CA4009: Search Technologies:
Research and Development Project on *Social Network Search*
Programme: Enterprise Computing
Module: CA4009, Search Technologies
Name(s): Maksims Kompanijecs - 15309671 - kompanijecs.maksims2@mail.dcu.ie
Matthew Farrelly - 15366246 - matthew.farrelly68@mail.dcu.ie
Michael Huben - 15396501 - michael.huben2@mail.dcu.ie
Josh Malone - 15357971 - josh.malone27@mail.dcu.ie
Date of submission: 13/12/2019

1.Plagiarism statement

A report submitted to Dublin City University, School of Computing for module CA4009: Search Technologies, 2019/2020.

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious. I have read and understood the DCU Academic Integrity and Plagiarism Policy.

I accept the penalties that may be imposed should I engage in practice or practices that breach this policy. I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

By signing this form or by submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and

Plagiarism Policy (available at: http://www.dcu.ie/registry/examinations/index.shtml)

Name(s): Maksims Kompanijecs, Matthew Farrelly, Michael Huben, Josh Malone
Date: 17/12/2019

2.Abstract

In this report we will describe our new social network search proposed application. GoSocial will allow users to search for content that is relevant to them, and will return the most popular/relevant results from multiple social media platforms. The platform will consist of content from five major social networking sites - Facebook, Instagram, YouTube, Twitter and Reddit.

The main goal for this social media search technology is to combine all the search techniques of the major social networks mentioned above into one. The motivation for this is that most of the content produced on different social media sites is the same, with users constantly refreshing their feed to view the exact same content on multiple platforms. Therefore, we believe that this system would allow users to eliminate time spent on different social media platforms and consume content more efficiently The system we have in place has two major functions:

1. Scan different social media network sites for images, text, trends based on the query provided
2. Return results based on the query provided that are most relevant and up to date

In the scientific functional description, we will go into the details of overall architecture and algorithms to be used, so that is clear on what components are being used and how they are combined into one search application system. We will also outline the pros and cons of the proposed system in terms of the technology and what limitations may occur.. Any assumptions or approximations that we make about any aspect of the system will be also outlined.

3.Introduction

Social network search is a method of retrieving information/data using different social media platforms search engines that usually only contain other user-generated content such as text, images and videos etc. Social networking sites have grown massively due to the use of the Internet and emerging technology. According to [1] Statista, approximately 2 billion users used social networking and social media apps back in 2015, since then, with the increased use of mobile devices, this number recently crossed the 2.6 billion mark in 2018.[1]

Five of the major social networking sites that we have chosen for our proposed social networking search system are:

1. Facebook - biggest social media networking site in the world
2. Twitter - largest microblogging social media network where people communicate in ‘tweets’
3. Instagram - largest visual social media networking media platform
4. Youtube - largest video social media networking platform
5. Reddit - social media networking and news aggregator platform

Our motivation and objective behind our proposed search system is to help people reduce the amount of time they spend searching for content within different social media networking sites. It is clear to see that nearly all of the content on the Internet is very similar, the difference is merely what social media platform it is posted on and in what form e.g. text, image, video etc.

A universal platform will save users time and allow them to consume relevant content more efficiently. In the simplest meaning, we hope to propose a system that will scan social networking sites and return results based on the most relevant and up-to-date information.

There will be quite a few limitations and possible constraints with proposed system in mind. Social media content has a number of different styles, it can be of different languages, formats and length. It is also full of poor quality content, that can possibly contain typos and poor choice of vocabulary. There is also quite a lot of misinformation and duplicate content, this can be mainly used to attract attention and publish the same content on the same site.

4.User Analysis

Our proposed system is primarily for any people with that browse social networking sites regularly. We also assume that the users of our system will have a good understanding and knowledge of the 5 social networking platforms that we have chosen and will know how to use and freely navigate through them.

And with this system implemented we feel like users will have the ability to process and look at different types of contents more efficiently. For our proposed system the main goal for this search engine is to combine and provide users with the ability to look at the most popular and relevant content across different social media platforms. According to [2] BroadBandSearch, in 2018 users spent an average of 144 minutes a day using social networking sites. That's an increase of 1 hour a day, or 62.5% over 2012.[2] People use social media for a number of reasons. It allows people to stay in touch, across different continents and timezones, but also with friends that could live next door down from us. It's a very powerful tool it connects it's users in a way that previous generations couldn't even dream about.

