaluation was one of the first performance tests carried out through the Cranfield experiments.(Cranfield, 2012) InformationRetrieval (IR) is also monitored through U.S Government departments. National Institute of Standards and Department of Defense co-sponsor test workshops through the Text Retrieval Conference(TREC).(Demeester, 2014) Many evaluations have attempted a quantitative analysis even though an objective metric has been difficult to define. A standard Search engine comparison is made with the quantitative metrics such as Precision and Recall. (Mowshowitz and Kawaguchi, 2005) Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved. (Voorhees and Harman, 2001) While stability an another metric which is an aggregation of various characteristics such as number of pages retrieved, results of top 20 ranking order etc(Vaughan, 2013) Then other measures among others, such as coverage or pages indexed and response time.

2.1 Engine Bias

A dominant theme in search engine evaluation is bias in search results which can take many forms and technical manipulations of metatags, hyperlinks, pointer text, freshness, argument exclusion, unbalanced returns etc. Its a large area of research and many evaluations have been carried out under legal or anti-trust obligation. But a high proportion of these tests are carried out by high cost human subjects.(Can et al., 2001). A typical scenario is, Twenty-four subjects ranked four sets of Web pages and their rankings were used as benchmarks against which to compare search engine performance.(Vaughan, 2013) Google has had several cases of bias and monopo-

lisation charged against it, with its two tier system of organic and paid results, the claims are that Google does not deliver the results that would best serve consumers, but instead alters those results to serve its own competitive interests.(Patterson, 2013) Hchsttter and Lewandowski (2009) conducted a comparison on thestructure of different search engine results pages. There findings showed or verified that results did tend to favour offerings from their own respective business units Google - YouTube, Yahoo -Yahoo Answers etc There methodology was to compare five search engines under headings of Position of organic results, Absolute Position whether organic or paid, AdWord position, Trigger term for Ask.com result. Their emphases was in finding what elements occur most in search results, paid,organic, links or ads. Some of the same problems occurred in their study as in this project i.e. a limit on amount of returns obtained per term. (Hochstotter and Lewandowski, 2009) Another approach retrieved URLs and analysed for bias in the results. It was a comparative analysis against other engines with a cluster analysis of the types of URLs returned. Similarity testing of URLs was a feature but was testing for measurements of bias against recall, precision. The URL distribution were then measured against an ideal distribution for that query. The evaluation then had a judgement process of bias occurrence, rather than having a view of the actual returns unlike the exercise here. (Mowshowitz and Kawaguchi, 2005) A group of librarians were used in one case of a search engine analysis. 16 librarians were asked to access their experience inlooking their intended material. Librarians are probably the most skilled to determine quality of returns and their analysis.But the assessment was confined to Greek librarians and again it was a subjective manual exercise. (Garoufallou, 2012)

Search evaluations have been undertaken on the quality of information retrieved by search engines. Medical experts conducted quality of returns for a weight-loss search. Again quality was judged by subjective medical personal and a time consuming manual process. There findings were that returns were of substandard quality, but there was no engine comparison of concrete facts.(Modave et al., 2001)

A similar evaluation was conducted on search engines on how information on prescription drugs are accessed. Search terms were scientifically chosen by professionals and the results showed that Wikipedia was the most prominent retuned page. The evaluation was on the web sites returned but again the interest was on the page rank.(Law et al., 2011)

Another exercise carried out five performance evaluations on five engines. The evaluations had an element of rigour in that queries were selected from a library of information science. Complex queries were used and the experiment was conducted over two time frames. The comparison was again on

retrieval, coverage, relevance, stability. Both Excel and Avova were used for analysis but emphases was on first page results. Allocation of returned engine categories and proportions was not performed.(Sanjib et al., 2009) Automatic evaluation has also been attempted, automatic Web search engine evaluation method (AWSEEM) (Can et al. 2004). The evaluation is based on the first200 hundred pages returned. But returns are judged for relevancy in an obscure method for inclusion in the count. This entails entering the complexities of the page using vector space and idf models.(Can et al., 2001) Many other comparisons of search engine results use transaction logs and search engine optimisation (SEO) comparisons. There are some comparative analysis, between search engines but they tend to mostly focus on technical details such as precision, relevancy, response time and other parameters.(Edosomwan and Taiwo, 2010)

2.2 Google Page Rank

A core element of the Google search engine is PageRank which involves a mixture of matrix theory, numerical analysis, Graph Theory, a random web surfer model, probability to give a PageRank Score. The manipulation of these parameters and others such as page links, word frequency, Damping factor, Personalisation factor, Rank vector, Order Node etc. is not divulged by Google.(Wills, 2013)

There have been numerous studies or evaluations on different Search Engines of Page Rank positions and what factors influence getting to the top position. This is mostly driven from a business or marketing perspective. Bar-Ilan, Mat-Hassan and Levene (2006) carried an evaluation to find how in practice are ranking algorithms with the attempt to derive a measurable dimension or quantify changes over time. (Bar et al., 2001) Metrics used were the standard over-lap measure between search engine results, and Spearmans footrule. An extension of spearmans rule a G measure to help overcome a problem encountered in comparing ranking when two search engines return different or non-identical document sets. This extension seems overly complicated, and a fourth M measure involving intuition is involved. This measure tries to capture the intuition that identical or near identical rankings among the top documents. (Bar-Ilan et al. 2006) As can be seen attempting to achieve accurate quantifiable metrics of Search Engine rankings is not an accurate or well defined science.

