CURSORY SEARCH ENGINE

hdn

PROBLEMS:

- **get URL & html parser:**

1.^@

2. <p>

- **Web programming & info retrieval**

1.python :web +

C++ : index +

python(jieba) : seg-word. *too slow (2)*

2.py -> index

PROBLEMS:

- **tf-idf:**

1.key word = [“info”, “ruc”, “ACM”]

doc1 = “info ..... ruc .... ACM”

doc2 = “info ruc ACM ... ACM ... ACM”

=> score[doc1] > score[doc2]

2.consider that there are mainly news and introductions in http://info.ruc.edu.cn

so that the title could be important

=> *Increase the score of the words which occur in title*