5. Problems with Social Network content

There are quite a few problems with social media content and it's very important that we address these issues, so that our search engine can provide the most relevant and up to date content for its users. The main problems with social media content are as follows:

- Content can be of different variety, language and style
- Poor quality content is also possible, content with typos and poor choices of words
- Deliberate misinformation known as spam, usually done to attract attention
- Duplicate content, very similar or duplicate information published

6.Scientific Functional Description

The following section will describe different components that our system contains and go into overall architecture and algorithms used in the system. Four main components of our proposed system are:

1. Document collection - collection of the documents
2. Document preprocessing - processing and understanding the contents of the documents
3. Document indexing - storage of documents for efficient information retrieval
4. Document ranking - ranking of the most relevant documents

6.1 Document collection - web crawling

Our proposed system will need a functionality that must be able to navigate the WWW to collect documents (HTML pages) based on 4 social media networking sites that we have chosen. Primary job of this function is to create a copy of all the visited pages, which we can parse and index them which will allow our engine to provide faster searches. Having so many HTML pages on the internet, the process of web crawling could go on for a very very long time, and most web crawlers only crawl pages that are relevant. This is achieved by following different policies, that will make web crawler much more selective about what pages it should crawl and in which order. It is also important for a web crawler to follow a policy that will outline on how often they should check the same pages for any content updates. An example of a 'polite' web crawler' includes a number of good practices[3].

1. Respecting robots.txt, which is usually stored in **url/**robots.txt
2. Don't degrade websites performance, which can be achieved by having time delay between requests
3. Identify website being crawled with your personal information[3]

A web crawling algorithm that we have chosen for our proposed search application is A* search algorithm, which is an example of best first search algorithm.[4]A* uses Best First Search. It calculates the relevancy of each link and the difference between expected relevancy of the goal web-page and the current link. The sums of these two values serve as the measure for selecting the best path. [4]

We chose this algorithm, because in this approach the relevancy values are calculated for each link and this is important for social network search engine to be providing the most relevant content for its users.

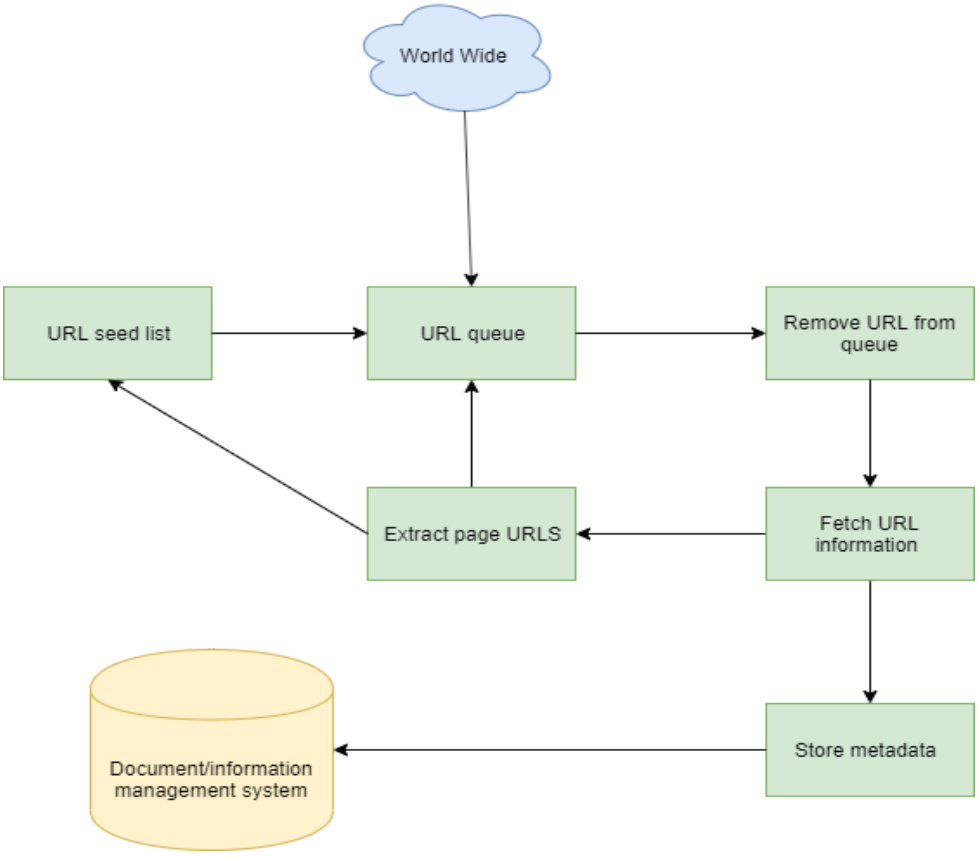

Fig 1, web crawler diagram

6.2 Document pre-processing

It is important that we process and understand the document contents. In simplest form during preprocessing we will take the documents that we have collected using web crawler and split them into indexable text tokens. Our proposed system will achieve this by carrying a number of preprocessing tasks which will be summarised below.