2.3 DuckDuckGo

A keyword search engine such as Google it is claimed does not take into account the meaning of the search term and expression used in the web page. While Google uses its Rank Algorithm to give its results. Semantic Search uses the science of meaning.(Singh, 2013) to predict user search term intent. Semantic search attempt to interpret what the user intends to know by the search term. (Singh, 2013) carried out a comparison between a Keyword search engine ssuch as Google and Yahoo and a semantic based engine like DuckDuckGo, Hakia, Bing. The comparison was made on precision ratio and natural language processing between the keyword and semantic search engines. Semantic search engine scored highest in precision ratio. Initially, two keyword based search engines (Google and Yahoo) and three semantic search engines (Hakia, DuckDuckGo and Bing) are selected to compare their search performance on the basis of precision ratio and how they handle natural language queries. Ten queries, from various topics was run on each search engine, the first twenty documents on each retrieval output was classified as being relevant or non-relevant. Afterwards, precision ratios were calculated forthe first 20 document retrieved to evaluate performance of these search engines. Again this comparison uses ambiguous judgements of what is relevant non-relevant.(Singh, 2013)

2.4 Business Model

One approach to Search Engine analysis is from an economic motivation point of view identify economic forces ..as a robust critique of the current situation. (Rieder and Sire, 2013) The thesis is that the Google is built on a three-sided market. (Rieder and Sire, 2013) business model. The Google threefold structure is one of an exchange around users , content providers and advertisers. The results page is the visible outcome of a dynamic procedure of query-results-ads matching (QRAM). (Rieder and Sire, 2013) TheQRAM itself is a field of study in itself, a complex interaction of various technologies and expertise which shift business relations to an engineering focus or problem. But the end objective is to maximise actor gains. In a sense, the company is the technology-focused shark in a pond of content-focused fish. (Rieder and Sire, 2013). Google then attempts to satisfy or relies on all three sides of the market of user queries, content providers and advertisers. It's a balancing act to engage all actors but with one subsidising the other two. users search for free, content providers get indexed for free.advertisers pay.(Rieder and Sire, 2013) It has multiple media services such as You Tube, Google maps, Google books etc.along with operating systems, cloud service,

social network, mobile applications email. A hardware division with products such as Google glass, Nexus, Chromebook, OnHub, etc so it has many reasons and incentives totrack its users, their economic resource. Advertisement is at the heart of its existence - advertising indeed accounts for virtually all revenues collected by Google, the dominant actor.(Rieder and Sire, 2013)

DuckDuckGo is also a free service like Google and needs to generate revenue , which it does through advertising and affiliate revenue. It does this by the search term that is entered rather than from a profile of the users previous search history. There is an association with Yahoo and Microsoft as part of the Yahoo-Microsoft search alliance and DuckDuckGo is part or one of the Bing distribution channels. It is also an affiliate of Amazon and eBay and if these are accessed through DuckDuckGo with a subsequent purchase then commission is paid.(Duckduckgo, 2013) Ads can occur in the search results which is provided by a third party 'Microsoft's Ad Center' or Bing Ads. Bing Ads are similar to Google the Google AdWords model where revenue is based on a relationship of an advertiser's willingness to pay a certain amount an the CTR(Click through rate).(Online, 2016). Both the Bing ads service and the affiliate programs for Amazon and eBay are open to the general public for integration into their own sites. To advertise on Duckduckgo a customer needs to be signed up to a Bing Ads account. The Ad will now appear on Duckduckgo search.(Duckduckgo, 2013) The claim is that there is no identification involved and a previous purchase will not influence subsequent ranking of search returns. if your browsing history shows you visit high-end sites, some sites will increase prices. (That's why plane fares can drop if you delete the "cookie" files in your browser.)(Arthur, 2013) Duckduckgo differentiates itself on its privacy policy. Monetising its business in the agressive approach or other players would mitagate its unique sector. To date it relies on 3rd party sources for its reveue.

3 Methodology

3.1 Concept

Rather than presuppose an hypothesis, latent design bias or approach from a refined theoretical angle, the end results per search term are derived. As seen from the Literature Review some previous search engine analysis took a qualitative approach. Human subjects were used for judgments, a subjective process on ambiguous metrics such as relevancy, representativeness, bias and others. While other analysis took a more quantitative and automated approach deriving metrics such as precision, recall, rank position, Ad position

among others. Subjectivity of human judgements are bypassed, and to an extent internal engine complexity as its end result of search engine algorithms that are evaluated.

3.2 Design

The exercise relies on the web service of Similarsites.com for categorisation of URLs. There are other such services and may show a variation of category allocation from Similarsites.com. Google search is through Selenium Webdriver which allows for programmatic browser drive. All three search configurations of Google Signed In - Google Signed Out and Duckduckgo are carried out programmatically. Results for a search term are then compared graphically on a single display. Both Jaccard distance and Cosine similarity are calculated and also displayed. A DataFrame of Jaccard Distance and Cosine are written to a csv file for each search. This will allow for construction or build of an historic profile of a search term and further analysis. A test by a search term may be an input from keyboard or from file depending on which configuration of a batch file is selected.

3.3 Portability

The test may be executed independent of location and machine. Variation of results with location or machine may then accessed. The programs may be uploaded to different machines with Linux O.S . The Google Signed In results may have a bias depending on user profile history. (For these tests a new account was opened and a search profile built with selected search terms. This profile build was automated and executed on a daily basis). Again portability allows testing and assessment in different time frames.

3.4 Output

A first evaluation is made through pie charts which will give an immediate indication of any variation in categories per search term between search engine results. A second evaluation is then made by finding the Jaccard distance and cosine distance between returned categories for the same search term per search engine. Taken together these two evaluations can reveal trends.

4 Implementation

Six programmes in Python and Bash shell script automate the testing and analysis, with a further file the gathers data for further analysis that will enables a historic analysis perspective. A new Google account is opened and the selected terms are searched through this account on a daily basis in order to build an account profile. The profile building is an automated process which is activated once or twice daily. The terms are shown below in Table 1.

