G52GRP Interim Group Report

Project Title:

Crowdsourcing and Data Validation Tools

Date: 04/12/2017

Prepared by **Group 1**:

Anas Nader Almasri	khcy6ana	023789
Amirul Umar Pandai	khcy6aup	023484
Mayur Gunputh	khcy6mgg	025584
Stephen Kua	khcy6skk	016806
Oh Lean Kai	khcy6olk	025344

Supervisor: Mr. KR Selvaraj

Table of Contents:

1. Abstract

Data Science is a relatively new field of study. Agricultural research is one of those mildly marginalized industries when it comes to Data Science. This report goes over a project which was initiated by the need to collect data about certain crops from people worldwide via Crowdsourcing to ultimately help utilize crops that are currently underutilized. The project specifies that a fully-functioning system shall be built. This system is going to be asked text-based questions. The answers will be derived from the data returned by the automated analysis algorithms integrated into the system. These answers are going to help fill up the gaps in the client's database and, possibly, pivot attention towards new aspects of crop utilization. This report is equivalent to documentation of the development stage so far in the project.

2. Introduction

2.1 Purpose

The aim of this project is to collect data about crops from people around the world through Crowdsourcing from social media platforms and blog forums. This data will then be analyzed, sorted and then represented in graphical form to help people using the research tool of Crops For The Future to get an idea of what crop would grow most effectively in what region and in which seasons of the year.

2.2 Client

Crops For the Future (CFF), the world's first research center dedicated to study and analyze underutilized crops for food and non-food purposes. CFF's research focuses on the uses of underutilized crops and agricultural biodiversity to diversify crop and agricultural systems, address changing climates, improve food security, economic well-being and nutrition.

2.3 Project Scope and Deliverables

Group members intend to constantly develop previews of their work to present to the client, ask for their feedback and amend the group's project accordingly. Members shall utilize required datasets and implement a script in MySQL, PHP, HTML, CSS, JavaScript amongst other languages to look up through the datasets for specific keywords entered by the user. The results returned are then to be analysed and sorted. This project requires digital and physical deliverables. This includes reports, software and presentations.

3. Literature Review

3.1 Data Science Industry

Data Science encompasses a wide range of fields involving the extraction of knowledge or insights from data in various forms, either structured or unstructured, employing techniques drawn from mathematics, statistics, Information Science and Computer Science. Jim Gray, American computer scientist and 1998 Turing award winner, imagined Data Science as a "fourth paradigm" of science, the first three being empirical, theoretical and computational, and now data-driven, asserting that "everything about science is changing because of the impact of Information Technology." **[1]**

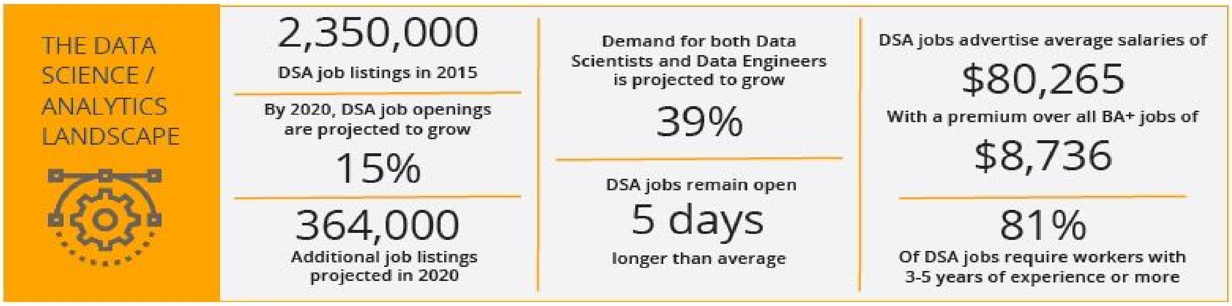
Data Science as an industry usually refers to its ability to harness big data, data sets so large or so complex that traditional data processing application software would be inadequate to deal them. Peter Norvig, Google's AI expert, summarises the power of big data as "Simple data and a lot of data trump more elaborate models based on less data," referring to the value big data has over traditional data handling. **[2]** The disruptiveness of data science is further amplified by the fact that data science has been used to provide more accurate answers, affect the way decisions are made on a business standpoint, and able to identify trends. **[3]**

3.2 Crowdsourcing as A Part of Data Science

Crowdsourcing is a specific mode of "sourcing" achieved by dividing work between participants to achieve a collective result, specifically with users from the Internet, although this mode of sourcing was already successful prior to the digital age. **[4]** Crowdsourcing can be distinguished from outsourcing in that the work can come from an undefined public instead of coming from a specific, named group; based on public data rather than personalized data, therefore there are no ethical overlaps with the ongoing process of development. Advantages of using crowdsourcing may include lower costs, improved speed, better quality, flexibility, scalability or diversity. **[5]** Crowdsourcing has also been used for noncommercial work and to develop common goods, such as the Wikipedia website which, coincidentally, is one of the main sources of data collection for this project. **[6]**

3.3 Growth of The Industry

The data science industry as a whole is one of the largest and most important industries in the technology sector, and it will only grow bigger over the coming years. By 2020, IBM projects the number of data professionals will increase by 364 thousand openings to 2.72 million in the United States alone, up from 2.35 million in 2015, with advertised salaries averaging US$80,265. **[7]**



THE DATA SCIENCE / ANALYTICS LANDSCAPE

2,350,000
DSA job listings in 2015

By 2020, DSA job openings are projected to grow
15%

364,000
Additional job listings projected in 2020

Demand for both Data Scientists and Data Engineers is projected to grow
39%

DSA jobs remain open
5 days
longer than average

DSA jobs advertise average salaries of
$80,265
With a premium over all BA+ jobs of
$8,736

81%
Of DSA jobs require workers with 3-5 years of experience or more

Image source: https://www.ibm.com/analytics/us/en/technology/data-science/images/The-data-savvy-job-landscape.jpg

3.4 Project Participation in The Industry

CropBASE, the database of this project's client, has a huge data set that this project aims to broaden through the use of automated algorithms by implementing Crowdsourcing. Although the goal of this project is to generate more meaningful data from an unstructured source, the ultimate goal of CFF is to contribute to the facilitation of the United Nations

Sustainable Development Goals (UNSDGs). This project is a small solution out of many others that can further the cause posed by the UNSDGs. This is a new application of Crowdsourcing that can help change the utilization of crops in order to save the environment and shift the focus towards environmentally-safe behaviour of human beings. This project is capable of making the public hear those innovative people who are far away from being heard through other means.

---

4. Project Description

4.1 Problem Description

Crops For The Future (CFF) has a huge database named CropBASE that stored 404 million pieces of data about one crop. All the information are to be found in the database including quality of a crop, soil details, advices of growing the particular crop, etc. However, more data is required in order to have more accurate information about crops. Thus, CFF wants to widen the database through machine learning algorithms that facilitate Crowdsourcing.

