PROJECT REPORT

QA SYSTEM

PARADIGM:

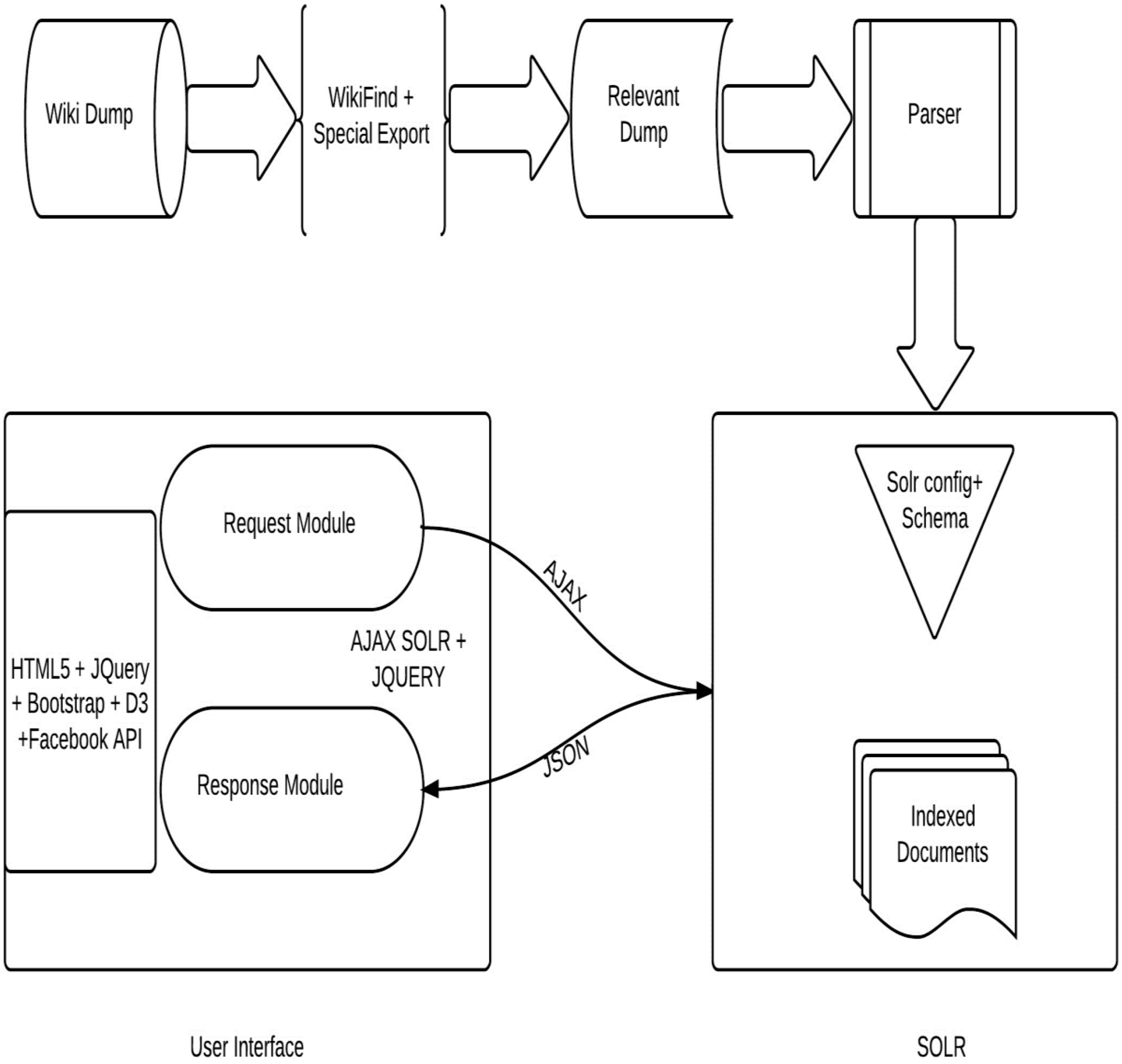
Aswin Bharadwaj Ramesh

Deepak Ravishankar Ramkumar

Aravindhan Thanigachalam

Vivekanandh Vel Rathinam

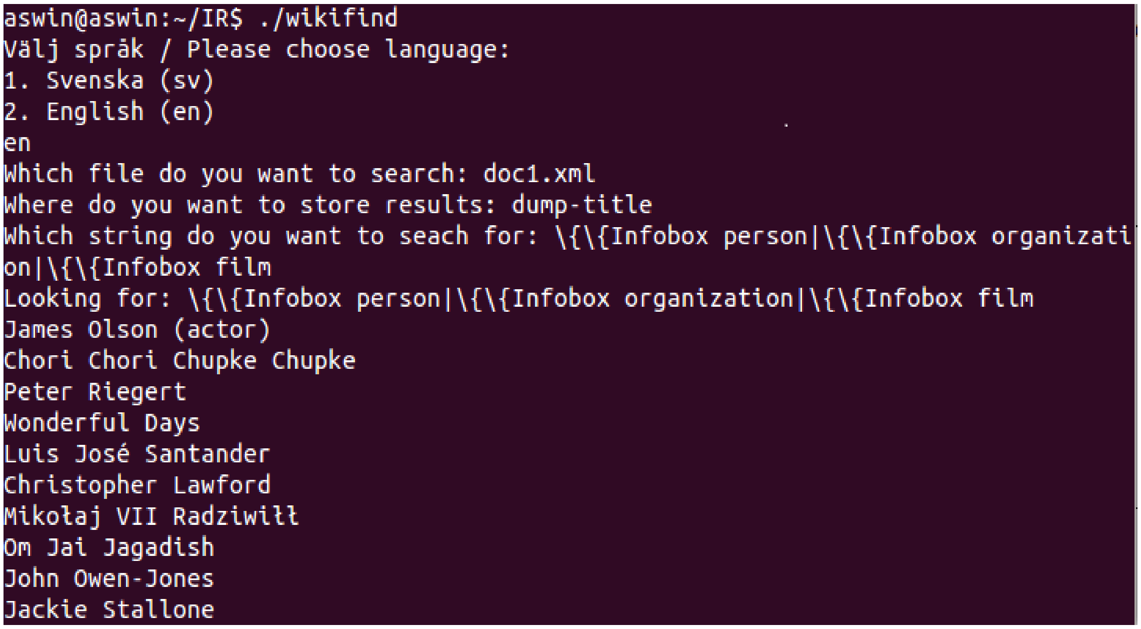
SYSTEM DIAGRAM



SYSTEM DESCRIPTION

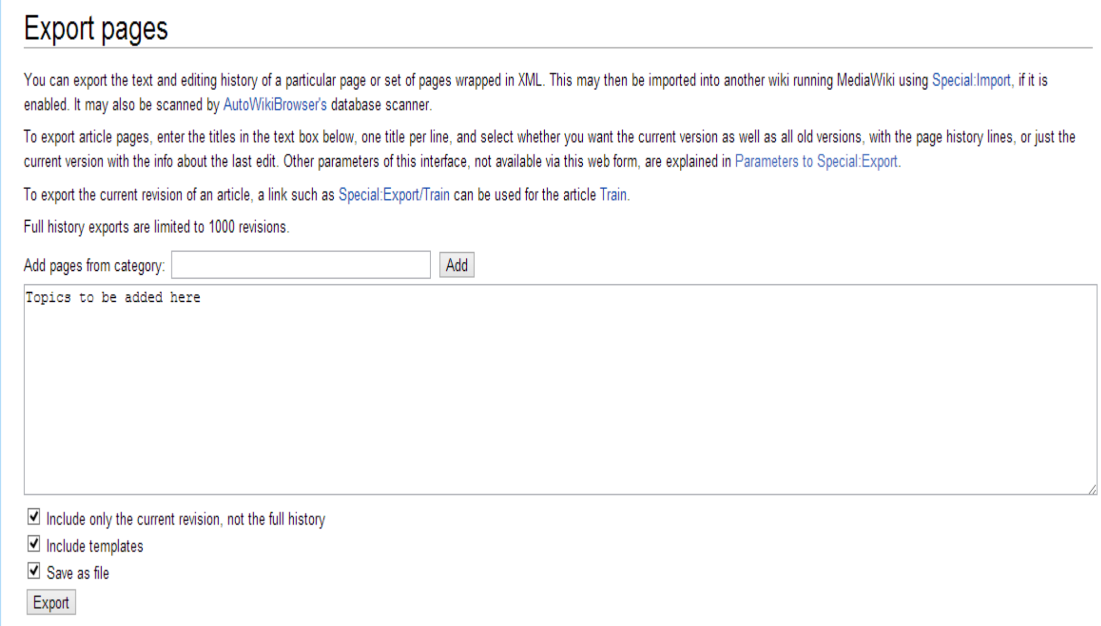
1) Processing of the Wikipedia dump

We are using an online tool 'WIKIFIND', a software written in C++ to find topics of Wikipedia pages related to People, Organization and Films.



```
aswin@aswin:~/IR$ ./wikifind
Välj språk / Please choose language:
1. Svenska (sv)
2. English (en)
en
Which file do you want to search: doc1.xml
Where do you want to store results: dump-title
Which string do you want to seach for: \{\{Infobox person|\{\{Infobox organizati
on|\{\{Infobox film
Looking for: \{\{Infobox person|\{\{Infobox organization|\{\{Infobox film
James Olson (actor)
Chori Chori Chupke Chupke
Peter Riegert
Wonderful Days
Luis José Santander
Christopher Lawford
Mikołaj VII Radziwiłł
Om Jai Jagadish
John Owen-Jones
Jackie Stallone
```

An equal proportion of these topics are given as input to another online tool 'EXPORT PAGES' in the Wikipedia site http://en.wikipedia.org/wiki/Special:Export to generate the XML dump.



Export pages