Suit
Tooth
Abortion

Tabel 1

Different variations of terms are used around the same topic. An example might be Suit Blue suit, Light suit, Tooth Implant, Pro-choice, Pro-Life, Best used car. This profile building is done from home location. A different machine in a different location is now used with the same search terms. (This machine is never powered up in the home location). A search is also carried out from a home location on a the same machine that was used for profiling. The setup is shown in below in Table 2

Home Machine	Other Machine
Build profile	Never power at home
Run Tests	Run Tests

Table 2

Both simple and compound terms are used in search e.g suit, blue suit.. Search terms are manually entered into a text file - searchTerms.txt - these terms are common input for all three engine configurations. Google search automation is facilitated through the Selenium webdriver, for both Google

Signed Out and In. (With search through Selenium there is an assumption that its independent of bias from either Google or Duckduckgo). While the Duckduckgo search is through wget package. The test setup is shown below in Table 3.

Search config	Returns
Google Signed Out	First 30
Google Signed In	First 30
Duckduckgo	First 30

Table 3

4.1 Dataset

Selenium is programmed to open sequentially on the first three page results from a search term with each page returning 10 page rankings. This gives a total of the first 30 returns per search. Likewise the Duckduckgo search returns the first 30 search results per page. Duckduckgo returns are in JSON format with the URL extracted through pharsing while the URL for selenium returns is derived from the page source elements and then pharsed.

4.2 Software

Both Selenium configurations of Goggle Signed Out and Google Signed In and search Ducduckgo are programmed to output the search term used and the resulting URL. The three resulting outputs require processing before URLs are sent to Similarsites.com for categorisation. The Selenium output is processed through ProcesSel.py to remove a line due to a trailing line left by the firefox Browser. While the Duckduckgo results output are in Json notation and require pharsing. The three result sets for each of the configurations are now processed through the simSite.sh program, which returns the categories of the URLs. The final outputs returned are the original search terms, their returned URLs and their (URL) categories. The analysis of the categories is

the processed through the python Pandas package. Using the Pandas package greatly reduces the manual time consuming cleaning and filtering of the raw data. The clean data can now be used to form a DataFrame, written to file and storage of the result sets for historic trends and future analysis.

4.3 Output

Output is a pie chart visualisation of all three configurations per search term. Together with the pie charts are the Jaccard distance and Cosine similarity of the retuned categories per search configuration of Google Signed Out/In and Duckduckgo. A typicaloutput is shown below in Figure 1

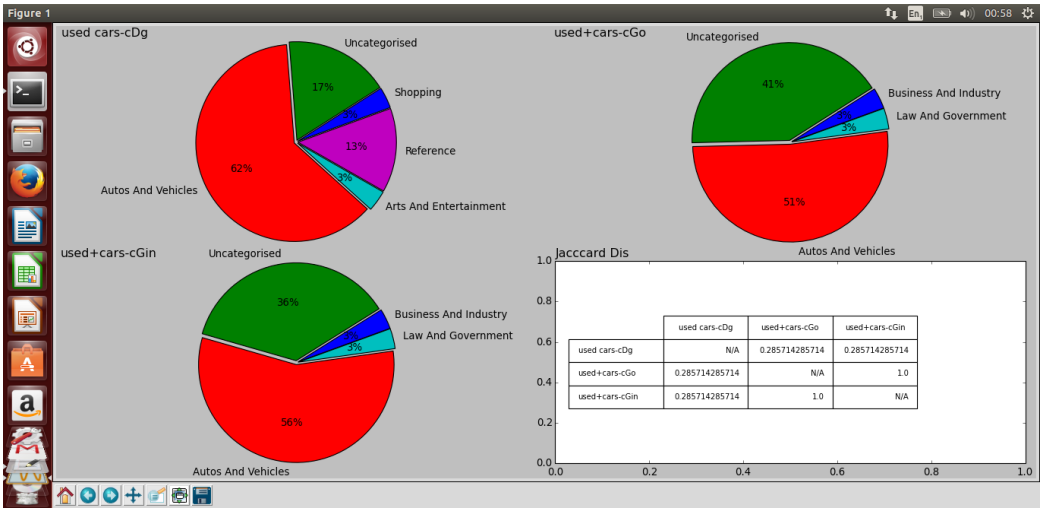

Figure 1: Tests output

4.4 Visuals

A visual comparative of all three configurations Google Signed Out, Google Signed In and Duckduckgo is the first output from a search term. Along with this output there is the Jaccard Distance and the Cosine similarity. These

comparison are for the immediate time frame and give an indication of the current status. Figure 2 below shows the output.