4.2 Underlying Assumptions

A problem with Crowdsourcing is that it includes millions of irrelevant pieces of data that are not useful to the industry. Through that huge amount of data, the algorithm should be able to answer defined questions (preferably ten questions). Questions like “Do people like plant X? In what country?”. Another goal is to get useful information from people far away in different continents (e.g. African farmers) on certain crops and how their utilization works. Manually, this is not possible, but through automated algorithms of current technologies, it probably will be.

4.3 Operating Environment

This project is a part of a bigger image into which it is going to be integrated by the clients. The client's operating environment that was disclosed to group members consists of a huge database, a data-based User Interface and a local server. On the other hand, group members are working on their personal computers (with their preferred Operating Systems). This work is carried into another environment. This is furtherly discussed in chapter 7, *Implementation Description*, section 7.2 *Development Environment*.

4.4 Design Implementation Constraints

Since this is considered a mildly large project, design consists of three different stages. There was no suggested design for the User Interface by the client. It was just made clear that the UI must be somewhat interactive, taking text-based questions as an input and returning answers on based on those questions.

4.5 Target Audience

As emphasized by the client, the system will only be used by the client, CFF, for now. They will be the ones validating the system in general and doing the final run by asking at least 10 questions to it. In the future, it could be integrated into other systems and made public, which is completely up to the clients themselves.

5. Group Organization Methodology

5.1 Workload Division

Workload is usually divided equally, among members depending on their capabilities and skills. Workload is also divided based on the expertise and main interests of each member. Prior to tackling any problem, the missions are reviewed by all members and members choose their tasks accordingly while the leader keeps track of the fair division.

5.2 Responsibilities

During the first group meeting, responsibilities of each member were made clear. Apart from the group leader being elected by the members, one member was made responsible for the User Interface design. Another member is responsible for any connection with the data-access layer on the server side. The other two members are considered the main developers of the project.

Group leader’s main responsibility is to ensure that the project is on the right track and that work is being carried out by all members. By setting a milestone for every task, the leader gives the other members a time frame and a clear picture of the task they are working on . The top priority of the front-end/UI designer is maximizing usability and accessibility to provide users with the best experience. The front-end designer works closely with the back-end developer to ensure all functionalities are implemented. The back-end/server-access developer is responsible for the interaction of data between the server and users. **[8]**

It is worth mentioning that the role of each member encapsulates and utilizes their skills, but members are not required to solely work on the matter at hand if it is within their field of expertise or interest. Everybody works on all the parts, but the last - and first - say for each part is up to the person responsible for that particular part.

5.3 Meeting System

Meetings in this project take different forms, depending on the objective of each, and therefore the attendees of each meeting differ. For formal meeting, it is made sure of that all members attend, along with client’s representatives. Supervisors are always welcome to attend, as they are always informed of every meeting before it is held.

On the other hand, informal meetings are only attended by group members. Most informal meetings are stand-up meetings. Those are held by the group members 2-3 times a week. The objective of these meetings is always to keep track of what every member is working on, what problems they are dealing with, and what is their next task on the list. Other informal meetings are normal meetings. Those usually are longer than stand-up meetings and have multiple objectives that are prepared beforehand by the group leader based on the progress of the project and the need for new actions.

Formal or informal, the most important meetings are recorded in the project log, enclosed within *Appendix A* of this report. These records include the objectives, discussions, action points, and attendance list of every meeting.

5.4 Development Methodology

As the group members facilitate Agile Extreme Programming (XP) methodology, they base the work on four simple values – simplicity, communication, feedback and courage. Extreme programming encourages starting with the simplest solution and relies on that extra functionality can then be added later on in programming. **[9]**

---

6. Requirement Specification

6.1 Functional Requirements

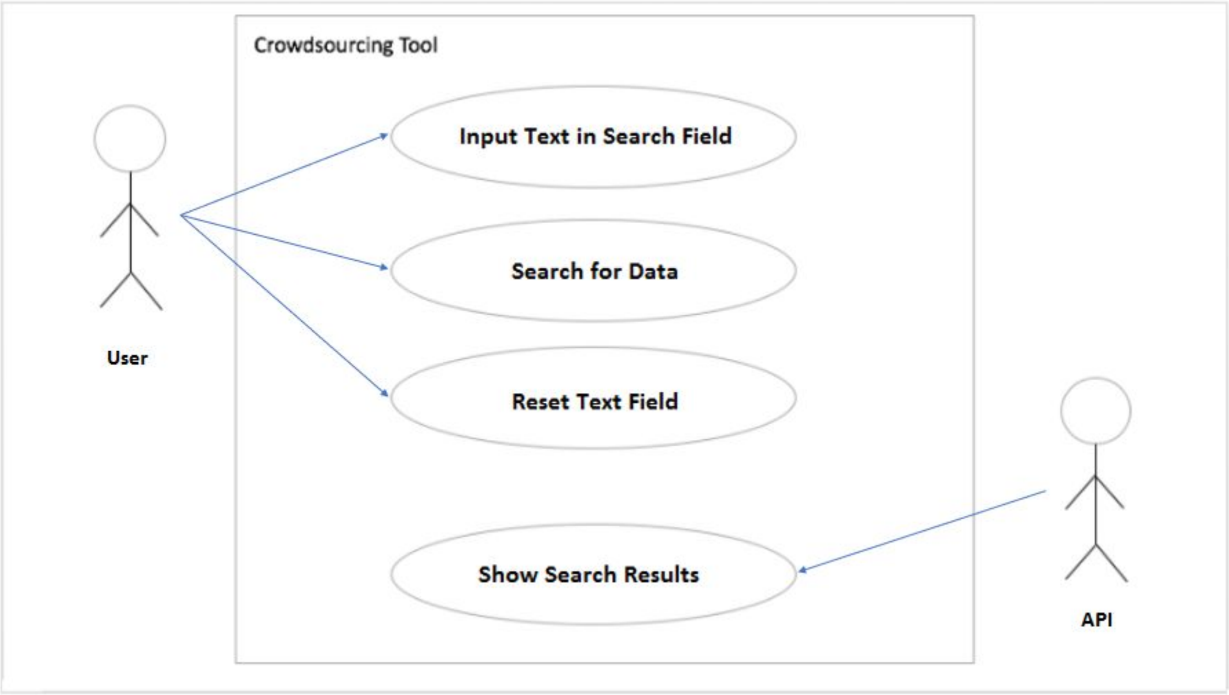
- Enter questions into search bar.
- Search for answers using a Search button.
- Reset text fields based on a button click.

