**Project Charter Document**

**Project Name:** **"Feature Extraction from Medical Journals"**

**Department:** Life Science and Health Care

**Focus Area:** Drug and Drug Composition Recognition

**Product/Process:** NLP

**Project Team:** DS\_Project\_Team\_59



**Prepared By**

|  |  |
| --- | --- |
| **Document Owner(s)** | **Project/Organization Role** |
| Sharat Chandra | Project Manager |
| Raju Kalyan | Project Manager |
| SP Pavani | Project Mentor |
| Mohammed Zaid Habibulla | Member |
|  |  |
|  |  |

**Project Charter Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Change Description** |
| 1.0 | 01/11/2021 | Sharat Chandra | Document created |
|  |  |  |  |

**TABLE OF CONTENTS**

[**1**](#_heading=h.3whwml4) **PROJECT CHARTER PURPOSE 3**

[**2**](#_heading=h.2bn6wsx) **PROJECT EXECUTIVE SUMMARY 3**

[**3**](#_heading=h.qsh70q) **PROJECT OVERVIEW 3**

[**4**](#_heading=h.3as4poj) **PROJECT SCOPE 4**

[4.1](#_heading=h.1pxezwc) Goals and Objectives 4

[4.2](#_heading=h.49x2ik5) Project Deliverables 4

[4.3](#_heading=h.2p2csry) Deliverables Out of Scope 4

[4.4](#_heading=h.147n2zr) Project Duration 4

[**5**](#_heading=h.3o7alnk) **PROJECT CONDITIONS 5**

[5.1](#_heading=h.23ckvvd) Project Assumptions 5

[5.2](#_heading=h.ihv636) Project Issues 5

[5.3](#_heading=h.32hioqz) Project Risks 5

[5.4](#_heading=h.1hmsyys) Project Constraints 5

[**6**](#_heading=h.41mghml) **Project Structure Approach 6**

[**7**](#_heading=h.2grqrue) **Project Team Organization Plans 6**

[**8**](#_heading=h.vx1227) **PROJECT REFERENCES 6**

[**9**](#_heading=h.3fwokq0) **APPROVALS 6**

[**10**](#_heading=h.1v1yuxt) **APPENDICES 8**

[10.1](#_heading=h.4f1mdlm) Document Guidelines 8

[10.2](#_heading=h.2u6wntf) Project Charter Document Sections Omitted 8

# PROJECT CHARTER PURPOSE

* Manual annotation and feature generation by biomedical experts are inefficient because they involve a complex process and require expensive and time-consuming labour. Therefore, efficient and accurate natural language processing (NLP) techniques are becoming increasingly important for use in computational data analysis, and advanced text mining techniques are necessary to automatically analyse the biomedical literature and extract useful information from texts. To bridge the gap between academic development and industrial requirements, we are developing a web application that recognizes and extracts the entities from the content which helps to minimize time commitments from domain experts and the manual efforts on researching content.



# PROJECT EXECUTIVE SUMMARY

* Project goals
* Objectives
  + Data prepossessing
* Scope
* Assumptions
* Risks
* Costs
* Timeline
* Approach
* Organization



# PROJECT OVERVIEW

* The increasing number of biomedical articles and resources, searching for and extracting valuable information has become challenging.
* Researchers consider multiple information sources and transform unstructured text data into refined knowledge to facilitate research productivity. However, manual annotation and feature generation by biomedical experts are inefficient because they involve a complex process and require expensive and time-consuming labour. Therefore, efficient and accurate natural language processing (NLP) techniques are becoming increasingly important for use in computational data analysis, and advanced text mining techniques are necessary to automatically analyse the biomedical literature and extract useful information from texts.
* Our project is about Information Extraction (IE), which is one of the important tasks in text analysis and Natural Language Processing (NLP). It involves extracting meaningful pieces of knowledge related to clinical subjects like drug names, drug compositions, diseases, etc. from Medical Journals as unstructured data and unstructured data is computationally opaque.
* We are building this custom tool which helps to minimize time commitments from domain experts and the manual efforts on researching content.



