2014/03/01

Text extractor

Then decide the websites we should use like:

BBC

CNN

We divide main tasks into two parts: crawler and filter

url

Websites

parsor

Crawler group:

Ajay, Son, Sharu, Paul

Source website

title

crawl-Description

RSS

Crawler

CSS 10

RSS parsor

Index

crawler

Raw category

NoSQL

RDBS

File

Crawled time

Article time

Rss-description

Filter group:

Hao, Victor, Scott

2014/03/07

One member is assigned to write page parser for one specific news website within one week.

CNN - Scott

BBS – Sharu

New York Time – Ajay

Implementation Suggestion:

Python 2.7

Beautiful Soup

Database used – solo, HBase

solo

Page

HBase