**COMSATS University Islamabad,   
Park Road, Chak Shahzad, Islamabad Pakistan**

Project Proposal  
(SCOPE DOCUMENT)

for

**Data Science Toolkit for Social Scientists**  
Version 1.0

***By***

**Umer Khan Wazir CIIT/SP17-BCS-098/ISB**

**Yusra Fatima CIIT/SP17-BCS-063 /ISB**

***Supervisor***  
**Mr. Qasim Malik**

*Bachelor of Science in Computer Science (2017-2021)*

|  |  |  |
| --- | --- | --- |
| **No.** | **Comment** | **Action** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**SCOPE DOCUMENT REVSION HISTORY**

**Supervisor Signature**

**Date:**

**Table of Contents**

1. Introduction 1

2. Problem Statement 1

3. Problem Solution for Proposed System 1

4. Related System Analysis/Literature Review 2

5. Advantages/Benefits of Proposed System 2

6. Scope 3

7. Modules 3

7.1 Module 1: User Management 3

7.2 Module 2: Data Collection and Pre Processing 3

7.3 Module 3: Development and Integration of Data Science Tools……………………………….4

7.4 Module 4: Report……………………………………………………………………………….4

7.5 Module 5: Data Repository Management and Sharing…………………………………………5

7.6 Module 6: Application to the Data on British Rule in South Asia………………………...5

8. System Limitations/Constraints 5

9. Software Process Methodology 5

10. Tools and Technologies 6

11. Project Stakeholders and Roles 6

12. Team Members Individual Tasks/Work Division 7

13. Data Gathering Approach 7

14. Concepts 7

[14.1 Concept-1: Web Development: 7](#_Toc20782168)

[14.2 Concept-2: Optical Character Recognition: 8](#_Toc20782169)

[14.3 Concept-3: Data Science Tools: 8](#_Toc20782170)

15. Gantt chart 8

16. Mockups 9

17. Conclusion 10

18. References 10

19. Plaragism Report 10

**Project Category:**

* **B**- Web Application.
* **C-** Problem Solving and Artificial Intelligence.
* **H-** Image Processing.

# Abstract

In today’s world, Social Scientists like Historians etc. requires information on which they perform analysis and get benefit from it. But the problem is that they do not have enough time in today’s busy world to extract information from complex and detailed documents like research papers, articles, old newspapers, books, thesis, magazines etc.

In our project, we are creating website for Social Scientists in which they will be provided data science toolkit to extract useful information from documents like providing summary of long documents, assigning topics to stack of long documents, providing graph of percentage of different emotions (happy, sad, angry, positive, negative etc.) expressed in documents and many more. Users will be allowed to upload a document (pdf, doc, scanned images, web scrapping, epub, mobi, divu) on which data science tools will be applied to first convert document into editable form and then extract useful information from document. System can convert any type of mentioned documents into editable form. Data cleaning tools will also be provided so that user can clean document before implementing Data Science Tools. User can store results of extracted information from documents in report section and the history of uploaded document will also be maintained in repository so that user can access it at any time. User can also share their repository other users. The profile of users will be maintained and users can also search other users and view their repository in feed section by following them. The main purpose of our website is making information easily available for Social Scientists.

Data Science tools like Topic Modeling, Data Mining, Text Summarization, Sentiment Analysis etc. will be applied on the document uploaded by user to extract useful information from document that user do not need to waste time on reading whole document.

# Introduction

Ninety percent of the data in the world today has been created only in the last two years, according to IBM. With the increase of mobile devices, social media networks, and the sharing of digital photos and videos, we are continuing to grow the world’s data at an astounding pace. Data is big and getting bigger. Data has always been used to develop high-level metrics and business intelligence. Smart organizations have long relied on data to help make strategic business decisions. The information that was previously ignored because there was no reasonable way to process it. As Data Science Tools are available now to extract information from data, we will use that to make information easily available for Social Scientists.