6.2 Non-functional Requirements

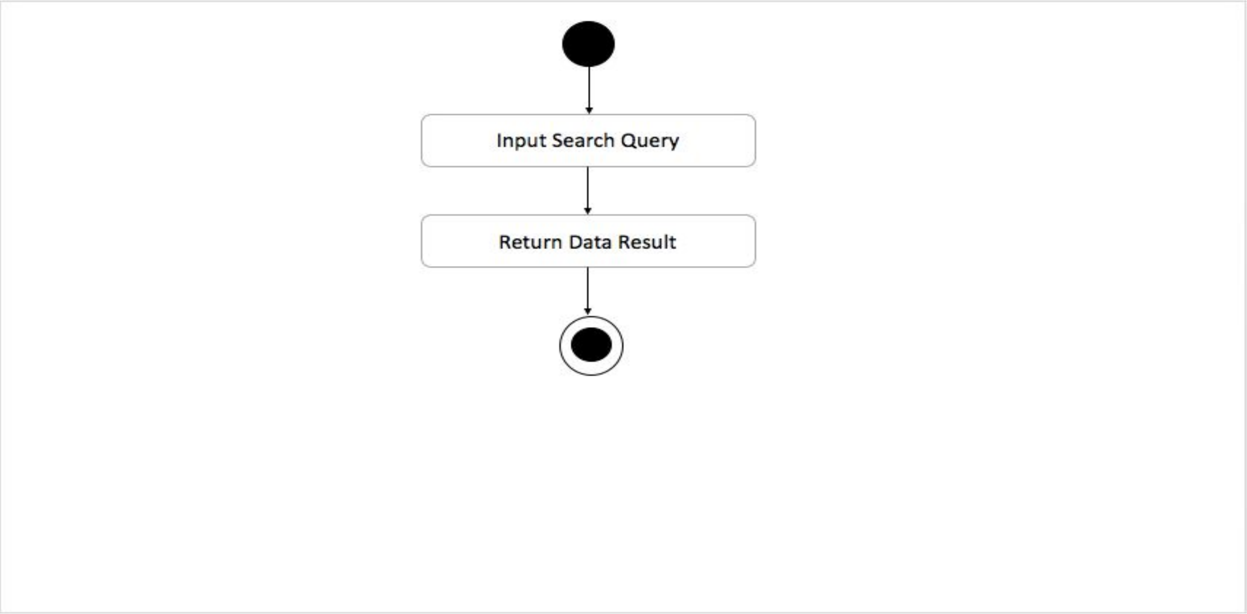
- Access dataset by using provided username and password (if required)
- Query the database for the question searched for.
- Perform analyses (inc. Sentiment Analysis) on data returned.

- Sort results obtained from the analysis algorithms.
- Display sorted results in representational answer-form.