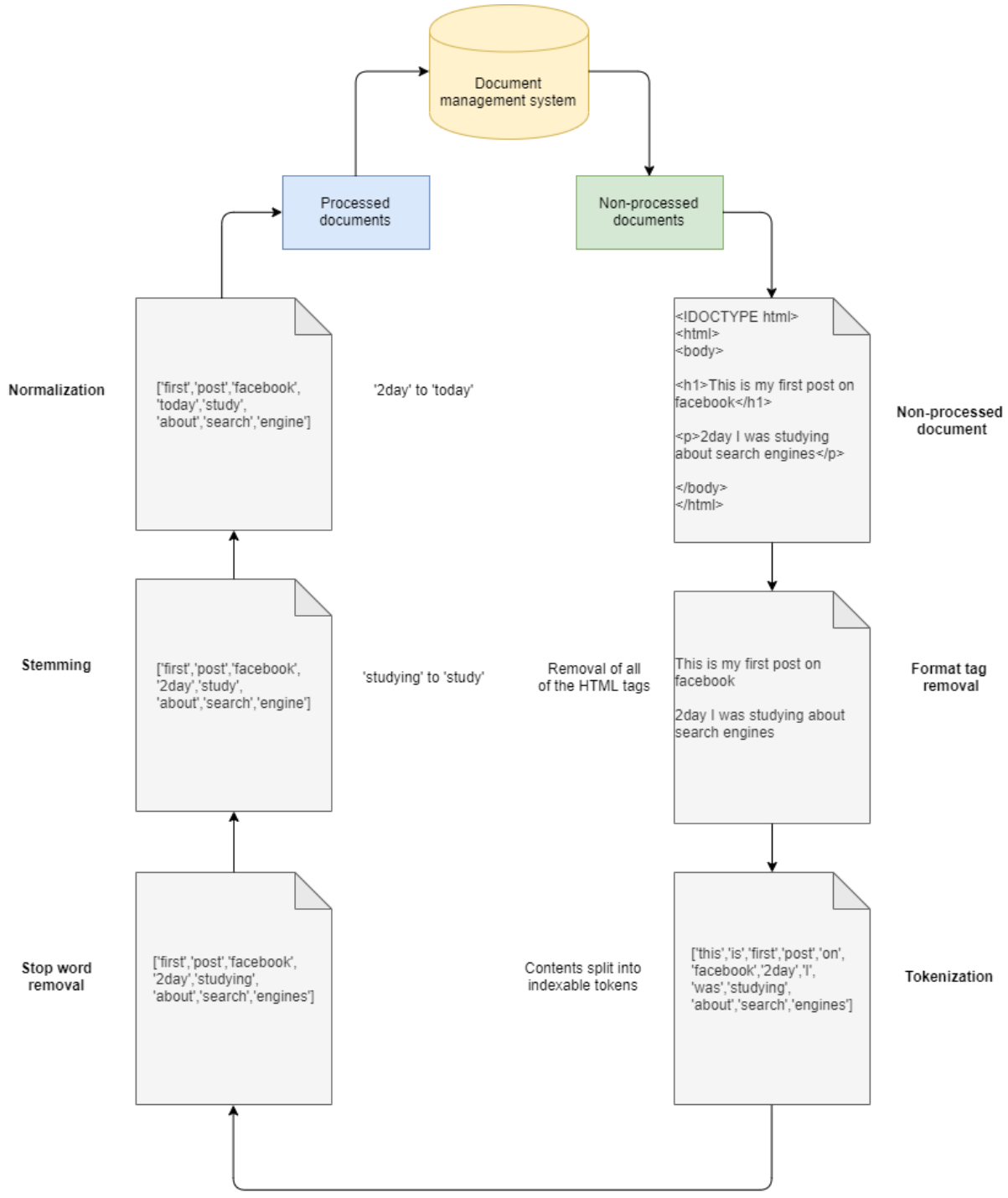

Fig 2, document pre-processing

6.2.1 Format tag removal

The documents that we collect using the web crawler will need to have any format tags removed. Format tags can include a lot of irrelevant information that is not directly related to

the document. Our proposed system can achieve by using a simple Python script and by using a regex or BeautifulSoup library.