You can export the text and editing history of a particular page or set of pages wrapped in XML. This may then be imported into another wiki running MediaWiki using Special:Import, if it is enabled. It may also be scanned by AutoWikiBrowser's database scanner.

To export article pages, enter the titles in the text box below, one title per line, and select whether you want the current version as well as all old versions, with the page history lines, or just the current version with the info about the last edit. Other parameters of this interface, not available via this web form, are explained in Parameters to Special:Export.

To export the current revision of an article, a link such as Special:Export/Train can be used for the article Train.

Full history exports are limited to 1000 revisions.

Add pages from category:

Topics to be added here

☑ Include only the current revision, not the full history

☑ Include templates

☑ Save as file

The infoboxes from each document in the XML dump are parsed using a program written in Java and then the parsed contents are converted to the SOLR XML format. Each SOLR XML document will have the following fields.

a) **Person**
Name, image, caption, birth name, birth date, birth place, death date, death place, residence, nationality, other names, citizenship, education, occupation, known for, organization, notable works, title, religion, spouse, children, parents and awards.

b) **Organization**
Name, former name, image, caption, abbreviation, motto, formation, founder, extinction, type, purpose, professional title, headquarters, location, coordinates, services, leader title and leader name.

c) **Films**
Name, image, caption, film name, director, producer, writer, screenplay, based on, narrator, starring, music, cinematography, studio, distributor, released, runtime, country, language, budget and gross.

2) Indexing the SOLR formatted XML documents

After setting up the SOLR instance in any of the servlet containers, the field entries that are present in the XML document are to be defined in schema.xml. The SOLR schema will be explained in the later sections.

After the field definitions are made in schema.xml, the SOLR formatted XML documents are indexed into SOLR using the command given below

```
java -jar post.jar *.xml
```

3) User Interface

The User Interface gives out the possible types of queries that the user can ask by loading it in drop boxes and also giving out a text box for the user to fill out the name of a person, organization or a film to be queried upon.