# PROJECT SCOPE

## Goals and Objectives

|  |  |
| --- | --- |
| **Goals** | **Objectives** |
| * Our major goal is recognizing Diseases, biomedical or Chemical composition names of drugs by named entity recognition (NER) which is an important task used to extract information and to recognize the Medical entities from Medical Journals. | * To maximise the utilisation of free-text medical journals, we focused on a particular sub-task of disease names,drug and it’s compositions information extraction and developing a dedicated named-entity recognition web application. |

## Project Deliverables

|  |  |
| --- | --- |
| **Milestone** | **Deliverable** |
| * PDF Extraction and Architecture Diagram | * Deliverable 1.1—Identifying Constraints and design the project architecture. * Deliverable 1.2— Data Extraction * Deliverable 1.3— Data Preparation |
| * EDA /Text preprocessing , Text parsing , Model Building | * Deliverable 2.1— EDA and Text Parsing * Deliverable 2.2— Model Building |
| * Model Evaluation and Deployment | * Deliverable 3.1— Model Evaluation, tuning and insights. * Deliverable 3. 2— Deployment |
| * Preparing for Final Presentation | * Deliverable4.1 – show case and review * Deliverable4.2 – Final Presentation and documentation * Deliverable4.3 – Handover and KT |

## Deliverables Out of Scope

* designs
* mobile app

## Project Duration (start date: 01/11/2021 End date: 01/12/2021)

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Milestone** | **Date Estimate** | **Deliverable(s) Included** | **Confidence Level** |
| * PDF Extraction and design the project architecture, explore Medical Journals which are in the form of PDFs. | [01/11/2021]  -  [08/11/2021] | * Deliverable 1.1—Identifying Constraints and design the project architecture. * Deliverable 1.2—Explore PDFs to extract relevant data. * Deliverable 1.3— Data Preparation | [High] |
| * Text Preprocessing/ EDA, Text Parsing, Model Building | [08/11/2021]  -  [15/11/2021] | * Deliverable 2.1— EDA and Parsing * Deliverable 2.2— Model Building | [High] |
| * Named Entity Recognition/Model Evaluation, tuning and insights, Deployment | [16/11/2021]  -  [22/11/2021] | * Deliverable 3.1— Model Evaluation, tuning and insights. * Deliverable 3. 2— Deployment | [High] |
| * PPT for Final Presentation and documentation | [23/11/2021]  -  [29/11/2021] | * Deliverable4.1 – show case and review * Deliverable4.2 – Final Presentation and documentation * Deliverable4.3 – Handover and KT | [Medium] |



# PROJECT CONDITIONS

## Project Assumptions

* Data is in an extractable format like pdf
* Data consists of features being extracted

## Project Issues

**Priority Criteria**

1 − High-priority/critical-path issue; requires immediate follow-up and resolution.

2 − Medium-priority issue; requires follow-up before completion of next project milestone.

3 − Low-priority issue; to be resolved prior to project completion.

4 − Closed issue.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Date** | **Priority** | **Owner** | **Description** | **Status & Resolution** |
| 1 | Evaluation | High |  |  |  |
| 2 | Deployment | High |  |  |  |

## Project Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Risk Area** | **Likelihood** | **Risk Owner** | **Project Impact-Mitigation Plan** |
| 1 | [Project Risk] | [High/Medium/Low] |  |  |
| 2 | [Project Risk] | [High/Medium/Low] |  |  |

## Project Constraints

* Maximize the quantity of features extracted from the medical journals
* Minimize incorrect feature detection
* Minimize the run time of the model training



# Project Structure Approach

* Text Extraction in most efficient way
* Data Preparation and EDA
* Model development
* Model evaluation and fine tuning
* Documentation



# Project Team Organization Plans

|  |
| --- |
| * Deliverable 1.1— * Deliverable 1.2— * Deliverable 1.3— |
| * Deliverable 2.1— * Deliverable 2.2— |
| * Deliverable 3.1— * Deliverable 3. 2— |
| * Deliverable4.1 – * Deliverable4.2 – * Deliverable4.3 – |



# PROJECT REFERENCES

<https://pdfminersix.readthedocs.io/en/latest/>

<https://pubmed.ncbi.nlm.nih.gov/34412834/>

<https://github.com/kormilitzin/med7>

|  |  |
| --- | --- |
| **Milestone** | **Deliverable** |
| [ |  |
|  |  |
|  |  |

# APPROVALS

**Prepared by** Mohammed Zaid Habibulla\_\_\_\_\_\_\_\_\_\_\_\_\_

Project Manager

**Approved by** Sharat Chandra M\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Project Sponsor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Executive Sponsor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Client Sponsor



# APPENDICES

## Document Guidelines

## Project Charter Document Sections Omitted

7. Project Team Organizational plans

