SIEMENS
Ingenuity for life

Project Title: Documents similarity for finding R packages

Name: Fazal Ahmad

Mentor: Mr. Sandeep Krishnan

Overview

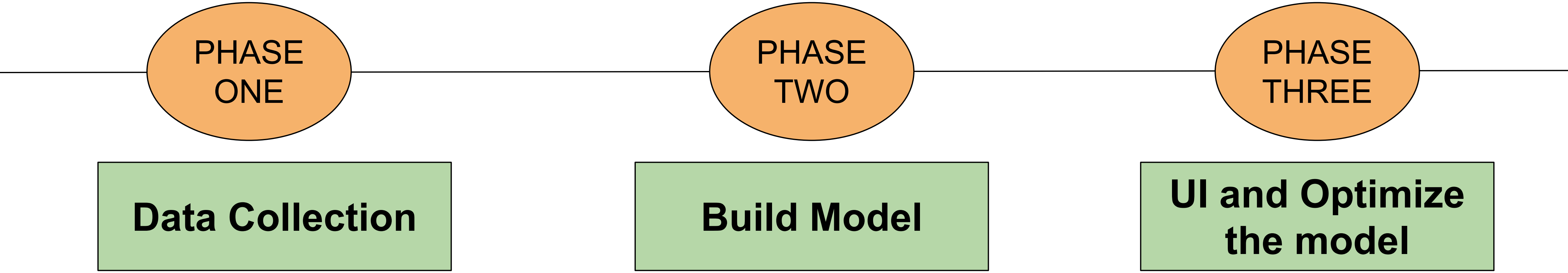
Problem Statement

Given a input text the system should suggest the relevant R package and in that R package it should also suggest the useful function of that package.

Why is it important?

For searching R package or function on google for specific work is time consuming. One need to go through the whole documentation of a specific package to get the relevant function.