6.3 Use Case Diagram

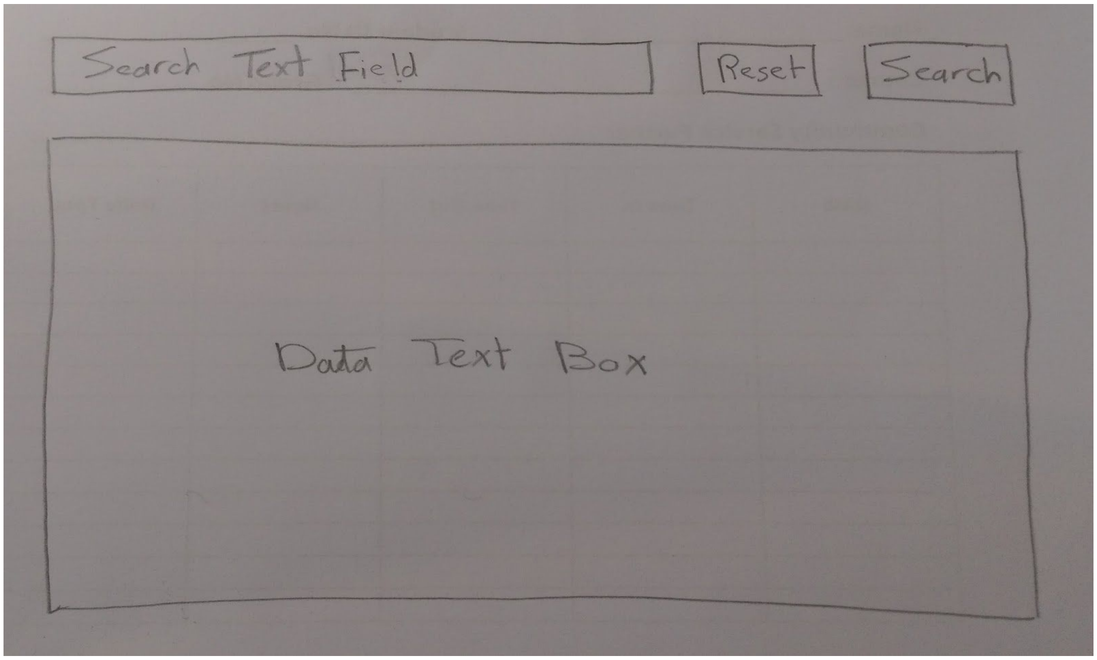


6.4 State Diagram

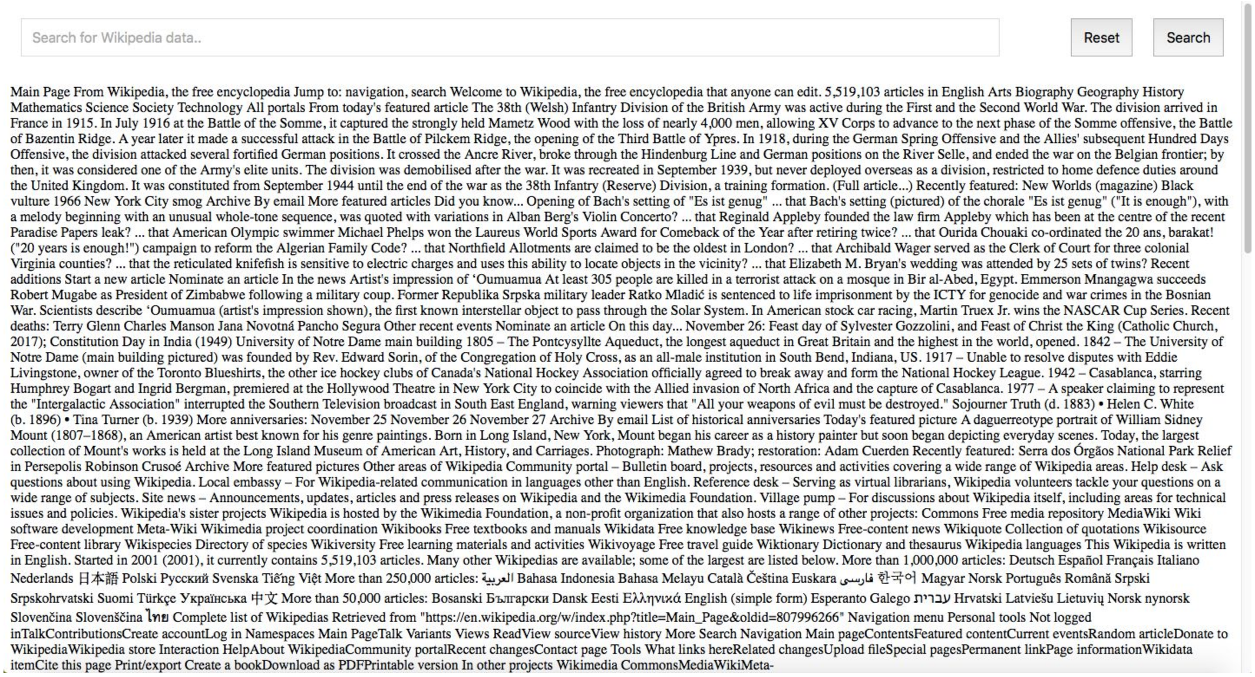


7. Prototypes and User Interface Design

7.1 Low-fidelity Prototype



7.2 Mid-fidelity Prototype



Main Page From Wikipedia, the free encyclopedia Jump to: navigation, search Welcome to Wikipedia, the free encyclopedia that anyone can edit. 5,519,103 articles in English Arts Biography Geography History Mathematics Science Society Technology All portals From today's featured article The 38th (Welsh) Infantry Division of the British Army was active during the First and the Second World War. The division arrived in France in 1915. In July 1916 at the Battle of the Somme, it captured the strongly held Mametz Wood with the loss of nearly 4,000 men, allowing XV Corps to advance to the next phase of the Somme offensive, the Battle of Bazentin Ridge. A year later it made a successful attack in the Battle of Pilckem Ridge, the opening of the Third Battle of Ypres. In 1918, during the German Spring Offensive and the Allies' subsequent Hundred Days Offensive, the division attacked several fortified German positions. It crossed the Ancre River, broke through the Hindenburg Line and German positions on the River Selle, and ended the war on the Belgian frontier; by then, it was considered one of the Army's elite units. The division was demobilised after the war. It was recreated in September 1939, but never deployed overseas as a division, restricted to home defence duties around the United Kingdom. It was constituted from September 1944 until the end of the war as the 38th Infantry (Reserve) Division, a training formation. (Full article...) Recently featured: New Worlds (magazine) Black vulture 1966 New York City smog Archive By email More featured articles Did you know... Opening of Bach's setting of "Es ist genug" ... that Bach's setting (pictured) of the chorale "Es ist genug" ("It is enough"), with a melody beginning with an unusual whole-tone sequence, was quoted with variations in Alban Berg's Violin Concerto? ... that Reginald Appleby founded the law firm Appleby which has been at the centre of the recent Paradise Papers leak? ... that American Olympic swimmer Michael Phelps won the Laureus World Sports Award for Comeback of the Year after retiring twice? ... that Ourida Chouaki co-ordinated the 20 ans, barakat! ("20 years is enough!") campaign to reform the Algerian Family Code? ... that Northfield Allotments are claimed to be the oldest in London? ... that Archibald Wager served as the Clerk of Court for three colonial Virginia counties? ... that the reticulated knifefish is sensitive to electric charges and uses this ability to locate objects in the vicinity? ... that Elizabeth M. Bryan's wedding was attended by 25 sets of twins? Recent additions Start a new article Nominate an article In the news Artist's impression of ‘Oumuamua At least 305 people are killed in a terrorist attack on a mosque in Bir al-Abed, Egypt. Emmerson Mnangagwa succeeds Robert Mugabe as President of Zimbabwe following a military coup. Former Republika Srpska military leader Ratko Mladić is sentenced to life imprisonment by the ICTY for genocide and war crimes in the Bosnian War. Scientists describe ‘Oumuamua (artist's impression shown), the first known interstellar object to pass through the Solar System. In American stock car racing, Martin Truex Jr. wins the NASCAR Cup Series. Recent deaths: Terry Glenn Charles Manson Jana Novotná Pancho Segura Other recent events Nominate an article On this day... November 26: Feast day of Sylvester Gozzolini, and Feast of Christ the King (Catholic Church, 2017); Constitution Day in India (1949) University of Notre Dame main building 1805 – The Pontcysyllte Aqueduct, the longest aqueduct in Great Britain and the highest in the world, opened. 1842 – The University of Notre Dame (main building pictured) was founded by Rev. Edward Sorin, of the Congregation of Holy Cross, as an all-male institution in South Bend, Indiana, US. 1917 – Unable to resolve disputes with Eddie Livingstone, owner of the Toronto Blueshirts, the other ice hockey clubs of Canada's National Hockey Association officially agreed to break away and form the National Hockey League. 1942 – Casablanca, starring Humphrey Bogart and Ingrid Bergman, premiered at the Hollywood Theatre in New York City to coincide with the Allied invasion of North Africa and the capture of Casablanca. 1977 – A speaker claiming to represent the "Intergalactic Association" interrupted the Southern Television broadcast in South East England, warning viewers that "All your weapons of evil must be destroyed." Sojourner Truth (d. 1883) • Helen C. White (b. 1896) • Tina Turner (b. 1939) More anniversaries: November 25 November 26 November 27 Archive By email List of historical anniversaries Today's featured picture A daguerreotype portrait of William Sidney Mount (1807–1868), an American artist best known for his genre paintings. Born in Long Island, New York, Mount began his career as a history painter but soon began depicting everyday scenes. Today, the largest collection of Mount's works is held at the Long Island Museum of American Art, History, and Carriages. Photograph: Mathew Brady; restoration: Adam Cuerden Recently featured: Serra dos Órgãos National Park Relief in Persepolis Robinson Crusoé Archive More featured pictures Other areas of Wikipedia Community portal – Bulletin board, projects, resources and activities covering a wide range of Wikipedia areas. Help desk – Ask questions about using Wikipedia. Local embassy – For Wikipedia-related communication in languages other than English. Reference desk – Serving as virtual librarians, Wikipedia volunteers tackle your questions on a wide range of subjects. Site news – Announcements, updates, articles and press releases on Wikipedia and the Wikimedia Foundation. Village pump – For discussions about Wikipedia itself, including areas for technical issues and policies. Wikipedia's sister projects Wikipedia is hosted by the Wikimedia Foundation, a non-profit organization that also hosts a range of other projects: Commons Free media repository MediaWiki Wiki software development Meta-Wiki Wikimedia project coordination Wikibooks Free textbooks and manuals Wikidata Free knowledge base Wikinews Free-content news Wikiquote Collection of quotations Wikisource Free-content library Wikispecies Directory of species Wikiversity Free learning materials and activities Wikivoyage Free travel guide Wiktionary Dictionary and thesaurus Wikipedia languages This Wikipedia is written in English. Started in 2001 (2001), it currently contains 5,519,103 articles. Many other Wikipedias are available; some of the largest are listed below. More than 1,000,000 articles: Deutsch Español Français Italiano Nederlands 日本語 Polski Русский Svenska Tiếng Việt More than 250,000 articles: العربية Bahasa Indonesia Bahasa Melayu Català Čeština Euskara فارسی 한국어 Magyar Norsk Português Română Srpski Srpskohrvatski Suomi Türkçe Українська 中文 More than 50,000 articles: Bosanski Български Dansk Eesti Ελληνικά English (simple form) Esperanto Galego עברית Hrvatski Latviešu Lietuvių Norsk nynorsk Slovenčina Slovenščina ไทย Complete list of Wikipedias Retrieved from "https://en.wikipedia.org/w/index.php?title=Main_Page&oldid=807996266" Navigation menu Personal tools Not logged inTalkContributionsCreate accountLog in Namespaces Main PageTalk Variants Views ReadView sourceView history More Search Navigation Main pageContentsFeatured contentCurrent eventsRandom articleDonate to WikipediaWikipedia store Interaction HelpAbout WikipediaCommunity portalRecent changesContact page Tools What links hereRelated changesUpload fileSpecial pagesPermanent linkPage informationWikidata itemCite this page Print/export Create a bookDownload as PDFPrintable version In other projects Wikimedia CommonsMediaWikiMeta-

8. Implementation Description

8.1 Sprint History

The client of this project divided the requirements into large sprints of tasks. Group members allocated parts of those tasks among themselves. The task sequence and outcomes are illustrated in the next subsections.