6.2.2 Tokenization

Tokenization is the act of breaking up any sentences, strings into words, numbers, symbols, phrases or symbols. Tokens can be anything, it can be numbers, words or even a full sentence. In simple terms, tokenization is [5] extracting indexing units from text. processed token to be used for search is referred to as a search term [5]

6.2.3 Stop word removal

One of the most important text-preprocessing techniques is the stop word removal. Stop words is the set of most commonly used words in any language. Example of stopwords in English are: ‘a’,’as’,’at’,’the’,’in’ and so on.

6.2.4 Stemming

Another text-preprocessing technique that allows for much more efficient inverted indexing technique is stemming. Stemming is transformation of the word back to its root form. For example, if I want to search for ‘play guitar’, and the document present that contains the information I’m looking for has words like ‘played guitar’ or ‘playing guitar’. In order to eliminate the possibility of this I can transform ‘playing’ and ‘played’ to ‘play’ and get to the root word.

6.2.5 Normalization

Normalization is another text-preprocessing technique that will benefit our information retrieval system. In simplest form, normalization is the transformation of text from internet slang into to its standard form.

6.3 Document indexing

Efficient information retrieval can be achieved by indexing. Next component of our system will be process of document indexing that we have collected using web crawler. This feature will help us to have indexed documents programmed into a document management system which will allow our search engine to easily access the data it requires. Documents in the management system can only be accessed by knowing via information that’s stored about them. The way that we have chosen the documents without our information management system is by using inverted index technique. Inverted index technique is a type of data storing technique that maps the documents words, numbers, text to its location in a document. The process of using inverted index technique will be illustrated below.

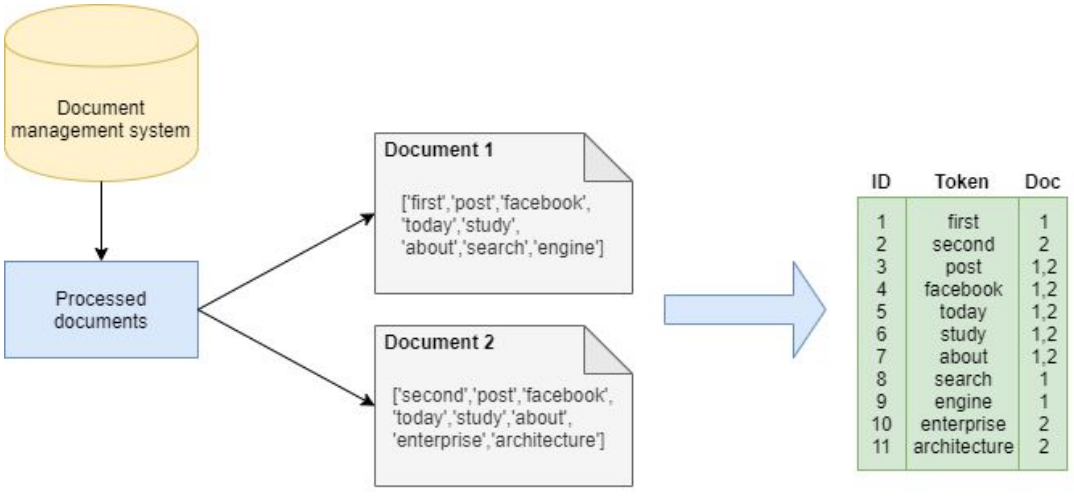

Fig 3, inverted index technique[7]

6.4 Document ranking

In order for our proposed system to display relevant results, there would need to be a process or an algorithm in place by which the stored inverted index information is ordered in the most relevant way for the user. A big problem that our users may encounter is the return of the results that can be spam or deliberate misinformation to attract attention. We feel like we could tackle this problem by using relevance feedback techniques. This technique would be the ability to give the users of this search engine platform the ability to upvote and downvote content based on their view if they think if the information returned is relevant or not. If the document in question receives a lot of upvotes from the users, we can associate a higher relevance number for the document for the given query and vice versa if the document received a significant number of downvotes. To summarise this, for the start we hope to determine relevance on information returned based on the search query to be determined by the users that use our platform.