Approach



PHASE ONE

Data Collection

PHASE TWO

Build Model

PHASE THREE

UI and Optimize the model

Data Collection

SIEMENS
Ingenuity for life

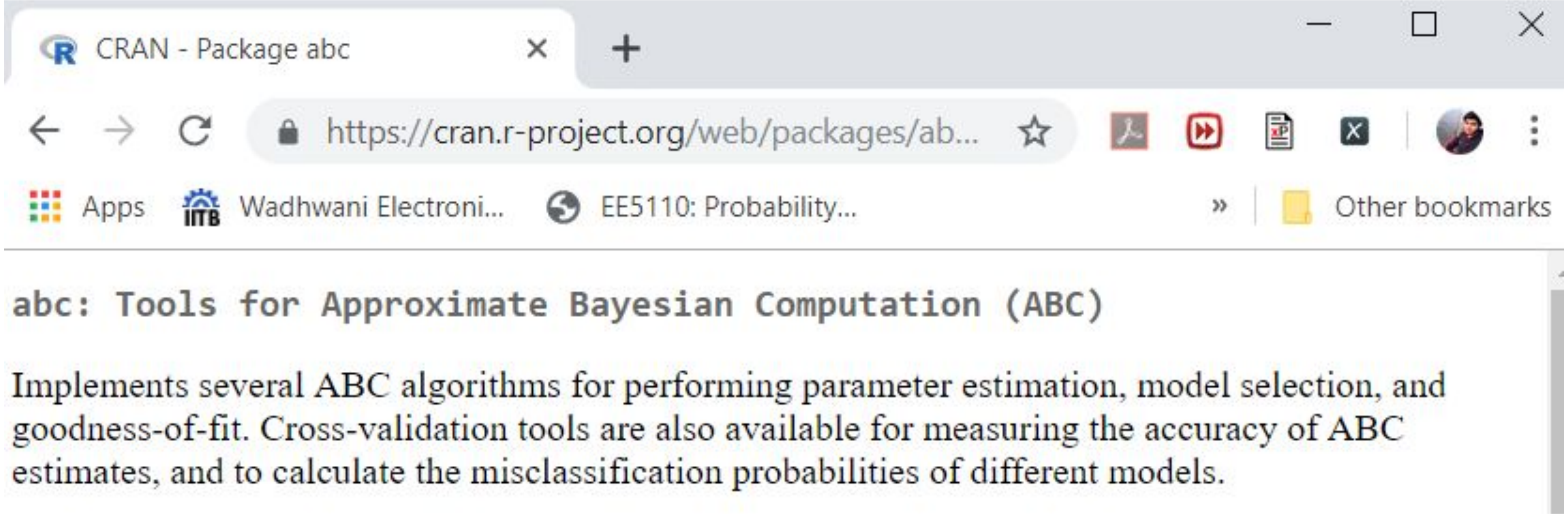
R package Data Set

I collected the description of all R packages from the offical website using web scraping.

Function Level Data Set

For every R package there is a reference manual in the form of pdf so I collected the description of all the functions of a given package from the pdf using PyPDF2 library in python.

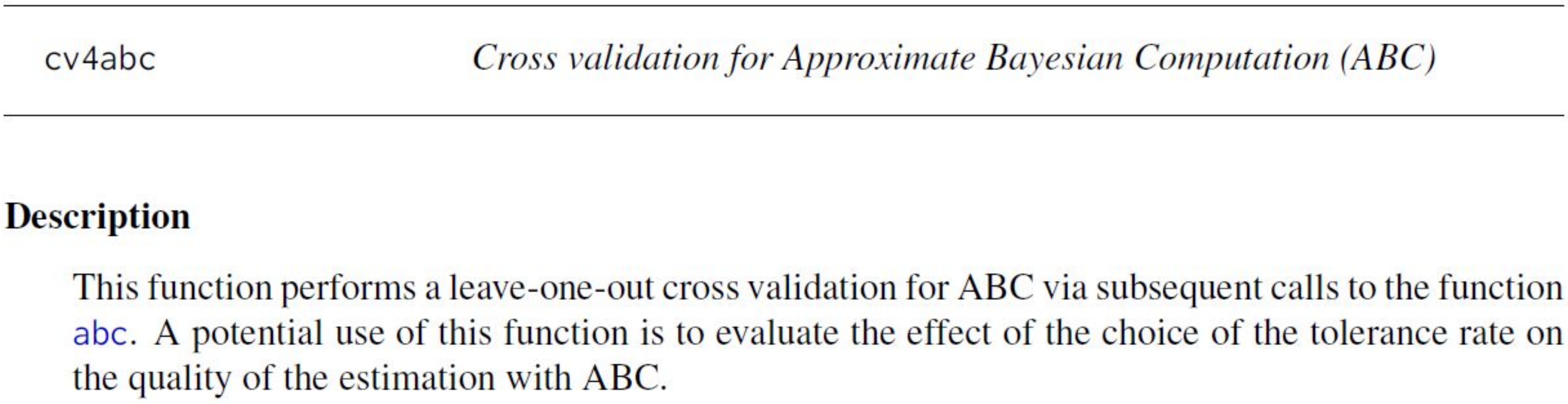
Description of package “abc”



abc: Tools for Approximate Bayesian Computation (ABC)

Implements several ABC algorithms for performing parameter estimation, model selection, and goodness-of-fit. Cross-validation tools are also available for measuring the accuracy of ABC estimates, and to calculate the misclassification probabilities of different models.

Function of package “abc”



cv4abc *Cross validation for Approximate Bayesian Computation (ABC)*

Description

This function performs a leave-one-out cross validation for ABC via subsequent calls to the function abc. A potential use of this function is to evaluate the effect of the choice of the tolerance rate on the quality of the estimation with ABC.

