Nostradamus

Term Project
CS60092: Information Retrieval

Aseem Patni **Nevin** Valsaraj **Sabyasachee** Baruah

Pramesh Gupta **Arkanath** Pathak **Pranjal** Pandey **Sanyam** Agarwal

Objectives

- Build a universal product search engine combining the results of most of the popular product sellers.

- A good scoring algorithm that takes into account a wide variety of features including the non-trivial elements like user’s sentiments in the comments and reviews for deciding the score of a product.

- Build a scalable system so that the product index can be updated easily with time.

The Backend

- The whole backend will be implemented in JAVA.
- **Apache Nutch**, v1.9, an open source Web crawler.
- **Apache Solr**, v4.0.0, open source indexing and search platform.
- **Semantria**, a cloud based Text and Sentiment Analysis API.

Apache nutch

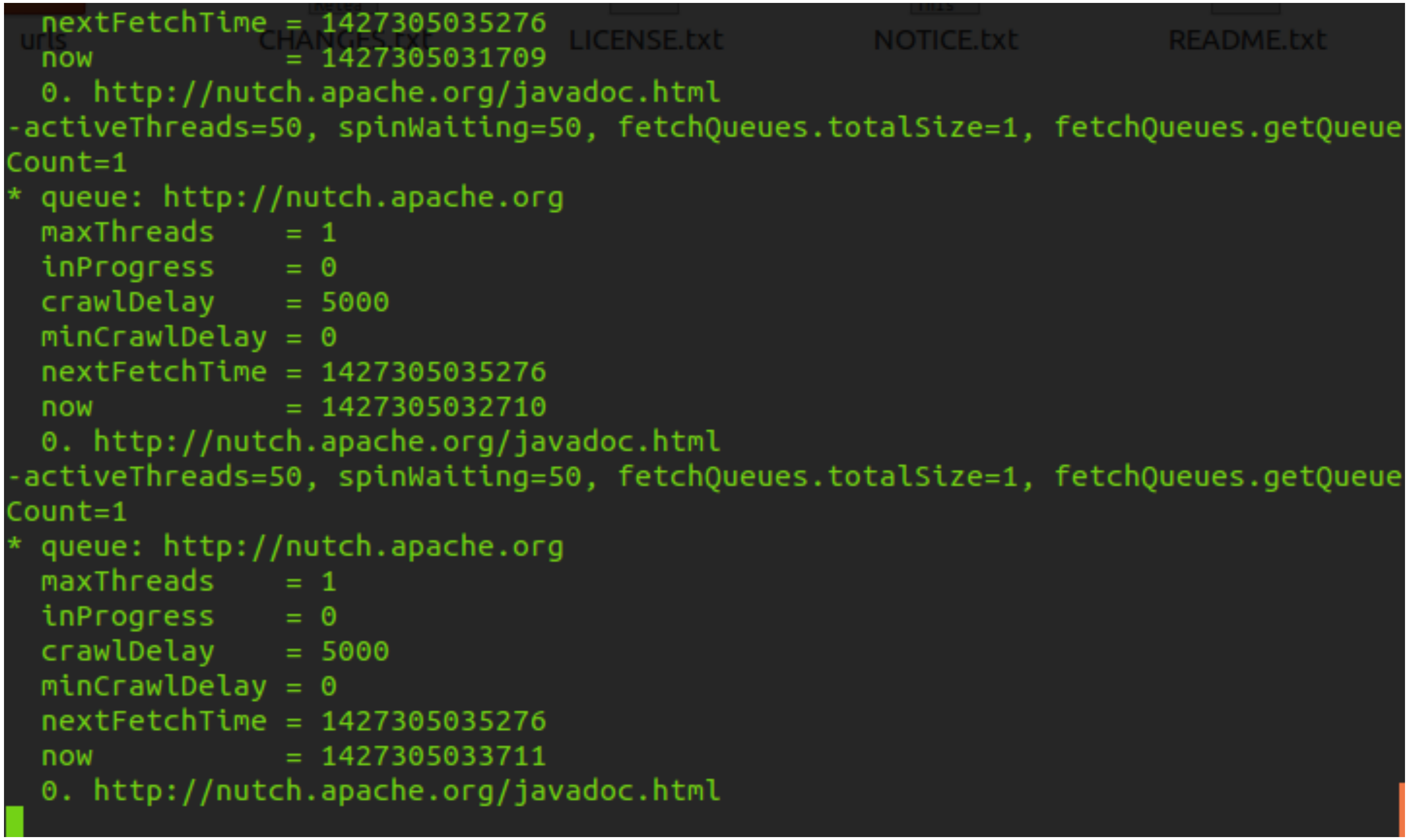
- Apache Nutch is an open source Web crawler written in Java.
- Provides a highly modular architecture
- Highly scalable and relatively feature rich crawler
- Can provide custom parsing to bias the important pages.