This project is related to Data Science and Web development. First, we will create a website using our web development skills. Then using our Data Science skills, we will create a data science toolkit, which will be available for users of our website. After uploading document which can be of many formats, system will convert document into editable form. User will be also provided data cleaning tools so that they can clean document before implementing data science tools. Now user can get information from that document very easily and in less time because of data science tools available in toolkit. We will create user-friendly environment for users, so it easy for them to use our website.

# Problem Statement

In today busy world, Social Scientists do not have enough time to read the whole research paper and other documents in one day, and the documents also contain details in which reader might not be interested. Reader may just require the specific information from document which might me hard to find in detailed documents. They may also want to compare multiple documents and also want to know about content of document in short time. User may also want to know about emotions expressed in documents which is very hard to know in long document in matter of seconds.

Social Scientists especially the Historians are very much interested in analyzing historical documents from various sources in order to generate lesson-oriented concrete and objective analysis of the past. Such documents, however, contain mostly the textual data and can be in various formats (e.g. pdf, doc, scanned images, web scrapping, epub, mobi, divu). The textual nature of document and their divergent formats make it quite tedious for social scientists to highlight or de-emphasize certain angles in order to generate the analysis in their own line of understanding.

System is required in which user can upload its document and just get the necessary information for which he/she is looking.

# Problem Solution for Proposed System

This project aims to develop a website in which data science toolkit will be available for Social Scientists after he/she creates his/her account, that allows them to feed documents of various formats like pdf, doc, scanned images, web scrapping, epub, mobi, divu, and provides them with a useful set of data science tools (Topic Modeling, Data Mining, Text Summarization, Sentiment Analysis etc.) to quickly extract useful information that in turn can be used to form their analysis or to uncover newer insights. Before applying Data Science Tools, Document of all mentioned types will be converted into editable form by system, so that user can update document or delete portion from it before implementing tools. User will be also provided data cleaning tools so that they can clean document before implementing data science tools. So now user do not need to read the whole document but the system will help them to get the required information in no time. The report of document will be provided to user. User can maintain history of the documents uploaded and access it at any time. User can maintain their repository and share their content with other users. Users can also search other users and follow their account so they can view their repository in feed section.

# Related System Analysis/Literature Review

Below is the application that is related to our system:

|  |  |  |
| --- | --- | --- |
| **Application Name** | **Application Description** | **Weakness** |
| Data Science Toolkit | * This is a website which contain data science tools. * Text Summarization and sentiment analysis can be implemented on documents. | * Text Summarization feature does not work. * Sentiment Analysis feature only gives result as positive and negative. * Interface is simple and not good. |
| Resoomer | * This is website which implement text summarization on documents. * User paste text into textbox and system provides summary of document in result. | * It does not have upload document features. * User can only copy and paste text. * This website is only for text summarization. * There is no repository to maintain results. |

Table 1: Related System Analysis

# Advantages/Benefits of Proposed System

System will help users to:

* Extract the required information easily from uploaded document.
* Form their analysis and uncover new insights easily because of user-friendly environment.
* Not read the whole document because of data science tools available.
* Easy to manage and access.
* View other people content and learn from their content.

# Scope

Extraction of useful information from detailed and complex data is time consuming as we are surrounded by a lot of data these days. Our system will help user solve this problem by getting required information in no time by creating user-friendly environment. This system is especially for Social Scientists, who are always looking for useful information to perform analysis on it and get benefits from it. User of application can create his account on website. After that user will be able to upload the document of various types like pdf, doc, scanned images, web scrapping, epub, mobi, divu, from which they want to extract useful information. After uploading the document, using OCR techniques system will convert all of the above-mentioned documents into editable form for users. User can now perform update or delete operation on document before applying Data Science Tools. User will be also provided data cleaning tools so that they can clean document before implementing data science tools. After processing the data, Data Science Toolkit will provide user many options like text summarization, topic modeling, sentiment analysis and many other to perform on uploaded document. Using toolkit, system will provide user summary of uploaded long documents, user can upload multiple documents of various types and system will categorize documents into different topics in short time, system can determine whether writer attitude towards particular topic, product etc. is positive, negative or neutral, system will display graph of percentage of different emotions like happy, sad, angry, positive, negative expressed in documents. User can view the important sentences highlighted in document; user can also go to topics in documents in which he/she is interested. System will provide results of tools implemented on documents. User can store results in Report section of website. User can keep history of document, so he/she can access it any time. User can maintain repository and share their content with other users. User can follow other users and view their repositories in feed section. User have to create and manage his/her account before accessing Data Science Toolkit.