User has to select the respective document type to be searched on. The input the user enters in the text box is the name of the document to be searched.

The answer to be returned is based upon the questions selected from the drop boxes. A query is formed and sent to the SOLR server and the response from the server is parsed in the client side to show the results. Various features of this system will be explained in the next section.

ADDITIONAL FEATURES OF THE QA SYSTEM

1) ***Auto Suggest:***

We are providing the user with a set of autosuggested results which pops down in a combo as he types ahead in the search box. A facet query is sent to the SOLR server as and when the search box is changed.

2) ***Automatic Spell Correction:***

A request handler has been created in SOLR to handle spelling corrections. Whenever a query is sent to SOLR, the system checks for a JSON response with the results. If there are no results, a spell checking query is fired to SOLR which gives a response with a list of spelling suggestions. The system takes the first suggestion from the response and answers the question.

3) ***Related Trivia:***

Information related to the particular query is also displayed to the user.

4) ***Tag Cloud:***

This feature has been provided in our user interface. It lists the top 100 personalities, organizations and films. The heuristic for selecting the top 100 documents is to sort the documents in the decreasing order of information present in them.

5) ***Clustering:***

The name of the persons that are indexed are grouped based on the person's occupation and displayed on the UI using a widget. The Facet pivoting in SOLR is used to perform the grouping.

6) ***Film Statistics:***

The top most films are displayed for each year based on their gross and budget. A combo has been presented in the User interface to select the year. A pop up window displays the name of the films and their gross. A pie chart is plotted with the gross or budget of the top six films for the selected year.

7) **Share on Facebook :**

The users will be allowed to share the application on facebook. The application has been hosted at http://qasysapp.herokuapp.com/.

CONFIGURATION DETAILS IN SOLR

1) Schema

All the fields in the SOLR XML document are indexed and stored as a string value using the SOLR StrField class. A unique ID is also created for each document dynamically using the SOLR UUIDField class.

Field types has been created for the autocomplete and the spelling suggestion feature. For the autocomplete feature, the corresponding field type indexes the names as n-grams which makes retrieval possible in the autosuggestion implementation.

```
<fieldType name="text_autocomplete" class="solr.TextField">
<analyzer type="index">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.LowerCaseFilterFactory"/>
<filter class="solr.EdgeNGramFilterFactory" minGramSize="1" maxGramSize="25" />
</analyzer>
<analyzer type="query">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
</fieldType>
```

Another field type for spelling correction has been created to split the names of the documents before indexing it. The white space tokenizer has been used to perform this task. The spell correction component performs the corrections comparing with the individual tokens in the index.

```
<fieldType name="text_general" class="solr.TextField" >
```

```
<analyzer type="index">
<tokenizer class="solr.StandardTokenizerFactory"/>
<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
<filter class="solr.SynonymFilterFactory" synonyms="index_synonyms.txt" />
<filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
<analyzer type="query">
<tokenizer class="solr.StandardTokenizerFactory"/>
<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
<filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" />
<filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
</fieldType>
```

Field names which are comma delimited like the occupation of a person are delimited and indexed using the pattern tokenizer factory.

```
<fieldType name="str_split_on_comma" class="solr.TextField">
<analyzer>
<tokenizer class="solr.PatternTokenizerFactory" pattern="[,/]s*" />
<filter class="solr.LowerCaseFilterFactory" />
<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
<filter class="solr.TrimFilterFactory" />
</analyzer>
</fieldType>
```

The location field type is defined to hold the coordinates of a location.

```
<fieldType name="location" class="solr.LatLonType" subFieldSuffix="_coordinate"/>
<dynamicField name="*_coordinate” type="tdouble" indexed="true” stored="false" />
```

2) Request Handlers

A request handler named select has been created to perform the default querying in SOLR. Number of rows to be returned and the default field to be searched has been specified as parameters.