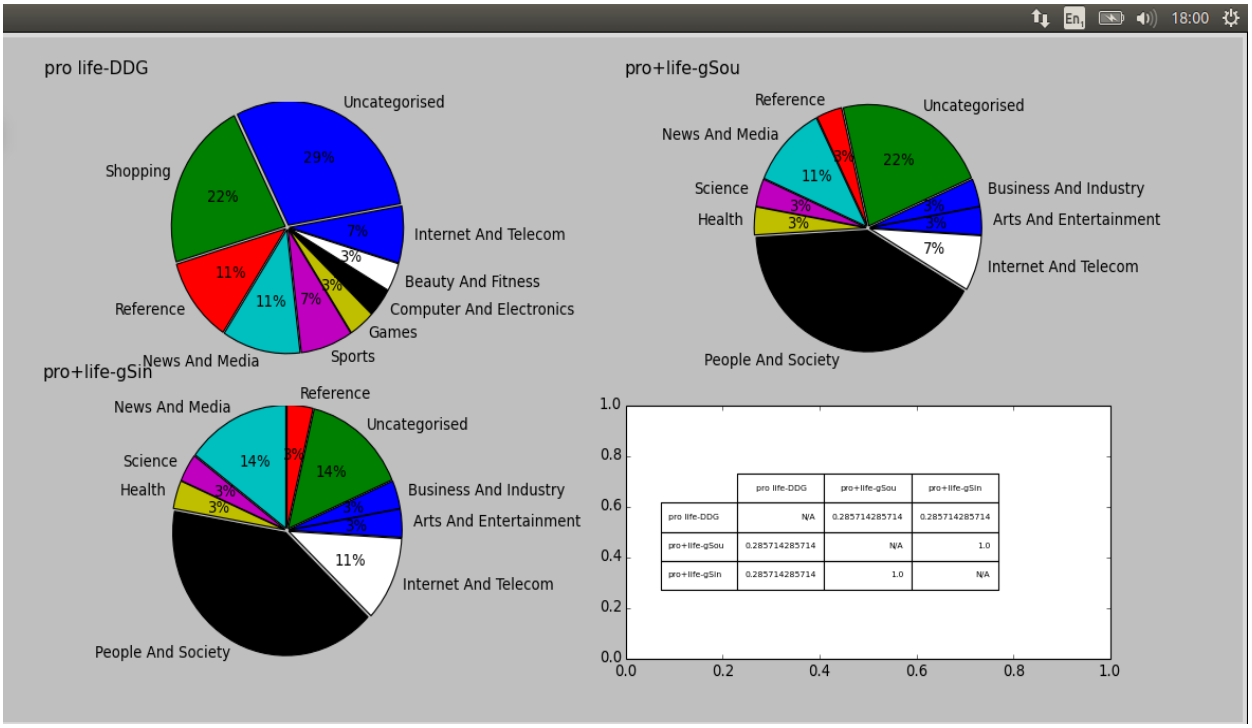

Figure 2: Output for Term Abortion arguments

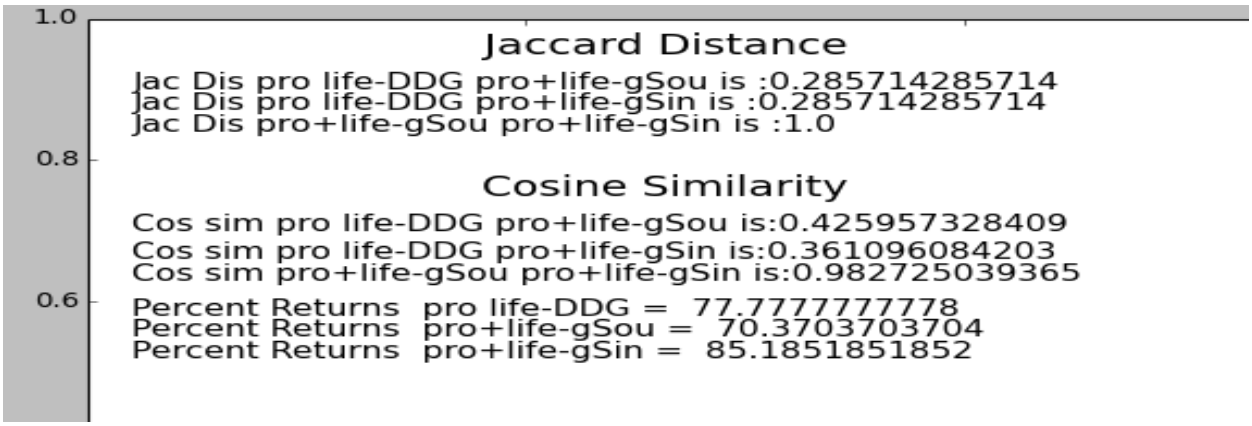

Figure 3: Jaccard and Cosine Output. Term pro-life

4.4.1 Abortion argument visuals

profile on this topic using variations of search terms around a central theme , has previously been built for approximately 8 weeks from a home location. This is under the Google signed in account of 'Tony Byrne'. Search results for these two terms from a home location on different dates are collected. Along side this a search using the same terms with a different machine (M2) and location are collected. (Machine M2 is never powered up apart from test location). The City center Central Libary is the location used here in these tests. The results for three different dates are shown below in Figure 4 and Figure 5 for the selected search terms 'pro-life' and 'pro choice'.
Both 'pro life' and 'pro choice' search terms show a deviation of results when the search is changed to a different location. A search is made from a Home location and a City center location on the same date. The results are shown in Figues 4 and 5 respectively. The difference is calculated in corresponding Table 4 and Table5.
The Table metrics consist of Difference = Home - Library, SAD= sum of absolute difference,SSD = sum of squared differences,Corelation=corealation coefficient.

Figure 4 and 5 Legend. H=Home, L=Library,M2=Machine 2(Different machine than Home machine)

	A	B	C	D	E	F	G
1	Abortion topic	Configuration	H- Nov 28	H- Dec 5	H- Dec 15	L-M2 -Dec 15	
2	jacrd_ab	pro life-catDdg catGin	0.23	0.23	0.29	0.6	
3	jacrd_ac	pro life-catDdg catGout	0.25	0.23	0.29	0.55	
4	jacrd_bc	pro+life-catGin catGout	0.86	0.75	1	0.9	
5	cosin_ab	pro life-catDdg catGin	0.35	0.46	0.38	0.96	
6	cosin_ac	pro life-catDdg catGout	0.35	0.26	0.36	0.94	
7	cosin_bc	pro+life-catGin catGout	0.97	0.93	0.99	1	

Figure 4: Abortion Topic. Term: pro life

Term: pro life	Home v Library Location - Dec15-2016
Difference	[-0.31, -0.26 ,0.1, -0.58, -0.58, -0.01]
SAD	1.84
SSD	0.8466
Corelation	0.576

Table 4

The comparision for Home location v Library location for the term 'pro life' is shown in Table 4 above. The corealation for the term is weak for the different locations. (Note that - Different location also uses a different machine.)