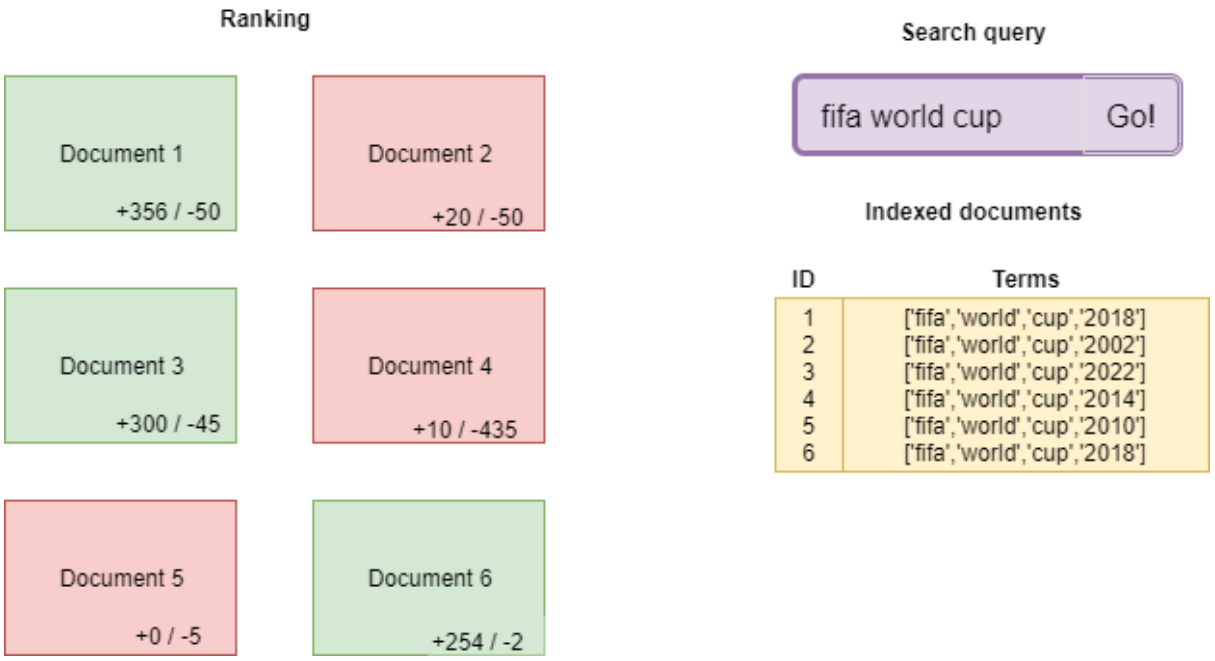

Fig 4, document ranking

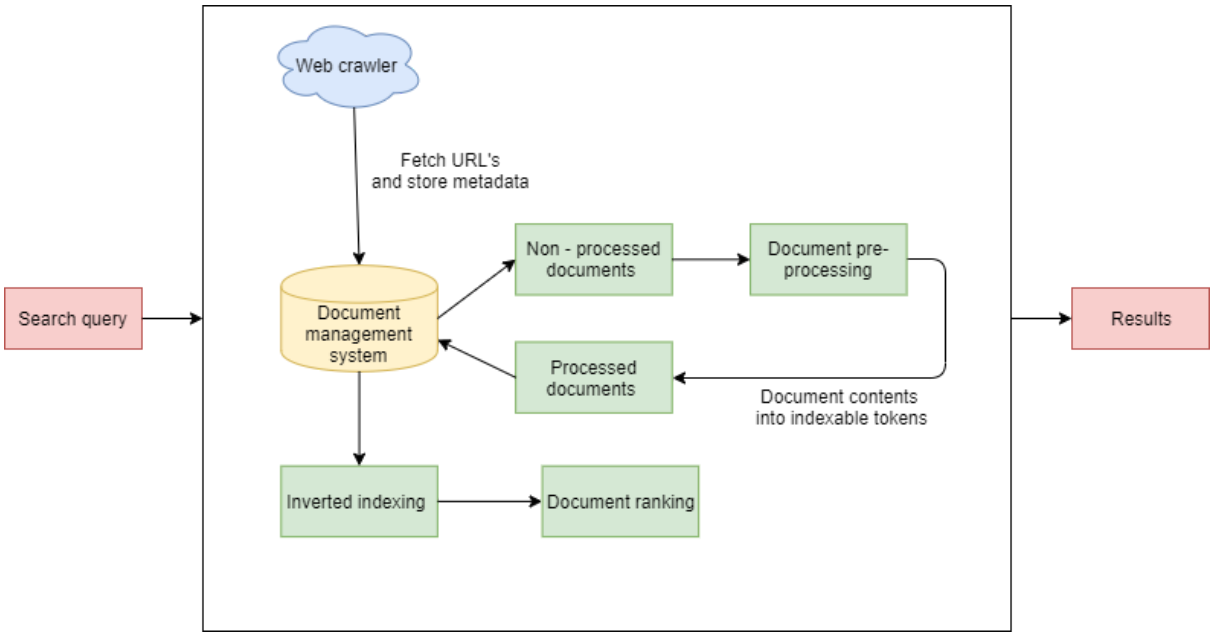

Fig 5, overall architecture of the proposed system

Evaluation

The purpose of this proposed system is to provide users with the ability to minimise the amount of time spent on different social media platforms and consume the content more efficiently. Essentially with this proposed system in mind, we are hoping to combine the content from all of the different social media platforms into one which will hopefully benefit the users by allowing them to reduce the time they spend on social media platforms looking at the same content. We believe that this proposed system will work best with the users that have somewhat experience using social media platforms and would know their way around them.

One of the biggest challenges that comes with our proposed search engine is the relevance of the content returned, this is due to the fact that the content relevance is calculated based solely by users and how they interact with it. This is definitely a big step forward to automate the process of determining the relevance and it would be very interesting to look into how this could be done. We could also maybe assess the relevance data by collecting a random number of search queries and the number of documents that they are returned when the search query is executed.

We could also evaluate the effectiveness of our system by using query expansion, expansion of query with additional terms could tell us a lot about how efficient information retrieval of our system actually is. To accurately evaluate our proposed system further, we must calculate both precision and recall values and to further fine tune our technology from the beginning of its life-cycle, we intend to implement BM25 test collections. These collections will begin with standardised b and k values. The product we have in mind is heavily used focused on the users that use this system, so it's important that we their values and preferences are always taken into consideration first.

Concluding Section

To conclude the report, we believe that this search engine could be an efficient and convenient app for people. We made sure that the use of our search platform is as easy as can be for all types of users like splitting the search into different filtered sections like people, pages and articles