8.1.1 Sprint 1

The first part of this sprint was for the members to be able to grasp on this totally new topic and get to know the field - and industry - they were getting into. After intensive research and multiple group meetings, members were confident enough to begin executing the next task. For the second part of this sprint, members were required to find a proper Crowdsourcing platform - a “gold mine”, as the client called it - and be comfortable with it. This task required a lot more research than the previous tasks. Members examined numerous databases for days until they came up with a new idea, which was using an Application Programming Interface instead of downloading a whole Crowdsourcing platform. Client had not mentioned APIs in the previous meeting. After further discussions among members, using APIs seemed to be more advantageous, and the work was resumed based on this critical decision. This decision was taken based on the fact that the system can use multiple APIs at a time, and therefore, have more data to be examined and analyzed. It was also obvious that using an API would require almost no space on the web host, which is a major advantage, in contrast to downloading a whole database (e.g. Wikipedia English XML database is as big as 62GB uncompressed).

Different APIs of Wikipedia, Facebook, Twitter, Reddit and Quora were examined closely and for several days. Using Quora or Reddit APIs would not have been efficient, as the companies themselves did not release official APIs. On the other hand, Facebook and Twitter had their own Developer websites that contain official APIs. Unfortunately, the Facebook API was deprecated by the company, and an alternative API was released, but the latter did not provide enough information.

Twitter API seemed more promising. It was downloaded, used and tested on the system when it was still simple. Subsequently, it was obvious that Twitter does not allow access to big amounts of information through their API. Therefore, the API would only retrieve less than 200 tweets per run. The only option members were left with was Wikipedia. A simple PHP API was integrated into the system and tested. It was chosen to be the right API as it worked perfectly.
A simple JavaScript code was written for the User Interface, which included a search bar that would search through the Wikipedia database and return huge amounts of data for minutes.

The API used in the system took a different approach. It comprised of two parts; a simple search-return functionality and a Web Crawler. The search part would return the first Wikipedia page that has the keyword searched for in its title. The Web Crawler would then visit the page and make a list of all the URLs in it (including related topics). Subsequently, the Web Crawler would return the text of each page it visited to the UI. The text would then be displayed and the Web Crawler would visit the next URL on the list. It would run in a long loop that would retrieve more information than needed.

In this first sprint, the client also aimed for the group members to be able to set up a development environment and request access to the UNMC server, as the system would mainly be hosted on said server.

8.1.2 Sprint 2

In the second formal meeting, the clients did not approve of the API system. They believed that it would be a better idea to make a system that does not rely on the internet connection's bandwidth rather than a system that works online. Clients were willing to sacrifice saving major storage space for working offline. In this sprint, tasks included the members re-doing what they had done through the previous five weeks. This time, building a system that works offline even though one of the clients' first requirements was to allocate a part on the UNMC server and have a UI as mentioned.

In the next few days, members were doing more research on the matter, and downloading a whole database - and working with it - was not looking good. Downloading the Twitter database turned out to be illegal, as one cannot keep other users' tweets for over 24 hours. However, Twitter has a Streaming API that gives real-time access to a portion of the database available for data scraping. Twitter was, once again, taken out of the equation.

Wikipedia database was downloaded, but it was too large for any well-known text editor to open. Members needed to know the information layout in order to build a search script for the database. Eventually, members decided to deal with it as a "Black Box" and find a script that already does the searching. Research took a major shift into finding a search engine for the particular database that does not rely on the local command prompt. Some tools were found (e.g. BzCat, BzGrep, etc) but no publicly documented ones were found, especially when it comes to XML parsing, which was needed since the Wikipedia database was in XML form. Professional help was seeked and, based on that, members came to the conclusion that those tools should not be used in the system, as they are actually more time consuming than the previous API approach. Let alone that some functions in those scripts were as long as 5000 lines! Development was put off for a few days for further research until the arrangement for the next meeting with the client to discuss the recent developments and future decisions.

8.2 Development Environment

During the first meeting, technologies were discussed among the members and the clients. It was obvious to only use the necessary tools rather than the personally preferred ones. The development environment of this project consists of two parts; local and public. Local development environment comprises of compilers and IDEs installed on each member's computer, as well as the open-source web servers, such as XAMPP.

The bigger portion of the public development environment is the server access granted by the university, whose IT solutions office uses the Mercury server. Access granted to the members by the IT office is considered an online web hosting service for the project. The other part of the development environment is the code sharing service (e.g. GitHub private repository of the group).

8.3 Front-end Functionality

The User Interface of this project is as simple as it can be, since the focus is mainly on the extraction and analysis of the data on the server side. The current UI uses HTML, CSS, JavaScript and PHP code, and includes a search bar and two buttons; Search and Reset. Below this horizontal alignment of objects, there is a box with transparent borders onto which all the search results will be displayed as plain text.

So far in the project, the search bar can be used for any term or phrase. It is not restricted to questions. The keyword(s) searched for is passed onto the PHP code from the HTML using the POST method. It is then used in the back-end code for the actual searching. The Search button is what triggers this operation. The Reset button deletes whatever is in the search bar the time of the click.

8.4 Back-end Functionality

As mentioned in section 7.1.1 *Sprint 1*, the back-end code of this project consists of two parts; a Wikipedia API and a Web Crawler. The overall responsibility of this code is to search for the keyword(s) in the database and return results to the UI to be displayed. More information about this matter was disclosed in section 7.1.1.

---

9. Software Testing

9.1 Testing Scheme

Testing of this system was done by the group members, specifically the leader, as the client requested in the first official meeting. The testing depended on the results being proportional to how generic the keywords searched for are. This testing was based on searching for random words and verifying whether the returned results have those keywords in common. Another criteria followed was how often the keywords searched for actually appeared in the search results. It is probably worth mentioning that the testing phase took place at multiple times throughout the development of this project, so far. A relatively sufficient number of tests were performed after every addition of code to the project in general.

