# CSCI 6370 IR and Web Search Project

## Part 1: Html Parser (a simple one)

### Due date: Monday, June 8 @ 12:00 PM.

### Note 1: The course project “MySearchEngine.com” is composed of 5 part with one part per week.

### Note 2: The project is a group project. That is, you can organize a group of at most 3 members to work together.

### Note 3: Only Python is allowed for completing the course project.

## Homework Problem:

You are given a zipped file “Jan.zip” which contains the following files:

* aol.html
* armed.html
* baptist.html
* bill.html
* birdnbee.html
* bunker.html
* cache.html
* child.html
* creditcard.html
* debug.html
* edwardii.html
* explain.html
* fab.html
* galant.html
* gravies.html
* harley.html
* heartprob.html
* hippos.html
* jesus.html
* kitty.html
* marriedplay.html
* phone.html
* problem.html
* qc.html
* quickies.html
* snow.html
* superbowl.html
* topten.html
* y2k.html
* y2kfollow.html
* y2kms.html

You need to write a Python to do the following two parts:

### Part 1

Repeatedly read the text of each of these files and extract index terms ( or words, or strings) that contain ONLY alphabets. Save these words (or strings) in a list (or some other data structure).

For examples, the following words ( or strings) are not index terms, hence shall not be extracted:

<title>

[rec.humor.funny]

type="text/css"

media="screen">

bgcolor="#ffffff"

text="#000000"

link="#0000ee"

vlink="#551a8b">

<!--

"/include/rhf/top.ofi"

-->

<map

name="joke-header-map">

shape="rect"

coords="103,52,194,71"

href="../../../best.html"

alt="best

jokes">

shape="rect"

coords="205,52,299,71"

href="../../../current.html"

alt="current

jokes">

href="../../../images/joke-header.html"><img

src="../../../images/joke-header.gif"

alt="fun

However, the following words (or strings) are index terms, hence must be extracted:

subject

provided

start

stuff

ismap

end

rhf

joke

archives

subject

much

was

hosting

survey

worst

music

videos

beating

poison

house

milli

vanilli

took

award

for

cheesiest

band

said

from

milli

vanilli

For ease of work, you shall convert all words (strings) into lower case.

### Part 2.

Write a loop to do:

* Ask the user to enter a word, called “search key.”
* Search for this search key in the extracted index terms of all the files. If a file contains the search key, then print “found a match” and display the names of the files containing the search key.
* If none of the files contains the search key, then print “No match”.
* The loop ends when an empty search key is entered.

For example, the following shows some iterations of the loop:

Now the search begins:

enter a search key=> music

found a match: ./Jan/fab.html

enter a search key=> cat

no match

enter a search key=>

Bye

### Part 3.

Special requirements for your program: To test your program, our TA would save it in a folder where the zipped file “Jan.zip” resides. When our TA runs your program, it shall know to read files from “Jan.zip” without asking for any additional input. Otherwise, your program is considered not working correctly.

## How to submit your work?

Submit your python codes to Blackboard before the deadline.

One group only needs to submit one copy. Please make sure to include group members’ names and ID’s in your submission.

## Grading

The TA would do test runs of your program to check whether it works.