	A	B	C	D	E	F
1	Abortion topic	Configuration	H-Nov28	H -Dec 5	H- Dec15	L- m2-Dec15
2	jacrd_ab	pro choice-catDdg catGin	0.69	0.31	0.31	0.67
3	jacrd_ac	pro choice-catDdg catGout	0.69	0.31	0.38	0.67
4	jacrd_bc	pro+choice-catGin catGout	1	1	0.89	1
5	cosin_ab	pro choice-catDdg catGin	0.77	0.7	0.66	0.85
6	cosin_ac	pro choice-catDdg catGout	0.68	0.67	0.72	0.84
7	cosin_bc	pro+choice-catGin catGout	0.96	0.99	0.99	1
8						

Figure 5: Abortion Topic. Term: pro choice

Term: pro choice	Home v Library Location Dec 15-2016
Difference	[-0.36, -0.29, -0.11, -0.19, -0.12, -0.01]
SAD	1.08
SSD	0.2764
Corelation	0.9839

Table 5

Comparision for Home location v Library location for term 'pro choice' is shown above in Table 5.
Comparing the results for each term of 'pro life' and 'pro choice' . The 'pro choice' term for the Library Location has a closer corelation to the Home Location . A corelation of 0.9839 as compared to 0.576 for the pro life term. Taking Google Signed In/out - for both locations and machine the Cosine Similarity shows no difference while the Jaccard Distance shows a 0.1 difference for Home to Library. But most of the difference comes from the Duckduckgo variation- a 0.58 variation for Cosine Similarity between Google Signed In and Duckduckgo. This weaker corelation between Home location and Library Location would support the claim of Duckduckgo that it dose not track the user.

4.4.3 Suit visuals

Figure 6 below shows the visuals for term 'linen suit' while figure 7 shows the Jaccard Distance and Cosine similarity. Both figures are for a recent serach on December 19th 2016.

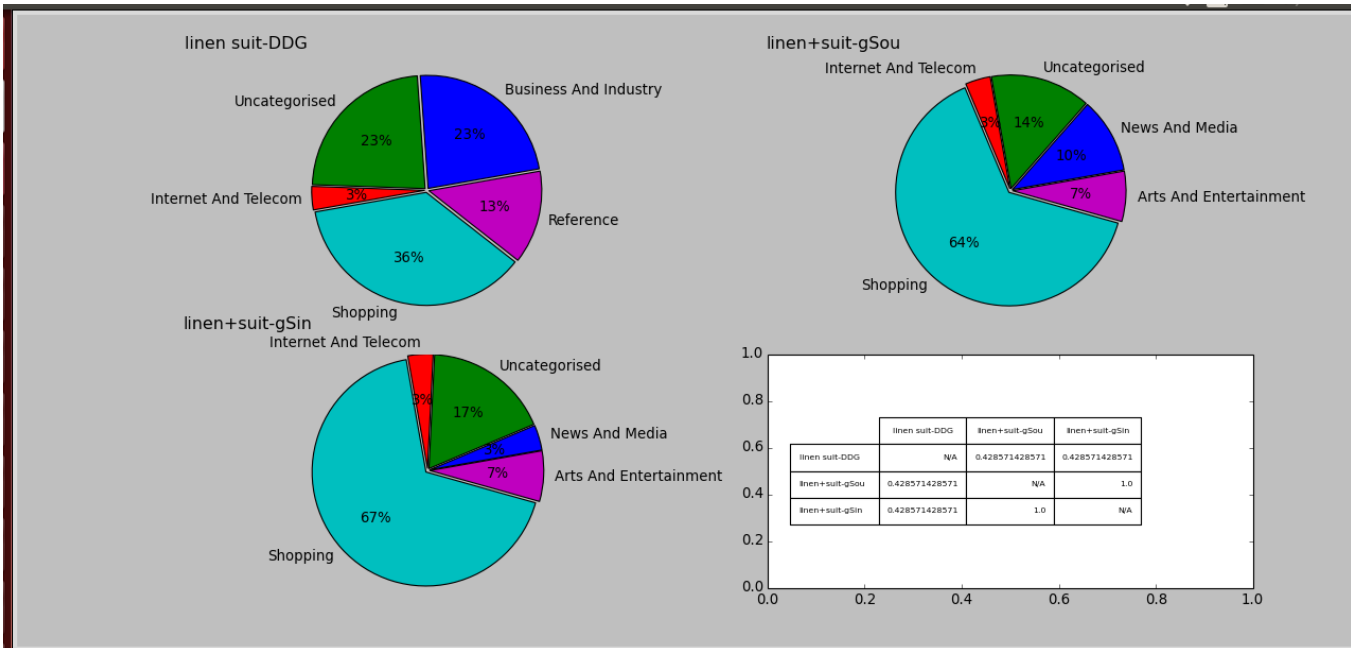

Figure 6: Suit Topic.

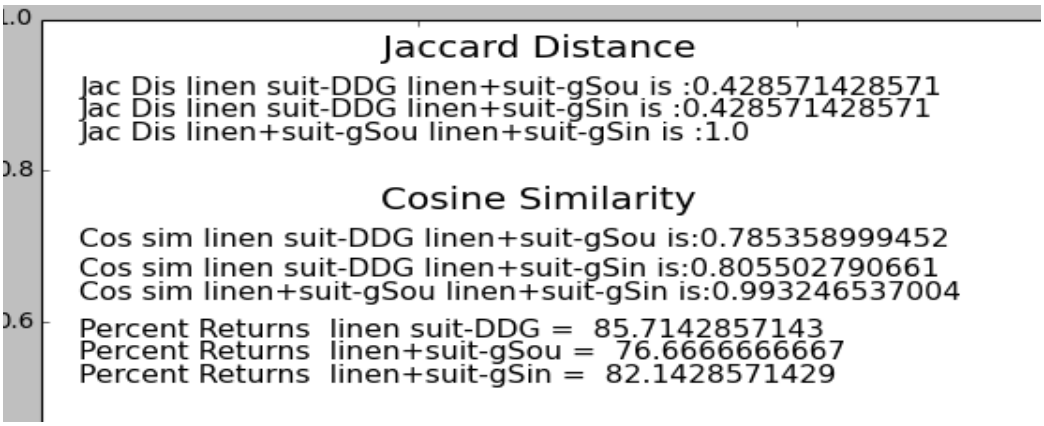

Figure 7: Suit Topic. Term:l linen suit

4.4.4 Suit Topic

Term selected for evaluation on this topic 'linen cloth' shows little variation on the time line of this test, as can be seen from Figure 8 below. Different machine and location or dates gives very little deviation of results.