9.2 Testing Phase Expectations

Prior to carrying out the testing phase, expectations included the search results being more precise. This made more sense to the group members as it was a matter of testing rather than prediction to know how bulky and colossal the database was. It was also anticipated that the system would get to the point when it comes to results, as it is usually the case with search engines.

9.3 Results

The trade-off of the results being able to match the expectations is a major factor when it comes to proper software development. After the testing had been completely gone through, the test results were reviewed closely. Those results illustrated that the system returns bulky information. Huge amounts of data get returned for even keywords that have the narrowest field of information. Furthermore, it is somewhat obvious for the system to return irrelevant information if it returns tremendous amounts of data, which was definitely the case.

10. Discussion

As the purpose of this project is relatively futuristic, anticipating that the aims have been fulfilled after just finishing the first stage is incomprehensible. According to the client, group members are on the right track and the pace they are moving at is proficiently suitable. Since the purpose of this project is to collect data from people far away in the world and use that data to fill up the gaps in the Crops For The Future database, having huge amounts of information in the database of this project is ultimately beneficial. This explains why the client is confident with the current development plan.

Nevertheless, it is rather justifiable that the results acquired through testing did not match the expectations set by the group members. Since this system works somewhat like an online real-time viewer, it is more digestible for it to explicitly have more results than expected. This explains the nullification of the first expectation which specified that the data would be more precise. On the other hand, prior expectations found it more likely for the system to return to-the-point results. However, the system returns some irrelevant text within the contextualized results. This is merely because the Web Crawler facilitated in this system returns to the UI whatever is enclosed within the <*body*> HTML tags. This definitely includes irrelevant text (e.g. text related to buttons/drop-down menus).

Even though tremendous amounts of data would include irrelevant pieces of information, this works for the project's benefit. As opposed to what one would think, it is benevolent for this project to have more data than needed. At the end of the day, it is a Crowdsourcing project, and Crowdsourcing relies on having lots of data to start with. In fact, having too little information on some matter would have a counter-effect on the outcomes, as it would bring the system to finalize search operations after having searched through insufficient amounts of data. With Crowdsourcing, available information must be examined rather than generalized upon full acquisition.

Reasonably enough, group members faced some obstacles throughout the development of this first major stage of the project. One of the first difficulties was dealing with clients and blending into the workplace when needed. In the real world, client-based projects are run by a hierarchy of people based on their roles. This is rather strenuous to deal with at first. Some of the group members had some experience working in a real workplace before, so this issue was quickly overcome. On the other hand, introducing one's self to new concepts while being responsible for any major decisions related to those particular concepts is an ongoing difficulty. This is being overcome by extensive research, but it is justifiable if it remains for as long as the project takes. At the end of the day, all parties involved in this project have one mutual goal. Therefore, all the problems encountered along the way will be resolved within.

Although it is, in some measure, early in the development to come to this conclusion, scarcity of related Crowdsourcing platforms and resources made this project a whole lot more challenging. It is probably worthwhile for more people to get involved in Crowdsourcing in the future. Not just because Data Science is growing fast, but because nobody knows the importance of Crowdsourcing except those who get their hands dirty with it. **[10]**

11. Time Plan

11.1 Work Hours Table

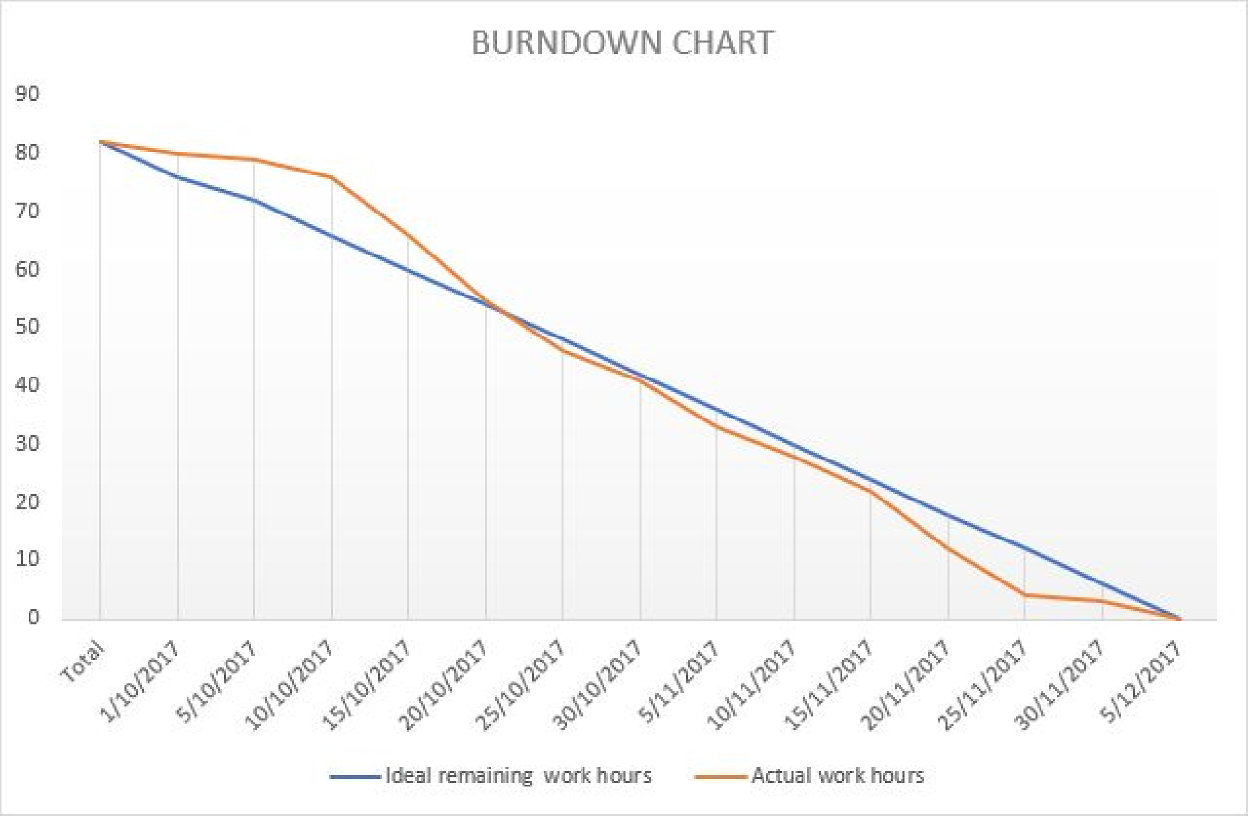
Following is a table that includes most of the tasks group members have worked on and, therefore, achieved. For each task, start and end dates are included as well as an estimate of the hours spent on the particular task during that period of time.

