**Technical Design Document (TDD)**

**Title: Feature Extraction from Medical Journals**

|  |  |
| --- | --- |
| **Mentors*:*** | **Authors:** |
| ***Raju***  **Mentor (Client):**  ***Sharat Chandra*** | ***Anandakrishnan K V*** |
|  |  |

**Table of Contents**

[1. Introduction 3](#_Toc90209619)

[1.1 Purpose 3](#_Toc90209620)

[1.2 Scope 3](#_Toc90209621)

[1.3 Document Organization 3](#_Toc90209622)

[1.4 Audience 3](#_Toc90209623)

[2 Overview 4](#_Toc90209624)

[2.1 Context 4](#_Toc90209625)

[2.2 Product Feature 4](#_Toc90209626)

[2.3 Libraries 4](#_Toc90209627)

[3.1 Architecture Diagram 4](#_Toc90209628)

[3 Approach 5](#_Toc90209629)

[3.1 PyPdf2 5](#_Toc90209630)

[3.2 PdfMiner 5](#_Toc90209631)

[3.3 Spacy 5](#_Toc90209632)

[3.4 En-Core-Med7-Lg 5](#_Toc90209633)

[3.5 Pandas 5](#_Toc90209634)

[4. References 5](#_Toc90209635)

# Introduction

## **1.1 Purpose**

The purpose of this document is to outline the high-level design of the **Feature Extraction from Medical Journals** and provide an overview for the tool implementation.

Its main purpose is to –

* Recommend relevant: Extraction of medical terms from given medical journals.

## **1.2 Scope**

The Project aims to Feature Extraction from Medical Journals.

Feature Extraction is manually a complex task/execution. Hence Named Entity Recognition technique helps to abridge such difficulty in order to understand different types of Drug compositions and dosage and also their strength in the medical fields. Also, it helps to reduce further harm to people with different health conditions depending on what type of disease the person suffers.

## **1.3 Document Organization**

This document is organized into the following section:

|  |  |
| --- | --- |
| Introduction | Provides information related to the document |
| System Overview | Describes the approach, architectural goals and constraints, Guiding principles |
| Application Architecture | Describe the application architecture in terms of different layers of application. Description of the presentation layer, business layer, data access layer and resource layer and their relationship to each other. |
| Database Architecture | Describes the overall Data model and entity relationship diagram |
| Assumptions and Constraints | Details various assumptions made during design and development of the Online Screening tool |

## **Audience**

The intended audiences for this document are: -

* Innodatatics Inc.
* The project development team
* Mentors

# Overview

## **2.1 Context**

The evolving tech industry serves the business needs. The first key area for a business to grow is to identify the need or demands of the user.

Using NLP helps in Extraction of Named Entities from Medical Journals.

## **2.2 Product Feature**

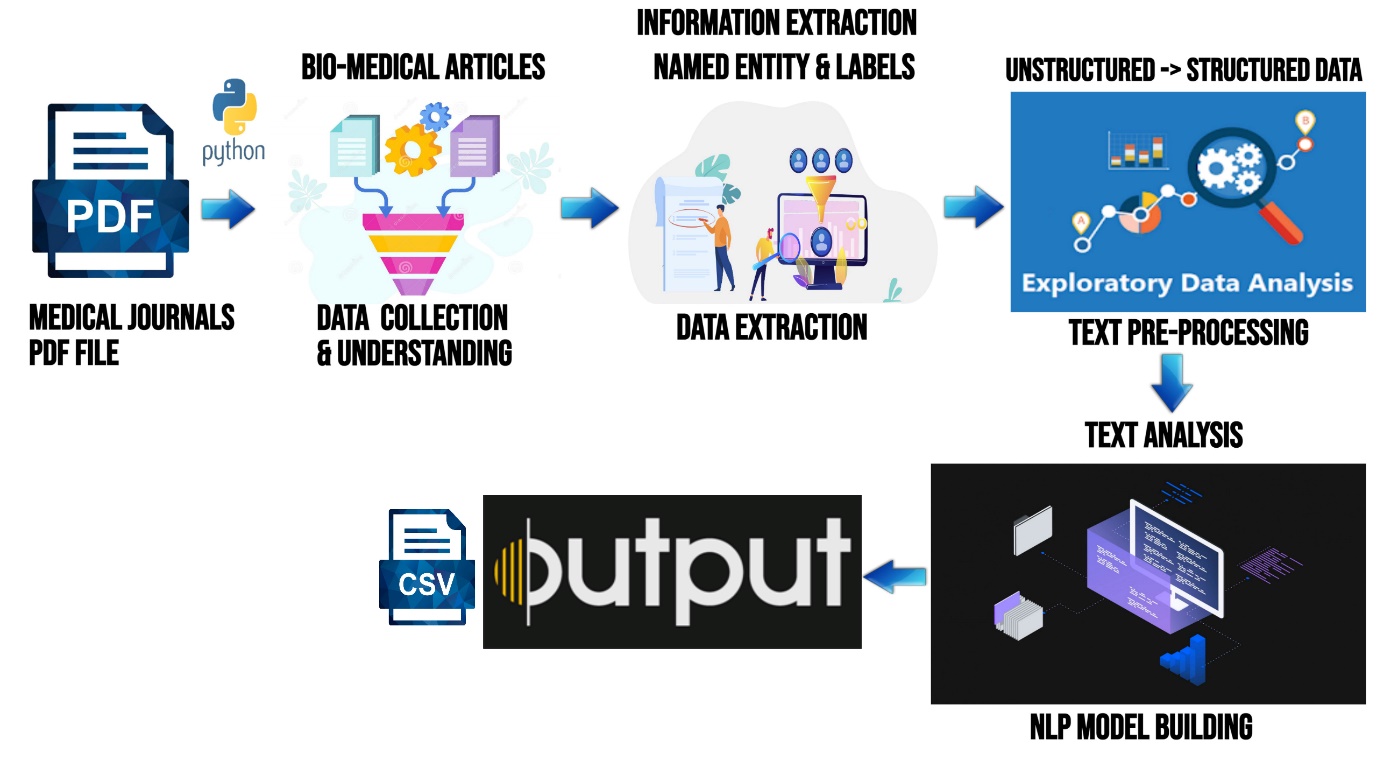
The major feature of the ***Project Title*** will be the following: -

* + **Natural Language Processing –** The question will take query process it extracts keywords from query string using NLP
  + **Information Extraction –** There will be a database containing all the information needed, populated using information extraction technique.

# 2.3 Libraries

* **Python Libraries** - PyPdf2, PdfMiner, Spacy, HuggingFace, En-Core-Med7-Lg and Pandas.

## **3.1 Architecture Diagram**



# Approach

We have tried ML approached to solve the issue which are listed below: -

## **3.1 PyPdf2**

**PyPdf2** is used for merging all medical journals.

## **3.2 PdfMiner**

**PdfMiner** is used for text extraction from pdf files.

## **3.3 Spacy**

**Spacy** is used for build information extraction or natural language understanding systems.

## **3.4 En-Core-Med7-Lg**

**En-Core-Med7-Lg** is used for extract medical entities.

## **3.5 Pandas**

**Pandas** is used for data analysis.

# 4. References

<https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/>

<https://githubmemory.com/repo/kormilitzin/med7/issues/11>