Nutch in Nostradamus

- We are using Nutch to index the following pages for each product:

  1. Product Profile Page
  2. Comments and Reviews Page (if different from Profile)
  3. Product Brand Profile Page (once for multiple products)

Work in Progress



```
  nextFetchTime = 1427305035276
  now           = 1427305031709
  0. http://nutch.apache.org/javadoc.html
-activeThreads=50, spinWaiting=50, fetchQueues.totalSize=1, fetchQueues.getQueue
Count=1
* queue: http://nutch.apache.org
  maxThreads     = 1
  inProgress     = 0
  crawlDelay     = 5000
  minCrawlDelay = 0
  nextFetchTime = 1427305035276
  now           = 1427305032710
  0. http://nutch.apache.org/javadoc.html
-activeThreads=50, spinWaiting=50, fetchQueues.totalSize=1, fetchQueues.getQueue
Count=1
* queue: http://nutch.apache.org
  maxThreads     = 1
  inProgress     = 0
  crawlDelay     = 5000
  minCrawlDelay = 0
  nextFetchTime = 1427305035276
  now           = 1427305033711
  0. http://nutch.apache.org/javadoc.html
```

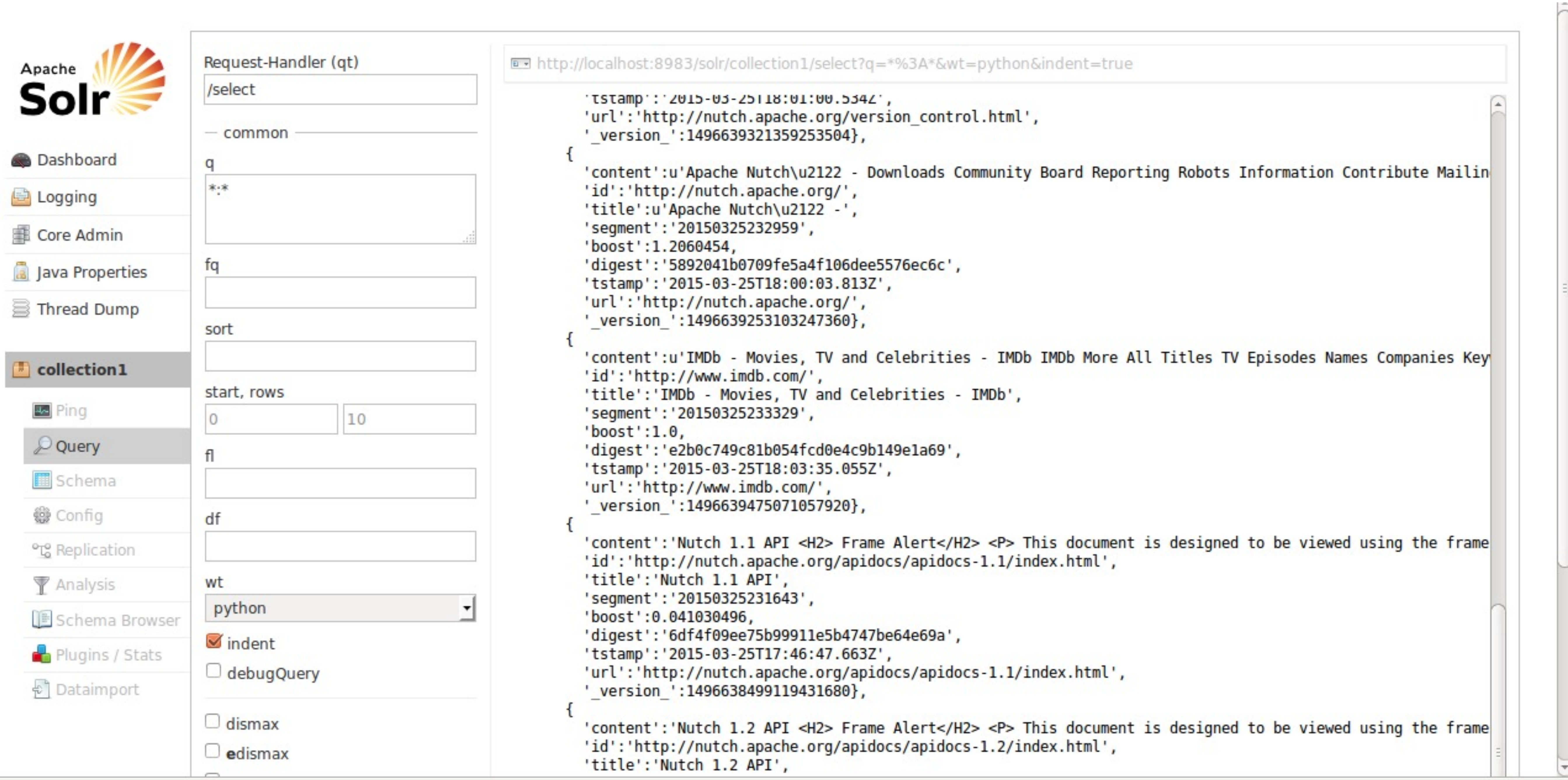
Apache Solr

- An open source enterprise search platform, written in Java, from the Apache Lucene project.
- Features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, and database integration.
- Highly scalable and fault tolerant
- HTML administration interface
- Schema-less mode and Schema REST API

Solr in Nostradamus

- We use Solr as a search engine on the back-end.
- We will then setup client-side JavaScript application that can access Solr via its REST-like interface.
- We also plan to use various different features of Solr like caching of queries, filters, and interface features like Auto-suggest.

Work in Progress



Request-Handler (qt): /select

common

q: `*:*`

fq:

sort:

start, rows: 0 10

fl:

df:

wt: python

☑ indent

□ debugQuery

□ dismax

□ edismax

http://localhost:8983/solr/collection1/select?q=*%3A*&wt=python&indent=true

```
        'tstamp':'2015-03-25T18:01:00.534Z',
        'url':'http://nutch.apache.org/version_control.html',
        '_version_':1496639321359253504},
      {
        'content':u'Apache Nutch\u2122 - Downloads Community Board Reporting Robots Information Contribute Mailin
        'id':'http://nutch.apache.org/',
        'title':u'Apache Nutch\u2122 -',
        'segment':'20150325232959',
        'boost':1.2060454,
        'digest':'5892041b0709fe5a4f106dee5576ec6c',