# Modules

The modules of our application are explained below:

## Module 1: User Management

This module deals with the user’s accounts. Before using system, user have to create account. User have to add his personal information like name, phone number, email address, location etc. to create his/her account. After signing, user can login to the system by entering personal information (email/phone) and password. Users profiles are also maintained. Users can search other users account and view their profiles and follow them.

## Module 2: Data Collection and Pre-Processing

This module allows user to upload the document on which he/she want to implement data science toolkit. Then OCR will be used to convert different documents like pdf, doc, scanned images, web scrapping, epub, mobi, divu into machine-encoded text. The document will be converted into editable form so that user can perform update and delete operations on document before implementing data science tools. User will be also provided data cleaning tools so that they can clean document before implementing data science tools. To prepare document for implementing tools on it, following are steps to do so:

### Module 2.1: Providing Editable Document

Social Scientists are provided option of uploading document which can be pdf, doc, scanned images, web scrapping, epub, mobi, divu. After that document will be converted into editable form, so user can change document content and perform delete and update operations on document. After user is satisfied with content of document, he/she can now apply data science tools that are provided.

### Module 2.2: Loading and Cleaning Data

After converting document into editable form, system will provide user options about document like normalize the case to lowercase, uppercase and proper case and remove any words that have non-alphabetic characters. Stemming and Lemmatization will also be applied in topic modeling. Putting this all together, we will create a function that will take a list of lines of text and return list of clean lines of text after user instructs system to apply this function.

## Module 3: Development and Integration of Data Science Tools

This module allows the user to have multiple options that they want to implement on document. User can search for summary of document, user can also upload multiple documents of various types and system will categorize documents into different topics in short time, system can determine whether writer attitude towards particular topic, product etc. is positive, negative or neutral, system will display graph of percentage of different emotions like happy, sad, angry, positive, negative expressed in documents.. For that many machine learning techniques will be used. Techniques that will be implemented on document are given below:

### Module 3.1: Text Summarization

This module allows user to upload document of various types like pdf, doc, scanned images, web scrapping, epub, mobi, divu and first convert it into editable form so that user can update and edit text in document. User can also copy and paste text into textbox to implement tools. Then using data cleaning tools user can also clean the document before using data science tools. Then text summarization will be used to display summary of long documents of various types. System will also allow user to manually select number of sentences in summary. User can also view important sentences in documents highlighted by system.

### Module 3.2: Topic Modeling

This module allows user to upload multiple long documents in system of various types. System will first convert all of them into editable form and then user will be displayed the topics generated by system which are extracted from uploaded document. Then stemming and lemmatization will also be used in topics. User can then view documents linked to each topic and the percentage to connection between topic and document. One document can have many topics.

### Module 3.1: Sentiment Analysis

This module allows user to upload document of various types like pdf, doc, scanned images, web scrapping, epub, mobi, divu and first convert it into editable form so that user can update and edit text in document. User can also copy and paste text into textbox to implement tools. Then using data cleaning tools user can also clean the document before using data science tools. Then sentiment analysis will be used to display percentage of different emotions like angry, happy, sad, positive, negative etc. expressed in uploaded document.

## Module 4: Report

This module allows user to get the report of the information they want to extract from the uploaded document. The results of all the data science tools implemented on documents will be saved in Report section of website. The reports of each document uploaded will be kept in database. So that when user login again he/she can access it again. User can also add his/her analysis of document and add description of document in report. User can also download the report and save it on computer.

