Report of Maliang

1.Function

a) To search for some queries and get some responses of URL in “info.ruc.edu.cn”

b) You can click the link and go into the webpage and look through the content.

2. How to design

a) Design the function to crawl all the URL and download the text file.

b) Get all the title and body in the text file and save them.

c) Use THULAC to make word segmentation .

d) Set dictionary that contains all the words disappear in the text file after segmentation.

e) Get the query and analyze . Score the URL and show the top 10 webs in the webpage .

3.Some skills .

a) I found that the time of setting a dictionary every time is almost the same reading the dictionary that saved in the file . So I set it every time .

b) Run the THULAC in the THULAC file instead of setting a path.

4.The query that can be analyze

a) Almost of the queries can be analyze and show some results.
b) But actually , some words that be departed can not show the most correct results .

5.Summary

a) Difficulty:
   i. The number of the URL is not enough
   ii. Have difficulty getting the dictionary
   iii. The character of the body is strange
   iv. The query that I get and the content I set are garbled.
b) What I get:
   i. More knowledge of Linux and some cmd .
   ii. Basic recognition of html 、 python 、LETAX.

6.How to use .

a) Crawling.cpp Content.cpp CWS.cpp Make_termlist.cpp Score.cpp Show_res.cpp
b) Run the cpp below one by one .