        'tstamp':'2015-03-25T18:00:03.813Z',
        'url':'http://nutch.apache.org/',
        '_version_':1496639253103247360},
      {
        'content':u'IMDb - Movies, TV and Celebrities - IMDb IMDb More All Titles TV Episodes Names Companies Key
        'id':'http://www.imdb.com/',
        'title':'IMDb - Movies, TV and Celebrities - IMDb',
        'segment':'20150325233329',
        'boost':1.0,
        'digest':'e2b0c749c81b054fcd0e4c9b149e1a69',
        'tstamp':'2015-03-25T18:03:35.055Z',
        'url':'http://www.imdb.com/',
        '_version_':1496639475071057920},
      {
        'content':'Nutch 1.1 API <H2> Frame Alert</H2> <P> This document is designed to be viewed using the frame
        'id':'http://nutch.apache.org/apidocs/apidocs-1.1/index.html',
        'title':'Nutch 1.1 API',
        'segment':'20150325231643',
        'boost':0.041030496,
        'digest':'6df4f09ee75b99911e5b4747be64e69a',
        'tstamp':'2015-03-25T17:46:47.663Z',
        'url':'http://nutch.apache.org/apidocs/apidocs-1.1/index.html',
        '_version_':1496638499119431680},
      {
        'content':'Nutch 1.2 API <H2> Frame Alert</H2> <P> This document is designed to be viewed using the frame
        'id':'http://nutch.apache.org/apidocs/apidocs-1.2/index.html',
        'title':'Nutch 1.2 API',
```

semantria

- Semantria applies Text and Sentiment Analysis to tweets, facebook posts, surveys, reviews or enterprise content.

- A cloud based Text and Sentiment Analysis API is available for multiple languages including JAVA.

- The Semantria API service returns a sentiment score with an out of the box precision rate of about 65-70%, without any model training

- Free for first 10,000 queries.

Semantria in Nostradamus

- "*This is an excellent product*" returns a sentiment score of **+0.60**
- This score is added as a bias to the overall score of the products

Websites to be crawled for testing

- Bosch Tools (http://boschtools.com)
- Amazon (http://amazon.com)
- Flipkart (http://flipkart.com)
- These will be crawled partially for test purposes only.

The Front-end

- We will be providing a web based GUI.
- We are planning to use the following libraries:

1. Twitter Bootstrap v3.0
2. JQuery
3. AJAX Solr library (https://github.com/evolvingweb/ajax-solr) , a JavaScript framework for creating user interfaces to Solr.

Thank You! :-)