## Module 5: Data Repository Management and Sharing

This module allows user to save all the documents they uploaded in repository. In repository user can easily access their documents as different portions will be made for documents to store. User can also share their documents with other users, so that other users can view documents they uploaded.

### Module 5.1: Data Repository

This module allows user to upload documents from computer and store them in repository section of website. User can also pick documents from repository section and implement data science tools on it. User can also view documents in repository section and also download document from there.

### Module 5.2: Sharing

This module allows user to share their whole repository with other users. Feed section is provided by system where user can view all the repositories shared by other users. User can follow other users account by searching their profile in search bar and after following them they can view their content in feed section. User can also view other users repository in repository section of users account.

## Module 6: Application to the Data on British Rule in South Asia

In the end, the usefulness of the development Data Science Toolkit will be demonstrated by applying it to researched textual data of British rule in South Asia from British and South Asian perspectives. After we get conformation that our Toolkit is working properly, we might launch our product in the market because it will be very useful for Social Scientists in their daily work routine.

# System Limitations/Constraints

Following are the constraints/limitations of our system:

* Document should be clear to read otherwise system would not be able to detect and read the document.
* Its preferred to have a computer typed document rather than hand written.
* User can use system only when internet is available because its web-based project.

# Software Process Methodology

The software process methodology used will be Incremental Process Methodology as it shows the progress at every state and is very easy to use and understand. Incremental Methodology is convenient for the user and the client too. It not only helps in detecting the issues in an initial state but also provides the flexibility to change at every stage.

# Tools and Technologies

Tools and Technologies used in developing this system are as follows:

Table 2 : Tools and Technologies for Proposed Project

|  |  |  |  |
| --- | --- | --- | --- |
| **Tools**  **And**  **Technologies** | **Tools** | **Version** | **Rationale** |
| Mongo dB | 4.0 | DBMS |
| OpenCV | 3 | Machine Learning |
| TensorFlow | 2 | Machine Learning |
| Bootstrap | 4 | Front-End  Web Development |
| node.js |  | Back-End  Web development |
| MS Word | 2016 | Documentation |
| MS Excel | 2016 | Presentation |
| Visual Studio Code | 16.4 | IDE |
| Moqups | Latest | Designing |
| **Technology** | **Version** | **Rationale** |
| Html | 5 | Language |
| CSS | 4 | Language |
| Javascript | 1.8 | Programming Language |
|  | Python | 2.8.1 | Programming Language |

# Project Stakeholders and Roles

Following are the Project Stake Holders of the system:

Table 3 Project Stakeholders for Proposed Project

|  |  |
| --- | --- |
| **Project Sponsor** | COMSATS University, Islamabad |
| **Stakeholder** | Umer Khan Wazir (SP17-BCS-098)  Yusra Fatima (SP17-BCS-063)  Project Supervisor Name: Mr. Qasim Malik  Final Year Project Committee: Evaluation of project |

# Team Members Individual Tasks/Work Division

Following are the team members and task division of the system:

Table 4 Team Member Work Division for Proposed Project

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student Registration Number** | **Responsibility/ Modules** |
| Umer Khan Wazir  Yusra Fatima | SP17-BCS-098  SP17-BCS-063 | Umer Khan Wazir (Module 3 - 4 - 6)  Development and Integration of Data Science Tools, Reporting, Application to the Data on British Rule in South Asia  Yusra Fatima (Module 1 - 2 - 5)  User Management, Data Collection and Pre-Processing, Data Repository Management and Sharing |

# Data Gathering Approach

The majority the data was collected from the following techniques:

**Interviews:** The interviews are a real experience to know the needs, demands and trends of the society and this was the reason of conducting interviews.

**Surveys:** Surveys were conducted to experience such problems of a society considering the condition and the solution of it too.

**Questionnaires:** Questionnaires were a source too as they covered a large amount of audience and provided the best yet quick results.

# Concepts

Following are the concepts used in my system:

## Concept-1: Web Development

The concepts that we will learn and data science tools that we implement in our system, all that will be web based. We will create website by using our web development skills on which we will run our system. We will learn how websites are build and how system will be linked to it.