Another request handler has been created for spelling corrections. This request handler uses a custom defined search component to perform the correction. This search component uses the DirectSolrSpellChecker and WordBreakSolrSpellChecker classes from SOLR.

```
<requestHandler name="/spell" class="solr.SearchHandler" startup="lazy">
<lst name="defaults">
  <str name="df">name_spell</str>
  <str name="spellcheck.dictionary">default</str>
  <str name="spellcheck.dictionary">wordbreak</str>
  <str name="spellcheck">on</str>
  <str name="spellcheck.extendedResults">true</str>
  <str name="spellcheck.count">100</str>
  <str name="spellcheck.alternativeTermCount">50</str>
  <str name="spellcheck.maxResultsForSuggest">50</str>
  <str name="spellcheck.collate">true</str>
  <str name="spellcheck.collateExtendedResults">true</str>
  <str name="spellcheck.maxCollationTries">100</str>
  <str name="spellcheck.maxCollations">50</str>
</lst>
<arr name="last-components">
  <str>spellcheck</str>
</arr>
</requestHandler>
```

```
<searchComponent name="spellcheck" class="solr.SpellCheckComponent">
<str name="queryAnalyzerFieldType">text_general</str>
<lst name="spellchecker">
<str name="name">default</str>
<str name="field">name_spell</str>
<str name="classname">solr.DirectSolrSpellChecker</str>
<str name="distanceMeasure">internal</str>
<float name="accuracy">0.3</float>
<int name="maxEdits">2</int>
<int name="minPrefix">1</int>
```

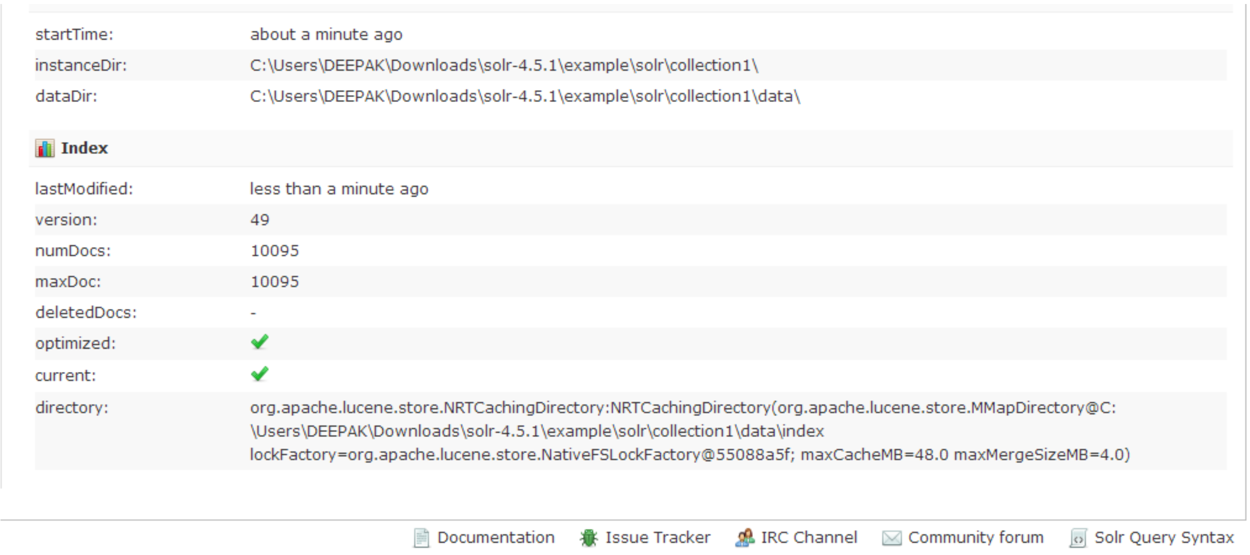
```
<int name="maxInspections">5</int>
<int name="minQueryLength">4</int>
<float name="maxQueryFrequency">0.01</float>
<float name="thresholdTokenFrequency">.01</float>
</lst>

<lst name="spellchecker">
<str name="name">wordbreak</str>
<str name="classname">solr.WordBreakSolrSpellChecker</str>
<str name="field">name_spell</str>
<str name="combineWords">true</str>
<str name="breakWords">true</str>
<int name="maxChanges">10</int>
</lst>
</searchComponent>
```

SOLR STATS

1) Index Size

We have indexed 10095 documents in the SOLR server.