	A	B	C	D	E	F	G
1	Suit topic	Configuration	H-Nov 28	H-Dec 5	H-Dec 15	Lib -m2 Dec15	
2	jacrd_ab	linen+cloth-resCatGout resCatGin	1	1	1	1	
3	jacrd_ac	linen+cloth-resCatGout choice-resCatDDG	0.64	0.64	0.71	0.71	
4	jacrd_bc	linen+cloth-resCatGin choice-resCatDDG	0.64	0.64	0.71	0.71	
5	cosin_ab	linen+cloth-resCatGout resCatGin	1	1	0.99	1	
6	cosin_ac	linen+cloth-resCatGout choice-resCatDDG	0.78	0.79	0.73	0.71	
7	cosin_bc	linen+cloth-resCatGin choice-resCatDDG	0.77	0.78	0.74	0.71	

Figure 8: Suit topic Term: linen cloth

Term: Linen cloth	Home v Library
Difference	[0, 0, 0, -0.01, 0.02, 0.03]
SAD	0.06
SSD	0.0014
Corelation	0.9963

Table 6

Term 'linen cloth' shows a strong corelation of resullts between Home and Library locations as shown in Table 6 above. Comparing different terms around the same topic might yield more meaningfull results.

Three different terms of 'Linen suit', 'Light suit, Linen cloth' show an increased divergence among terms. This divergence between Google search and Duckduckgo rather than between Google Signed Out/In. Jaccard distance for Google Signed Out/In remains the same among terms, while the Cosine

similarity moves by 0.02. The variation among terms is shown in Figure 9 below.

Topic suit	Configuration	Linen suit	Light suit	Linen cloth
jacrd_ab	catDdg catGin	0.43	0.46	0.64
jacrd_ac	catDdg catGout	0.43	0.46	0.64
jacrd_bc	catGin catGout	1	1	1
cosin_ab	catDdg catGin	0.79	0.64	0.78
cosin_ac	catDdg catGout	0.81	0.6	0.79
cosin_bc	catGin catGout	0.98	0.98	1

Figure 9: Suit topic terms

A comparision of terms ‘mens suit’ and ‘womans suit’ shows a greater deviation in results. The Jaccard distance here for Google Signed In/out between the two terms shows a difference of 0.5. The term ‘womans suit’ shows no variation of Jaccard Distance between Google Signed Out/In while the term ‘mens suit’ shows a Jaccard variation of 0.5 for the same configuration of Google Signed Out/In. These results were from a home machine and location where a profile was earlier established. The results are shown below in Figure 10.

Term:Suit	Mens suit v Womans suit
Difference	[-0.5, 0.25, 0.15, -0.02, 0.03, 0.0]
SAD	0.95
SSD	0.3363
Corelation	0.7198

	A	B	C	D	E
1	Topic suit	Configuration	Mens suit	Womans suit	
2	jacrd_ab	suit-goGi goGou	0.5	1	
3	jacrd_ac	suit-goGi DDG	0.5	0.25	
4	jacrd_bc	suit-goGou DDG	0.4	0.25	
5	cosin_ab	suit-goGi goGou	0.98	1	
6	cosin_ac	suit-goGi DDG	0.97	0.94	
7	cosin_bc	suit-goGou DDG	0.94	0.94	

Figure 10: Suit topic. Mens suit v Womans suit

Table 7

Table 7 above shows a deviation between the two terms. These results were both gathered in a Home location. The profile that was built for the suit search term were biased around clothing, such as 'light suit', 'travel suit', 'cotton suit'.

4.4.5 Tooth visuals

Figure 10 below shows the visuals for term 'tooth implant' while figure 11 shows the Jaccard Distance and Cosine similarity. Both figures are for a recent search on December 19th 2016.