It would best be described as a one stop social media browser that can search popular social media artefacts to the users desire.

Our main target demographic would be active social media users that don't need any particular skill-set. Having said this, it is still a platform older people can use to its full extent with a bit of practice and repetition.
It will parse the data from these different social media sites with all documents being indexed and ranked based on whether or not the user found it informative or helpful to what they were looking for. The results will be displayed to the end user with a link to the specific social media platform.

Next steps for our proposed system would include taking a better look at how we can improve the relevance of content that is returned.

References

[1] J. Clement, “Number of social network users worldwide from 2010 to 2021 (in billions)” **Available:** https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ **Accessed**: December, 6, 2019 **[Online]**

[2] BroadBandSearch, “Average Time Spent Daily on Social Media (with 2019 Data)” **Available:**https://www.broadbandsearch.net/blog/average-daily-time-on-social-media **Accessed:** December 6, 2019 **[Online]**

[3] Valdir Stumm Jr, The Scrapinghub Blog, “HOW TO CRAWL THE WEB POLITELY WITH SCRAPY” **Available:** https://blog.scrapinghub.com/2016/08/25/how-to-crawl-the-web-politely-with-scrapy **Accessed:** December 10, 2019 **[Online]**

[4]Aviral Nigam , “Web Crawling Algorithms” **Available:** http://www.academicpub.org/DownLoadPaper.aspx?paperid=16048 **Accessed:** December 15, 2019 [**Online]**

[5] Gareth Jones ,”Section 3: Text Retrieval” **Available:** https://loop.dcu.ie/pluginfile.php/2865816/mod_resource/content/15/ir.pdf **Accessed:** December 11, 2019 **[Online]**

[6] Wikipedia, “Search engine indexing” **Available:** https://en.wikipedia.org/wiki/Search_engine_indexing **Accessed:** December 16th, 2019 **[Online]**

[7] Hitachi, “Search: The Inverted Index” **Available:** https://community.hitachivantara.com/s/article/search-the-inverted-index **Accessed:** December 17th, 2019 **[Online]**