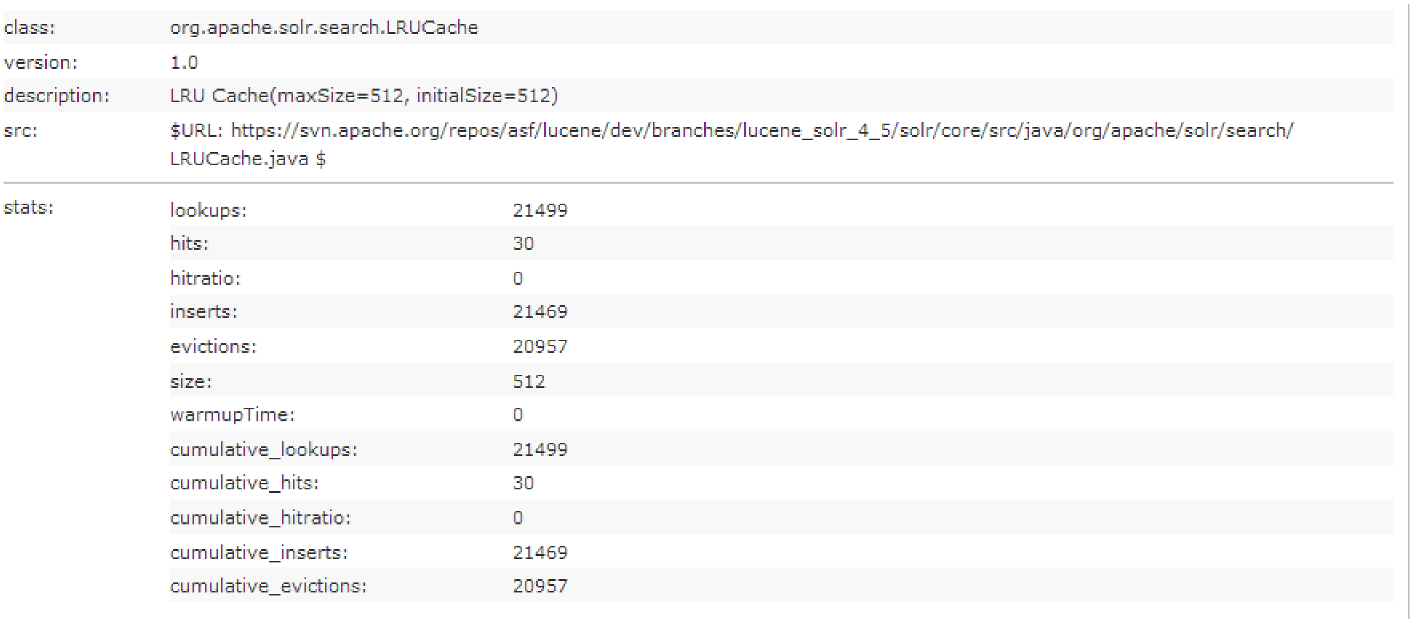
startTime:	about a minute ago
instanceDir:	C:\Users\DEEPAK\Downloads\solr-4.5.1\example\solr\collection1\
dataDir:	C:\Users\DEEPAK\Downloads\solr-4.5.1\example\solr\collection1\data\