NO	TASK	DURATION	START DATE	END DATE
1	Research on Data Science	8 hours	1/10/2017	12/10/2017
2	Reviewing requirements	6 hours	14/10/2017	18/10/2017
3	Finding a Crowdsourcing platform	5 hours	14/10/2017	16/10/2017
4	Setting up development environment	2 hours	16/10/2017	17/10/2017
5	Setting up server	2 hours	16/10/2017	22/10/2017
6	Further requirement review	4 hours	16/10/2017	20/10/2017
7	Developing the UI	2 hours	16/10/2017	19/10/2017
8	Finding proper APIs	11 hours	20/10/2017	1/11/2017

9	Connecting the parts together	2 hours	1/11/2017	4/11/2017
10	Uploading the work on the server	3 hours	5/11/2017	8/11/2017
11	Further research on downloadable DBs	7 hours	10/11/2017	21/11/2017
12	Research on API drawbacks	6 hours	10/11/2017	21/11/2017
13	Writing the interim report	18 hours	10/11/2017	28/11/2017
14	Testing	3 hours	17/11/2017	23/11/2017
15	Proof-reading the interim report	3 hours	30/11/2017	3/12/2017
TOTAL		82 HOURS		

11.2 Burndown Chart

The following Burndown chart relates to the table in the previous section. As illustrated in the chart, the team started becoming ahead of schedule after the first meeting with the client took place (i.e. mid-October).



12. Wrapping Up

Crowdsourcing is a relatively new field in the industry. The project enclosed within the plies of this report participates in helping this field of study grow by applying it to solve real problems that the client's industry is facing; lack of information coming from far away places. The client wants to fill up the gaps in their database through having a system that can be asked text-based, agriculture-related questions and answer those questions based on the analysis of data acquired from the huge database integrated into the system. This project is moving at the right pace, as the client suggested. Since the test results were far more informative than expected, the current database should be enough to start with. Even though the client has not fully approved of it, it is going to be discussed further and, accordingly, changed, improved or kept as is.

13. References

1. A. J. G. Hey, S. Tansley, and K. M. Tolle, *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, 2009.

2. A. Halevy, P. Norvig and F. Pereira, "The Unreasonable Effectiveness of Data," in *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8-12, March-April 2009.

3. “Why data science Is the fastest growing industry in tech right now,” Import.io, 30-Mar-2017. [Online]. Available: https://www.import.io/post/why-data-science-is-the-fastest-growing-industry-in-tech-right-now/. [Accessed: 01-Nov-2017].

4. J. Howe, “The Rise of Crowdsourcing,” Wired, 01-Jun-2006. [Online]. Available: https://www.wired.com/2006/06/crowds/. [Accessed: 30-Oct-2017].

5. R. Buettner, *A Systematic Literature Review of Crowdsourcing Research from a Human Resource Management Perspective*. 2015.

6. A. Taeihagh, “Crowdsourcing, Sharing Economies and Development,” *J. Dev. Soc.*, vol. 33, no. 2, pp. 191–222, Jun. 2017.

7. “IBM Analytics,” The Quant Crunch, 17-Mar-2016. [Online]. Available: https://www.ibm.com/analytics/us/en/technology/data-science/quant-crunch.html. [Accessed: 01-Nov-2017].

8. Rasnacis, A., & Berzisa, S. (2017). Method for Adaptation and Implementation of Agile Project Management Methodology. *Procedia Computer Science*, *104*, 43–50. https://doi.org/10.1016/J.PROCS.2017.01.055

9. Fojtik, R. (2011). Extreme Programming in development of specific software. *Procedia Computer Science*, *3*, 1464–1468. https://doi.org/10.1016/J.PROCS.2011.01.032

10. Dichev, C., & Dicheva, D. (2017). Towards Data Science Literacy. *Procedia Computer Science*, *108*, 2151–2160. https://doi.org/10.1016/J.PROCS.2017.05.240

Appendix A: Meetings Overview and Minutes

1	Date: 13/10/2017	Type: Formal	Attendees: Group members, clients and the supervisor

<u>- Objective:</u>

Attaining specified project information and requirements.

<u>- Discussion:</u>

After having an informative introduction about the clients and their field of work, the project was tackled from different angles. Here are the key points:

- CropBASE has a huge database that consists of 400 tables described by over 4000 variables. Each variable has Metadata (data describing collection of other data). We can store about 404 million pieces of data about one crop. There is an interactive tool that is connected directly to the tables inside that database. This tool works for extracting information from the database and analyzing and visualizing it in a more comprehensive way. The tool is capable of getting all the needed data from the tables to build a visualization of crops and their information in any place in the world. That information includes change of the quality of a crop based on climate change, soil details, advice about the best timing to start growing a particular crop, etc. A tool that is still under development is bound to expect how climate change will affect that crop in the future.
- Since the database is huge, data cannot all be obtained manually. Another challenge is using that huge amount of data.
- There are several online tools (Google AdWords, Wikipedia, Twitter, etc.) that can be used to navigate Crowdsourcing data to something that utilizes it. The first task set is to know where this Crowdsourcing data is available and what are the best tools for extracting that data for later uses. This tool is presumed to be the goldmine of this project.
- The project has been done before, but it was incomplete. Therefore, this project is considered a solution to a problem that CFF has, which is the need to broaden their database through automated algorithms facilitating Crowdsourcing.
- The previous implementation used Google AdWords, but that tool has blocked unpaid memberships.
- The data lifecycle from the Crowdsource to the database: extract, analyze, find patterns, verify, curate, visualize and submit.
- A problem with Crowdsourcing is that it includes millions of irrelevant pieces of data that are not useful to the industry. Through that huge amount of data, the

algorithm should be able to answer defined questions (preferably ten questions). Questions like “Do people like plant X? In what country?”.

- One of the goals is to get useful information from people far away in different continents (e.g. African farmers) on certain crops and how their utilization works. Manually this is not possible, but through automated algorithms of current technologies, it probably will be.
- It might be a necessity to translate data from other languages into English to collect as much data as possible.
- The system should be able to keep up with the trends and answer questions based on that. E.g. “Why is this particular crop being discussed more in April than in August?”.
- We are not meant to write a library about how to distinguish a topic among millions of topics and extract it. There are libraries that already exist and have that same functionality.
- The second task is setting a server for the environment. The server will be accessed by everyone involved in the development of this project. Clients should be given a URL to the interface of the database connected to the server.
- We shall be granted access to UNMC’s server (Mercury) in order to upload the database on it. This is a part of the second task.
- The database will be built using MySQL. The database must be connected to a Web Server like Apache or TomCat. We might not have to use a Web Server in case UNMC’s server is capable of providing Apache’s functionalities, for example.
- Get an account on Mercury and make it work. Upload the database and the website on it. Once we are able to use it comfortably, we shall proceed to upgrading the connection by linking those parts through UNMC’s server instead.
- Use XAMPP, LAMPP or MAMPP for establishing the connection configuration of the development environment, including the server connection.
- After choosing the data source, we should be able to extract the whole database and connect it to the website allocated through the server hosting on UNMC.
- Communication with the client will be through Slack and Doodle, and those channels should be emailed to the client once established. Email and Google Docs are other means for communication when it comes to this project.
- Clients are willing to share directories on Google Docs with group members, including the description of the project and tasks. Through those online Docs we should, as members, put our names and contact details and update the Minutes of every meeting right after the meeting is conducted. Minutes should include attendees, place, time, discussion and action points. This is to make sure that everyone is clear on the requirements. It is also useful for peer assessment.