## Concept-2: Optical Character Recognition

Optical Character Recognition techniques will be used to convert document which can be of various types like csv, pdf, scanned images, link to document etc. into editable form for user. After converting document into machine encoded text, user can perform deleting and updating operation on document before implementing Data Science Tools. While building our system, we will learn OCR techniques and concepts.

## Concept-3: Data Science Tool

While building our system, Data Science Toolkit will be used in which different tools like text summarization, topic modeling, sentiment analysis etc. will be implemented on document after converting it into machine encoded text. It will extract useful information from detailed document in no time. Now user can get information from document without wasting time on reading detailed documents. System will provide users report of document in which all information will be available. User will maintain history of documents so he/she can access it later. We will learn data science concepts and how to implement those concepts.

# Gantt chart

The Gantt Chart of Application is as follows:

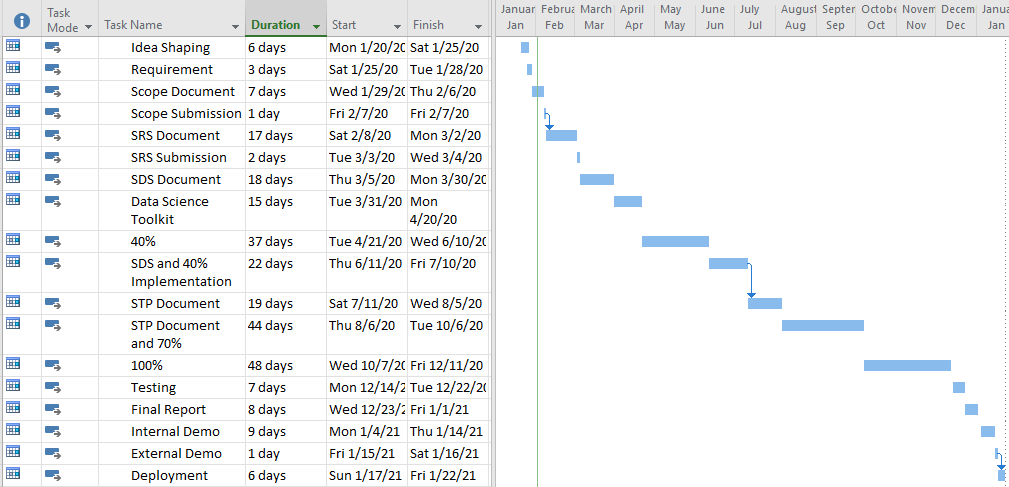
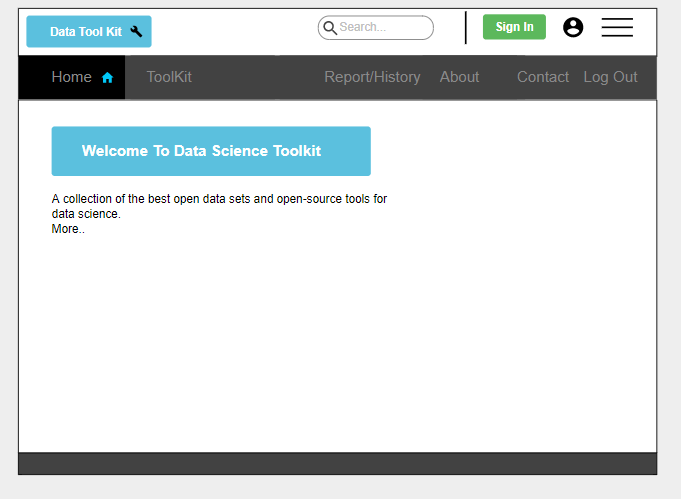
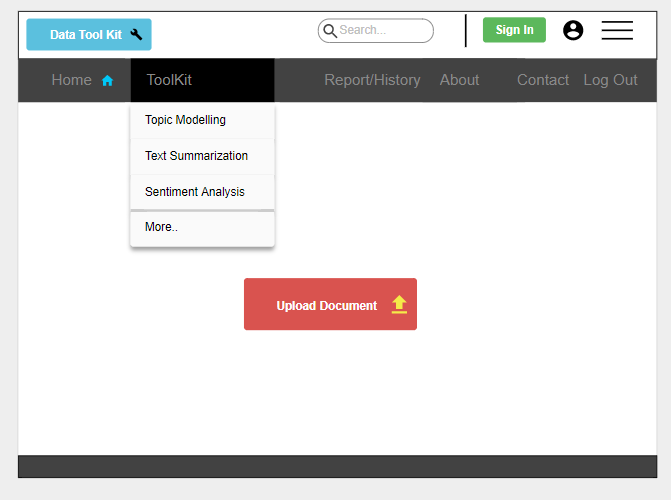