Model Building

Algorithm

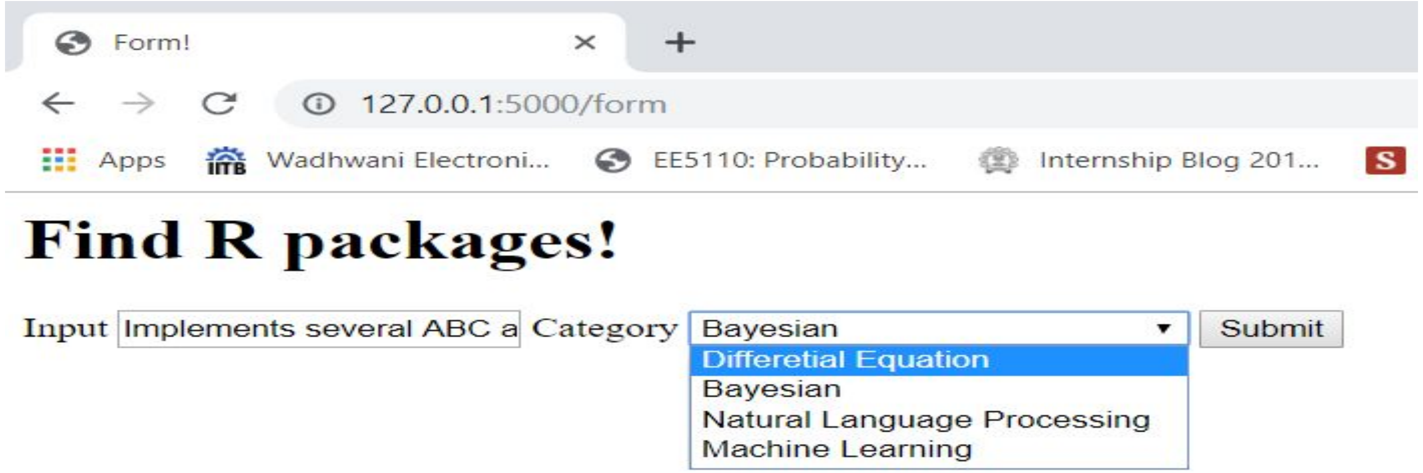
I converted the input text and the description into a matrix form using word embedding matrix. And tried to check the similarity between the matrix.For this I used “Spacy” library in python.

Drawbacks

Since the dimensions of the matrix is big so the complexity of matrix computation is high because of this it is taking around 15 to 20 sec for each search.

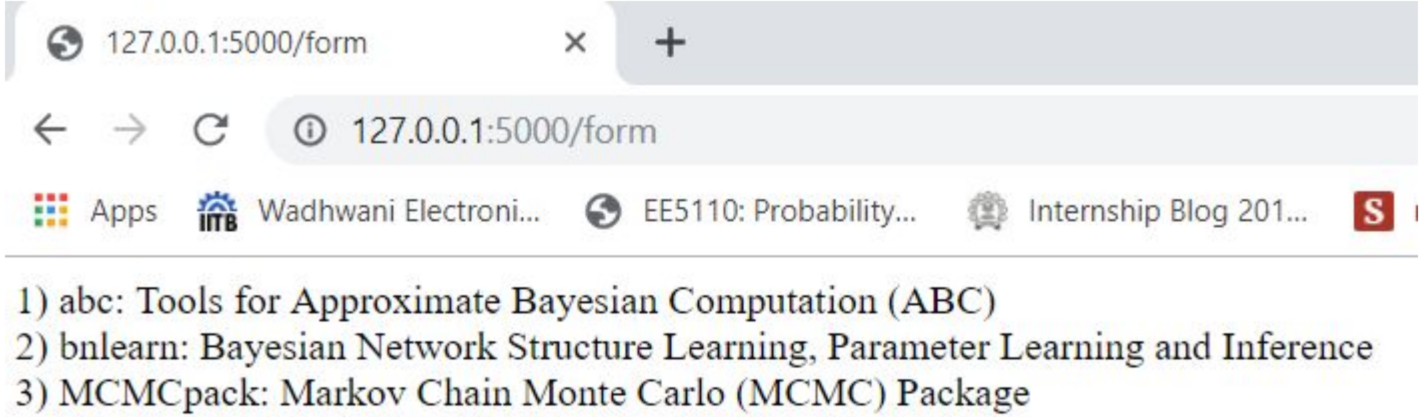
User Interface for searching R package

SIEMENS
Ingenuity for life



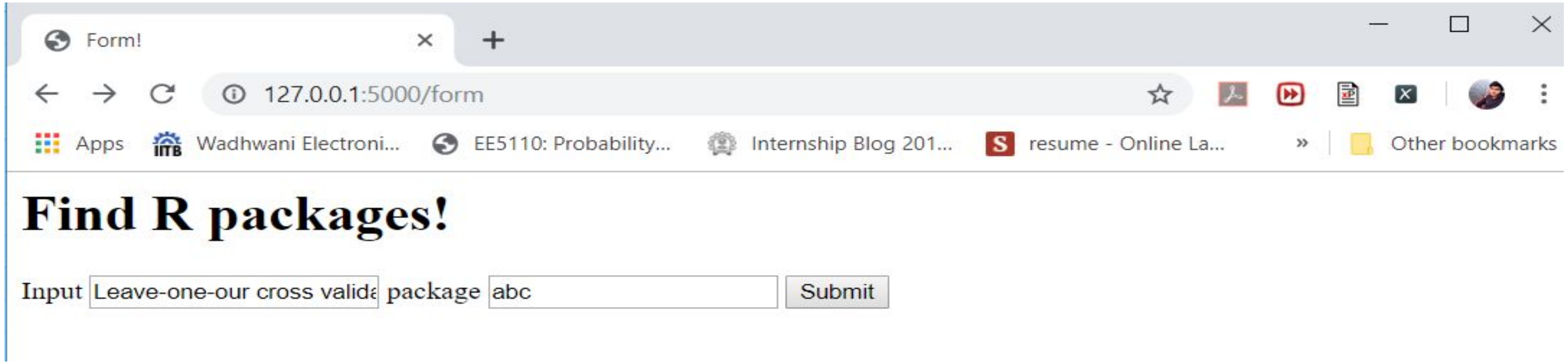
Find R packages!

Input: Implements several ABC a
Category: Bayesian
- Differetial Equation
- Bayesian
- Natural Language Processing
- Machine Learning



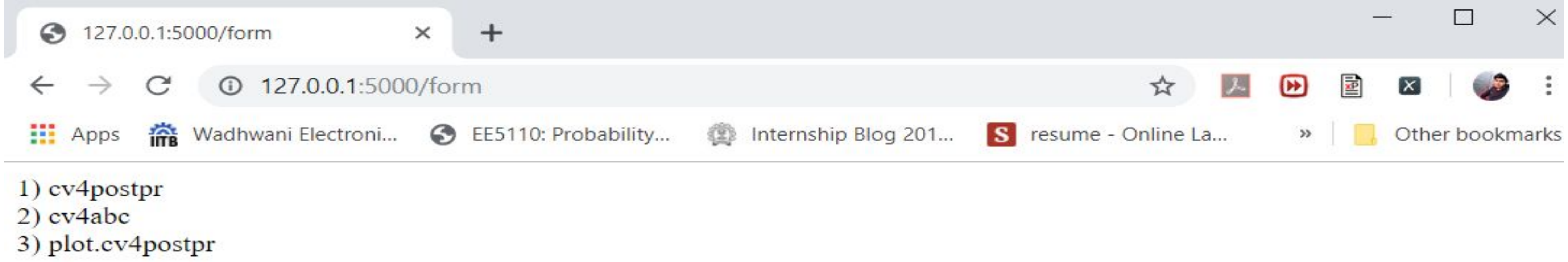
1) abc: Tools for Approximate Bayesian Computation (ABC)
2) bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference
3) MCMCpack: Markov Chain Monte Carlo (MCMC) Package

User Interface for searching function of a specific package



Find R packages!

Input Leave-one-our cross valida package abc Submit



```
1) cv4postpr
2) cv4abc
3) plot.cv4postpr
```

Optimization and results

Optimization

To reduce the dimensions of 2D matrix I took the average of words in the description so that the matrix computation process speeds up. After this optimization it is taking around 2 to 3 sec.

Results

To check the accuracy of model I generated the keyword for every description using "RAKE" library in python. And gave those keyword as input to the model and the accuracy is around 79%.

Possibilities to extend

We can extend this R package search engine to other programming language like python,java etc.

Also we can make the system dynamic, like whenever a new package comes it should directly add to our data set.

We can also optimize the algorithm by making clusters of all the description to reduce unnecessary matrix comparison.

Thank You