- Whenever an email is being sent, both clients and supervisors must be CCd in order to avoid confusion and to keep things as clear to everyone as possible.
- The focus is on the back-end rather than the front-end. The system is a user-access system, not a public-access one.
- The most important feature is the connection between what is trending and the data that is meant to be extracted. Livefeed is very much prefered to be utilized.
- Technologies are to be updated accordingly as we proceed with the project. Expected needed programming languages are: PHP, MySQL, R, Python, HTML, CSS, JavaScript and possibly MATLAB for the analysis.
- Data analysis is going to be done through libraries that might be specified by the client in the future.
- The most important step is bringing all the parts together into one system rather than focusing on data analysis which has already been done. The plan is to use what has been done and enhance it to work for our purposes and bring us the best outcome eventually.
- The ultimate goal of CFF is to achieve the United Nations Sustainable Development Goals. The current project actually fits into the domain of UN SDGs.
- It is very important to have a working interface for the project. However, the focus is mostly on data extraction and analysis.
- Testing will be done by the team (self-assigned testing). The clients will not be doing any tests except for the final run. There will be no human involvement in the testing stage. However, clients will be validating the tests run by the team members afterwards.
- Clients are willing to follow the timeline set by the university.
- The methodology used is dividing the project into tasks (milestones) set by the clients. The team is meant to finish the task within the specified timeframe.
- Meetings are to be held on a monthly basis, and tasks shall be updated accordingly.
- There are no projected competitions with other facilities in the world. If any are found, it would be better to make use of the creativity utilized in those tools rather than building an improved version of them.
- The project will be considered a success if it answers all the predefined questions asked.
- Crowdsourcing is based on public data rather than personalized data. Therefore there are no ethical overlaps with the ongoing process of development.

<u>- Conclusion:</u>

Members were given 3 weeks to accomplish two tasks:
1- Finding a Crowdsourcing platform that is both reliable and informative.

<table>
<tr><td></td><td colspan="3">2- Setting up the development environment.</td></tr>
<tr><td>2</td><td><b>Date:
16/10/2017</b></td><td><b>Type:
Informal</b></td><td><b>Attendees:
Group members</b></td></tr>
<tr><td></td><td colspan="3"><b><u>- Objectives:</u></b>
1- Exchanging information after intensive research.
2- Finding the “goldmine” of Crowdsourcing.
3- Setting-up the development environment.
4- Requesting for the server hosting.

<b><u>- Discussion:</u></b>
The meeting focused on choosing a “gold mine” as our database, as well as implementing the server. The key points are as follows:
● Client gave the task to find a “gold mine” and obtain a database. The meeting discussed the advantage for using a database which is the simplicity of obtaining information and less programming required to make it work. However the main disadvantage comes from the fact that a huge database, such as Wikipedia’s 40-45 GB worth of data, would be outdated as soon as it is retrieved, plus the huge database would be incompatible with the server that will be used. The solution proposed was to use an API of the database to save storage and be more efficient overall. The programming work may be more demanding than a traditional database approach, but it was decided the change in the approach would be worth it.
● When choosing an API, the website options were narrowed down to Wikipedia, Facebook, Twitter, Reddit, Quora, and various others. It was pointed out that each API uses different programming languages, but it should not matter since the data will be converted to JSON anyway, which will in turn be used in the PHP side of programming.
● Work was assigned to each member to search for a website API, build the user interface for the prototype, and to set up the hosting of the test server.

<b><u>- Conclusion:</u></b>
The database approach will be moved aside in favor of an API. Possible “gold mines” were identified and will be studied further.</td></tr>
<tr><td>3</td><td><b>Date:
26/10/2017</b></td><td><b>Type:
Informal</b></td><td><b>Attendees:
Group members</b></td></tr>
<tr><td></td><td colspan="3"><b><u>- Objectives:</u></b>
1 - Reviewing progress made in searching for APIs of various websites.</td></tr>
</table>

2 - Discussing the workload for the group report.

- Discussion:

A review of the progress made over the last ten days. The key points are as follows:

- While the server setup continues to be delayed, it was decided that work would begin for the interim group report rather than putting it off until nearing the date of the submission. Because of the early start, only the first half of the group report can actually be discussed, but it was decided that instead of waiting for the server setup, idle time would be used to work on what can be done now, which led to the proposal to start writing the group report.
- The Wikipedia API is able to deliver a sufficient amount of results as a replacement for the database, but the content is not organized in any way to make out any useful information. The Twitter API does not output as many results, so the idea of using Twitter as the “gold mine” were set aside. If implementation of one API works smoothly, there may be plans to further implement multiple APIs to increase the size of the gold mine, leading to more data to analyze.
- User interface is implemented. Although very simple in functionality, creates the premise for the prototype that will be built. The UI will be added to the group Github repository.

- Conclusion:

Work will begin for the group report while waiting for the server. Search for the API “gold mine” continues.

4	**Date: 09/11/2017**	**Type: Formal**	**Attendees: Group members and client**

- Objectives:

1- Discussing database options found.
2- Discussing actions taken for the APIs.
3- Presenting the new User Interface.
4- Receiving the next sprint of the backlog/tasks.

- Discussion:

The points and issues discussed were as follows:

- Client suggested that using APIs over downloadable databases as a Crowdsourcing platform is partially a correct way, but it is not the right approach. When it comes to efficiency, using a downloaded database and processing its data locally is the best choice, as it does not rely on how fast the internet is. Moreover, APIs change and a lot of them get deprecated after a while. We do not want these kinds of problems in our system.

- Some databases might look insufficient, but when it comes to reality, they might be the best fit for the project. Assuming and concluding that a database will not work without testing it in the existing system is a naive mistake.
- The problem of downloaded databases being out-dated can be solved using some tools that enable system administrators to schedule “update jobs” for the database to be updated accordingly and as frequently as needed.
- A good way to determine whether a database is sufficient for the project is categorizing the databases found into ‘relevant’, ‘somewhat relevant’, and ‘irrelevant’ and documenting the reasons behind the decisions taken.
- The project deliverables were revised on. Client emphasized on the system being dynamically able to answer questions based on the data it withholds and analyzes. The system depends on what is called Natural Language Processing which is based on text classification algorithms. There is an API, made by Google, that provides this capability.
- Client prefers not to provide members with the questions that the system is going to be asked. Otherwise, programmers might follow the method of answering those particular questions in code rather than the system being able to answer on its own.
- One of the members should be allocated to object to the approach followed (use of APIs). This person is responsible for proving everyone wrong as a way of finding the best solution to current and future problems. It is preferred that this role assignment will rotate throughout the implementation of the project.

- Conclusion:

Client suggested that the API approach might be invalid. Members were given 3 weeks to find a new Crowdsourcing platform with a downloadable database (preferably Twitter).