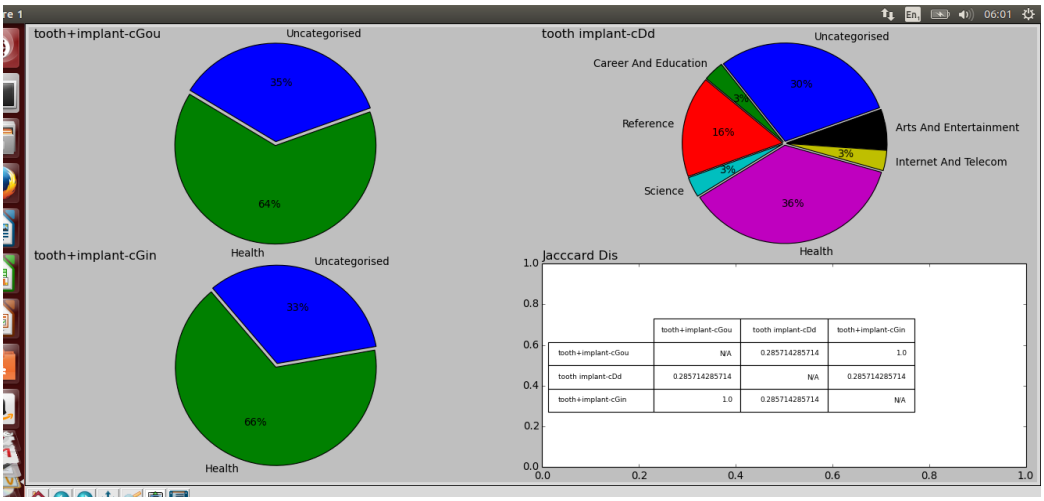

Figure 11: Tooth Topic: Tooth Implant

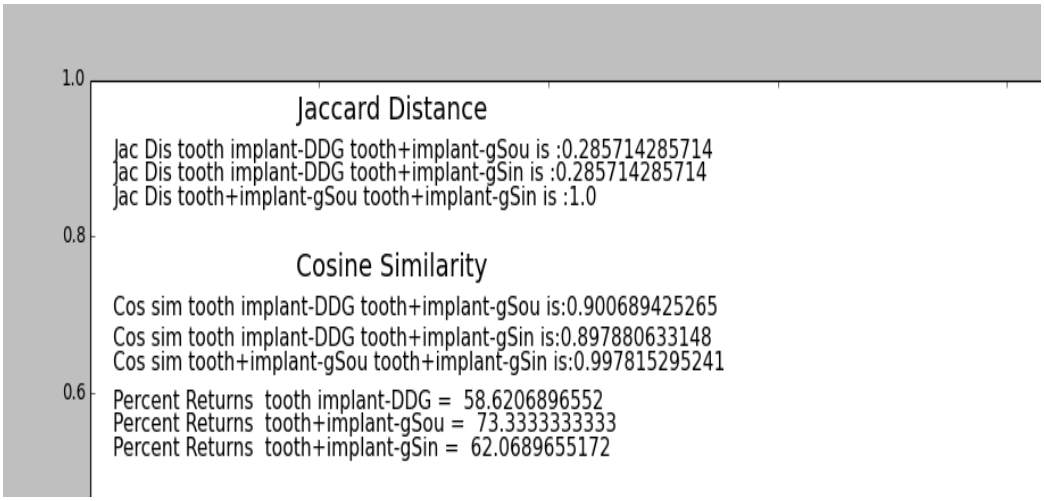

Figure 12: Tooth topic:Term Tooth implant

4.4.6 Tooth topic

This topic again showed little variation by time line or location and machine. The short period of these tests makes evaluation or establishment of trends on a time scale makes it difficult to access on a time scale. However related topics with search terms of 'tooth implant',tooth implant dublin,'tooth implant cost' are shown in the Table below. For these terms Jaccard Distance shows there is no difference between Google Signed Out/In. This is the case for either a home location or City Center location and a different machine. While Cosine Similarity only shows a dis-similarity of 0.01.

	A	B	C	D	E	F	G
1	Tooth topic	Configuration	H-tooth implant	H-tooth implant dublin	H-tooth implant cost-Dec 5	L-tooth implant cost-m2 Dec15	
2	jacrd_ab	catDdg catGin	0.29	0.38	0.25	0.25	
3	jacrd_ac	catDdg catGout	0.29	0.38	0.25	0.25	
4	jacrd_bc	cost-catGin catGout	1	1	1	1	
5	cosin_ab	catDdg catGin	0.89	0.85	0.92	0.9	
6	cosin_ac	catDdg catGout	0.89	0.89	0.91	0.9	
7	cosin_bc	cost-catGin catGout	0.99	0.99	1	1	
8							

Figure 13: Tooth topic:Term Tooth implant

Term:Tooth implant cost	Home -Tooth implant cost v Library-Tooth implant cost
Difference	[0, 0, 0.0, 0.02, 0.01, 0.0]
SAD	0.03
SSD	0.0005
Corelation	0.9997

Table 8

Table 8 shows a strong corelation of results for both Home and Library for the term 'tooth implant cost'

5 Discussion

Making this a managable project for the time period involved necessitated the selection of a small sample of test terms. Three topics of Suit,Tooth implant and Abortion were selected, and variations around these topics were used as test terms. This gave a starting point for the project and helped bring it into focus.

A different machine that was never powered up in a home location was used in a location other than home. This was to isolate or dis-entangle results from home or location identity.

Taking all three topics of abortion,suit and tooth implant, the abortion topic results indicate that using a different machine /different location combination gives a deviation from the home results. Results for the same date Dec15 indicate this with the greatest difference between Google and Duckduckgo results for both of the test terms 'pro life' and 'pro choice'. Term 'pro life' giving a corelation of 0.576 and term 'pro choice' a corelation of 0.983 with the results from the Home location on the same date. The evidence for these terms suggest that changing location and machine alter the results with greater difference for the 'pro life' term. While most of the change occurs in the Duckduckgo result, the Google Signed Out/In Jaccard distance changes by 0.1 also for the 'pro life' term. The Duckduckgo claim is that it does not track the user, the results agree with the claim as the same identity was used in Home machine as in other machine M2. But whether its the machine or location that causes the difference is stll ambigious. Taking the machine around different locations for the tests might help to dis-entangle machine from location. (Powering up machine M2 in the Home location might compromise its integrity as an isolated machine). Again this begs the question of why does Duckduckgo give different results by Location or machine.

The Suit topic showed strong or almost identical corelation on both search time line and different location/ machine combination. Corelation of 0.996

and a Sum of Squared Difference (SSD) of 0.0014. Taking different variations of search terms on the topic such as 'linen suit', 'Light suit', Linen cloth' showed more divergence among search themes. A stronger divergence on the suit topic among terms was from a gender perspective. Divergence between the term 'mens suit' and 'womans suit' gave a corelation of 0.719 and an SSD of 0.336. Most divergence was from Google Signed Out/In at 0.5 for mens suit as compared to 1 for womens suit(No difference between Google Signed Out/In). This result was from a Home location and suggests result filtering on the term 'mens suit. The profile was built on the name 'Tony Byrne' usually a man's name but along with this the profile build terms were around 'light suit', travel suit,' linen suit', 'blue suit', 'cotton suit', 'suit cloth'.
The tooth implant topic showed only slight divergence of results on either timeline or between Home and Location/Machine. Taking different themes of 'tooth implant' 'tooth implant cost' 'tooth implant dublin' gave divergence beween terms. Taking 'tooth implant costs' for Home and Library locations shows almost identical results. Corelation of 0.999 and SSD of 0.0005.
The short period for these tests and the small datasets would make it difficult to establish concrete outcomes or make definite conclusion. In the results obtained here, one conclusion is that the variation of search term used influences the outcome.(Assuming all else being equal - profile build was on both terms). Different location/machine term 'pro life' corelates at 0.576 to Home results while 'pro choice' corelates at 0.983 to Home results on same Date Dec 15. Terms 'Mens suit' v 'Womans suit' corelation of 0.719, if again assume all things equal result suggest filtering on the term 'Mens suit'. A claim that search results might have a gender bias might be too much for these initial tests.

