Birla Institute of Technology and Science-Pilani, Hyderabad Campus

First Semester 2018-2019

# 

# Information Retrieval (CS F469)

# Design Document

# Assignment 1

Domain Specific Information Retrieval System

Rohith Saranga (2017A7PS0034H)

Kasuba Badri Vishal (2017A7PS0270H)

Rikil Gajarla (2017A7PS0202H)

Under the guidance of

Dr. Aruna Malapati

#### Abstract:

This project aims to build a domain specific search engine based on vector space model. Initially the documents to be searched are pre-processed to generate tokens. In the vector space model, documents are represented as vectors containing weights of these tokens. When a query is given to the search engine, a vector is made from its tokens and similarity between all the documents is found. Then the documents are returned in order of decreasing similarity.

#### Architecture:

Programming Language: python

Dataset: Journals on Computer Science (350 journals scraped from archive.org)

Some of the python libraries used are nltk, pandas, numpy and pickle.

Folder Structure:

.

├── irstructures

│   ├── \_\_init\_\_.py

│   ├── document.py

│   ├── invertedindex.py

│   └── models

│       ├── \_\_init\_\_.py

│       ├── boolean\_retrieval.py

│       └── vector\_space.py

├── corpus

├── pickle\_files

└── main.py

Classes used are

1. Document
2. Tf\_Idf
3. InvertedIndex

The document object takes text from a file and does preprocessing like tokenization, case normalization, stemming and removal of stop words

Example Query:

