Milestone 1

Project : Web Search Engine

Members : Anil Shanbhag and Rahee Borade

Things Done till now

- Scraper

  To make it as close to real world as possible , the scraper scrapes pages from websites on www . The class “Scraper” is implemented in python for convenience as c++ [even with boost] doesnt provide a convenient way to download web pages from net and make some processing on them .

  The essence of it is 4 main components :

1 . Urls queue – the set of urls to be parsed

2 . Robots.txt file parser

3 . Set – Set of urls already downloaded

4 . Cache – Cache of robots.txt files already downloaded

In addition to this module also implements a savestate and loadstate methods to preserve the state of the parser in case we wish to halt it at a given point .

Using the parser a collection of 5.15 k html pages were downloaded into “Repository” ~ 300mb

- Indexer step started :

  Indexing involves many steps . Two of the routines have been implemented as c++ classes :

  1 . StopWords : There are two possible implementations of this classes , one using tries and other using map . The current implementation uses map . This class essentially exposes a boolean function to check if word is a stop word or not .

2 . URLResolver

Since urls in the html page tend to be in many forms like full links or relative paths wrt curdir or wrt root , it is necessary to have a url resolver for urls in page . This is necessary for building url index and page rank computation .