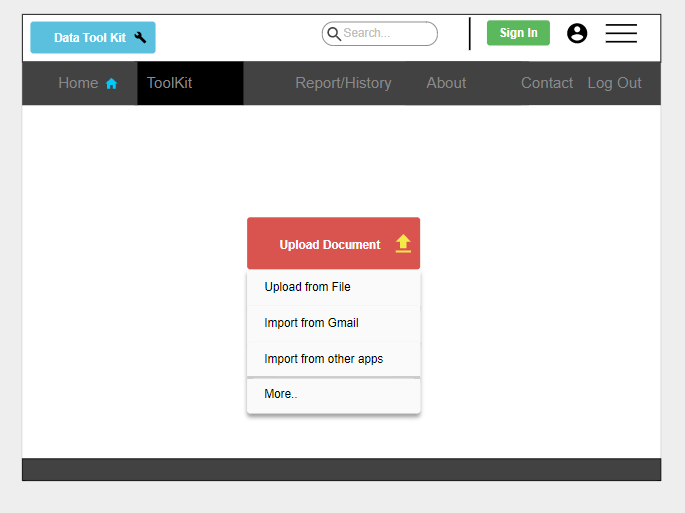
Figure 1Sample Gantt chart

Mockups:

Following are the mockups of our Application:



**Mockup 1: Home Page Mockup 2: Toolkit**

****

**Mockup 3: Upload Document**

# Conclusion

Our system will not only provide users tools to convert different documents into editable form but it will also allow them to implement many techniques like text summarization, topic modeling, sentiment analysis etc. to save their precious time and take document reading to whole new level. System will also provide data cleaning tools, so that user can clean data before implementing data science tools.

# References

**Web links:**

* <http://www.datasciencetoolkit.org/>
* <https://resoomer.com/en/>

**Book links:**

**Big Data Glossary A Guide to Public Data, First Edition, Pete Warden.**

* https://www.amazon.com/Big-Data-Glossary-Guide-Generation/dp/1449314597

**Data Source Handbook A Guide to the New Generation of Data Tools, First edition, Pete Warden.**

* <https://www.amazon.com/Data-Source-Handbook-Guide-Public/dp/1449303145>

# Plagiarism Report

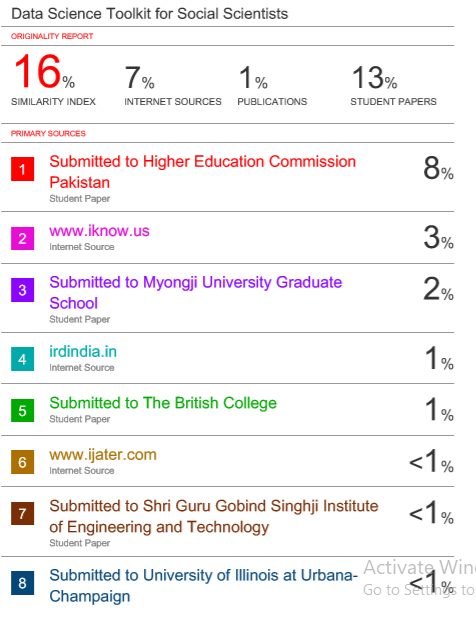
Following is the plagiarism report of the document:

Figure 2: Plagiarism Report