6 Future Work

Incorporation of a datbase, time stamped data gathering and a framework for results presentation would be a next development in the project. The accumulation of results over time would provide for a bigger dataset and allow for greater manipulation by data mining methods and techniques. Further automation in gathering this data through machine cycling over a time frame would give historic trends. Although the search tests were not confined to a single location or time they were confined to a single city location. A more varied location might give a more comprehensive dataset. The correlation of the results with social media debate, general media and political events may further aid comprehension of search engine dynamics. The tests here

were confined to a small number of terms in order to make it a manageable project for the time frame involved.
Other categorisation services could also be compared to the similarsites results. A more structured test regime for carrying out the tests and the results gathering would give a more quantative and measurable outcomes.

References

Albanese,R. Now you see it: What is the internet hiding from you? pw talks with author eli pariser. *Publishers Weekly*, 258:22 –52, 2013. URL `http://www.publishersweekly.com/pw/by-topic/digital/content-and-e-books/article/47579`. (Accessed on 5th Aug 2016).

Arthur,C. Nsa scandal delivers record numbers of internet users to duckduckgo. The Guardian., 2013. URL `https://www.theguardian.com/world/2013/jul/10/nsa-duckduckgo-gabriel-weinberg-prism`. (Accessed on 4 Dec 2016).

Bar,F., MatHassan,J., and Levine,M. Methods for comparing rankings of search engine results. *Computer Networks*, 50:1448–1463, 2001.

BibAli,R. and Beg,M. An overview of web search evaluation methods. *Computers and Electrical Engineering*, 37:835 – 848, 2011.

Can,F., Nuray,R., and Sevdik,A. Automatic performance evaluation of web search engines'. information processing and managementt. *American Journal of Public Health*, 40:495–514, 2001.

Cranfield. Test collections and the cranfield paradigm, 2012. URL `https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/6IR12.pdf`. (Accessed on 21th Aug 2016).

Demeester. Overview of the trec 2014 federated web search track, 2014. URL `http://wwwhome.cs.utwente.nl/~hiemstra/papers/trec2014fedweb-draft.pdf`. (Accessed on 20th Oct 2016).

Diaz,M. Through the google goggles. Master's thesis, Stanford University, 2005.

Duckduckgo. The search engine that doesn't track you., 2013. URL `https://duck.co/help/company/advertising-and-affiliates`. (Accessed on 23th Nov 2016).

Edosomwan,J. and Taiwo,T. Comparative analysis of some search engines. *South African Journal of Science*, 106:82 – 86, 2010.

Hochstotter,N. and Lewandowski,D. What users seestructures in search engine results pages. *Information Sciences*, 179:1769 – 1812, 2009.

Holone,H. The filter bubble and its effect on online personal health information. *Croatian Medical Journal.*, 57:298 –301, 2016.

Law,M., Mintzes,B., and Morgan,S. The sources and popularity of online drug information, 2011. URL `https://www.ncbi.nlm.nih.gov/pubmed/2134340`. (Accessed on 8th Sep 2016).

Machill,M., Beiler,M., and Zenker,M. Search-engine research: a european-american overview and systematization of an interdisciplinary and international research field. *Media, Culture and Society*, pages 591–608, 2008.

Modave,F., Shokar,K., Pearanda,E., and Nguyen,N. Analysis of the accuracy of weight loss information search engine results on the internet. *American Journal of Public Health*, 100:1971–1978, 2001.

Mowshowitz,A. and Kawaguchi,A. Measuring search engine bias. *Information Processing and Management*, 41:1193 – 1205, 2005.

Online,G. P. Google adwords, 2016. URL `https://adwords.google.com/home/#?modal_active=none`. (Accessed on 10th Dec 2016).

Patterson,M. Google and search-engine market power. *Harvard Journal of Law and Technology*, 34:1 –25, 2013.

Rieder,B. and Sire,G. Conflicts of interest and incentives to bias: A microeconomic critique of googles tangled position on the web. *New Media and Society*, 16:195–211, 2013.

Rogers,R. *Introduction: behind the practice of information politics'. Information politics on the Web.* The MIT Press, Cambridge MA, 2014. URL `http://site.ebrary.com.dbgw.lis.curtin.edu.au/lib/curtinuniv/docDetail.action`. (Accessed on 10th Dec 2016).

Sanjib,K., Narendra,D., and Lahkar,N. Performance evaluation and comparison of the five most used search engines in retrieving web resources. *Online Information Review.*, 34:757 – 771, 2009.

Sebastiani,F. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1 –47, 2002.

Singh,J. A comparative study between keyword and semantic based search engines,. In *International Conference on Cloud, Big Data and Trust 2013*, pages 130 –134, Nov 13-15 2013.

Vaughan,L. New measurements for search engine evaluation proposed and tested. *Information Processing and Management*, 40:677 –691, 2013.

Voorhees,E. and Harman,D. Overview of trec 2001, 2001. URL `http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf`. (Accessed on 14th Oct 2016).

Wills,R. Google's pagerank: The math behind the search engine. *search*, 28: 1 –15, 2013.

Zhang,B. A comparison of search engines for finding resources. information network applications taught by dr. gretchen whitney at the school of information science in the university of tennessee, knoxville, 2016. URL `http://www.yuanlei.com/studies/articles/is567-searchengine/page2.htm`. (Accessed on 23th Nov 2016).