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

Project Proposal  
(SCOPE DOCUMENT)

for

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

***By***

**Student Name 1 CIIT/SP17-BCS-098/ISB**

**Student Name 2 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 2

2. Problem Statement 2

3. Problem Solution for Proposed System 2

4. Related System Analysis/Literature Review 2

5. Advantages/Benefits of Proposed System 2

6. Scope 3

7. Modules 3

7.1 Module 1: Module Name 3

7.2 Module 2: Module Name 3

8. System Limitations/Constraints 3

9. Software Process Methodology 3

10. Tools and Technologies 4

11. Project Stakeholders and Roles 4

12. Team Members Individual Tasks/Work Division 4

13. Data Gathering Approach 5

14. Concepts 5

15. Gantt chart 5

16. Mockups 6

17. Conclusion 6

18. References 6

19. Plaragism Report 6

**Project Category:**

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

# Abstract

In modern world, the amount of data we produce every day is truly mind-boggling. There are [2.5 quintillion bytes of data](https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1) created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated.

In our project, we are creating website especially for social scientists in which they will be provided data science toolkit. Users will be allowed to upload a document on which data science tools will be applied to extract useful information from document. The main purpose of our website is make information easily available for users from uploaded document.

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. But the power and allure of Big Data is how it enables organizations to leverage unconventional data points: the information that was previously ignored because there was no reasonable way to process it.

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, 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, people do not have enough time to read the whole research paper 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 big documents.

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, csv, scanned images etc). 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 users after he/she creates his/her account, that allows them to feed document of various formats and provides them with a useful set of data science tool (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. 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.

# Related System Analysis/Literature Review

*None*

# 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

# Scope

Extraction of useful information is time consuming as we are surrounded by a lot of data. Our system will help user solve this problem by getting required information in no time by creating user-friendly environment. User of application can create his account on website. After that user will be able to upload the document from which they want to extract information. After uploading the document it will be recognized by system using image recognition techniques and the content of document will be read by system through natural language processing techniques. After processing the data, system will implement many machine learning techniques like text summarization, topic modeling and many other techniques to provide user an environment in which he/she can get the required information in no time. System will provide user document summary, user can go to topics in document in which he/she is interested, user can search what this document is about and many more. User can also keep history of document and the information extracted. User can manage his/her account and also keep their analysis on specific document and add description.

# Modules

The modules of our application are explained below:

## Module 1: User Management

This module deals with the users 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.

## Module 2: Pre-Processing

This module allows user to upload the document on which he/she want to implement data science toolkit. Then image processing will be used to identify document pages because document can be various types as mentioned above. After that through natural language processing, content of the document (text) will be read by system to get the content of the document.

## Module 3: Selection

This module allows the user to have multiple options that they want to implement on document. User can search for summary of document, knows what this document is about, go into specific section of document in which they are interested and many more. For that many machine learning techniques like text summarization, topic modeling, sentiment analysis etc will be used to help user get what they want with document.

## Module 4: Report

This module allows user to get the report of the information they want to extract from the uploaded document. User will also get statistics about document like which words were used most in document, summary of document etc. 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.

# System Limitations/Constraints

# Software Process Methodology

Write down your software methodology/ software process that will be used for projectdevelopment.Also mention why you have chosen this methodology. (Usually 3-5 sentences)

1. You can use Object oriented Methodology, or Procedural methodology.
2. Choice of methodology will affect choice of tools and technologies
3. Choice of methodology will affect nature of design (SDS)
4. Choice should be made depending on your expertise and your needs e.g. most simulation and device level software can only be programmed in procedural languages.

# Tools and Technologies

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Tools**  **And**  **Technologies** | **Tools** | **Version** | **Rationale** |
| OpenCV | 3 | Machine Learning |
| TensorFlow | 2 | Machine Learning |
| Bootstrap | 4 | Front-End Web Development |
| Node.js | 13.5.0 | Back-End Web development |
| MS Word | 2016 | Documentation |
| MS Excel | 2016 | Presentation |
| Visual Studio Code | 16.4 | IDE |
| **Technology** | **Version** | **Rationale** |
| html | 5 | Language |
| css | 4 | Language |
| javascript | 1.8 | Programming Language |
|  | sql | 2018 | Programming Language |
|  | python | 2.8.1 | Programming Language |

# Project Stakeholders and Roles

Write down the project stakeholders and their roles.

Table 3Project Stakeholders for Proposed Project

|  |  |
| --- | --- |
| **Project Sponsor** | All web applications and desktop applications should have real client. Mention your project sponsor.  Default option will be: COMSATS University, Islamabad |
| **Stakeholder** | Mention your stake holders with their roles and responsibilities.  Default option will be   * Students names * Project Supervisor Name: Mr./Miss … * Final Year Project Committee: Evaluation of project |

# Team Members Individual Tasks/Work Division

Table 4Team Member Work Division for Proposed Project

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student Registration Number** | **Responsibility/ Modules** |
| Student 1 Name | Student 1 Registration Number | Describe the work division of each  student along with modules  E.g.  Mr. Ali (Module1-Module3)  Augmented reality and Databasestasks. |

# Data Gathering Approach

Write down information and requirement gathering approaches for proposed project e.g. Interview, Questionnaire etc. (Usually 3-5 sentences)

# Concepts

Mention the concepts that you will learn while doing the proposed project.

For example: Augmented Reality, Virtual Reality, Algorithms, API’’s Code injection, Closures, VI technique etc.

Not more than 4 sentences for one concept.(Usually 3-5 concepts are briefly mentioned)

Example:

Concept-1: Concept Name E.g. Augmented Reality (Briefly give the overview of concept with respect to your project)

# Gantt chart

Create the Grant Chart and provide estimated start and end dates of all proposed modules/tasks for each team member. Also identify the dependencies (which tasks cannot be started/completed, until the dependent task is completed). Gantt chart can be created using MS Project.



Figure 1Sample Gantt chart

# Mockups

Insert minimum mockups (Usually 4-6 mockups) which show the major modules mentioned in the scope section of the document. Do not include mockups for Login, Signup, Forgot Password, Contact Us, About Us etc. If the project is a Web or a Smartphone Application, then include at-least three mockups from each part of the project. You can design mockup in any design tool for example pencil tool (<https://pencil.evolus.vn/>) or Balsamiq (<https://balsamiq.com/>)





# Conclusion

Conclude this document.(Usually 4-5 sentences)

# References

Mention the books, research papers, web links etc.

# Plagiarism Report

Attach the Plagiarism report of your project scope document from library staff of turnitin tool (<http://turnitin.com>)