Index

lastModified:	less than a minute ago
version:	49
numDocs:	10095
maxDoc:	10095
deletedDocs:	-
optimized:	✓
current:	✓
directory:	org.apache.lucene.store.NRTCachingDirectory:NRTCachingDirectory(org.apache.lucene.store.MMapDirectory@C:\Users\DEEPAK\Downloads\solr-4.5.1\example\solr\collection1\data\index lockFactory=org.apache.lucene.store.NativeFSLockFactory@55088a5f; maxCacheMB=48.0 maxMergeSizeMB=4.0)

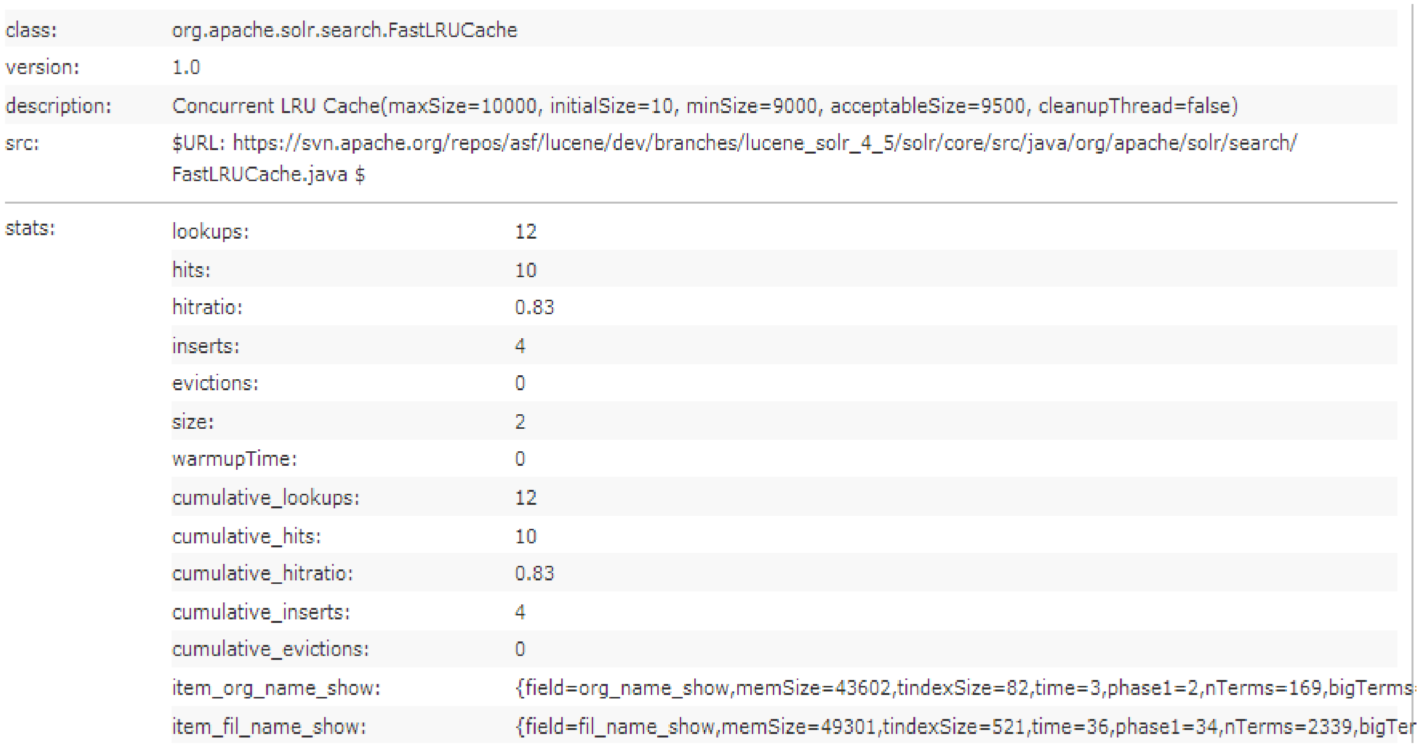
2) Stats on Caching

Document Cache:



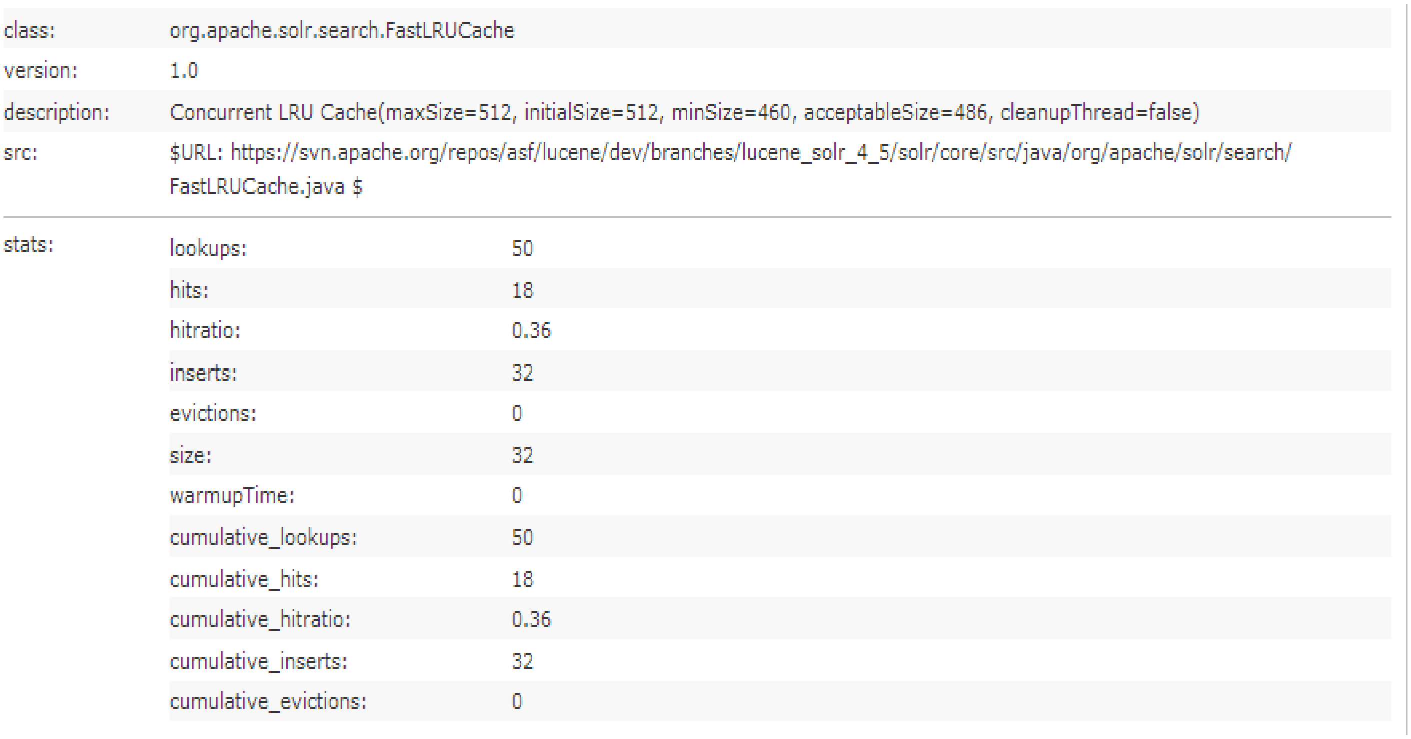
class:	org.apache.solr.search.LRUCache	
version:	1.0	
description:	LRU Cache(maxSize=512, initialSize=512)	
src:	$URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/search/LRUCache.java $	
stats:	lookups:	21499
	hits:	30
	hitratio:	0
	inserts:	21469
	evictions:	20957
	size:	512
	warmupTime:	0
	cumulative_lookups:	21499
	cumulative_hits:	30
	cumulative_hitratio:	0
	cumulative_inserts:	21469
	cumulative_evictions:	20957

Field Value Cache:



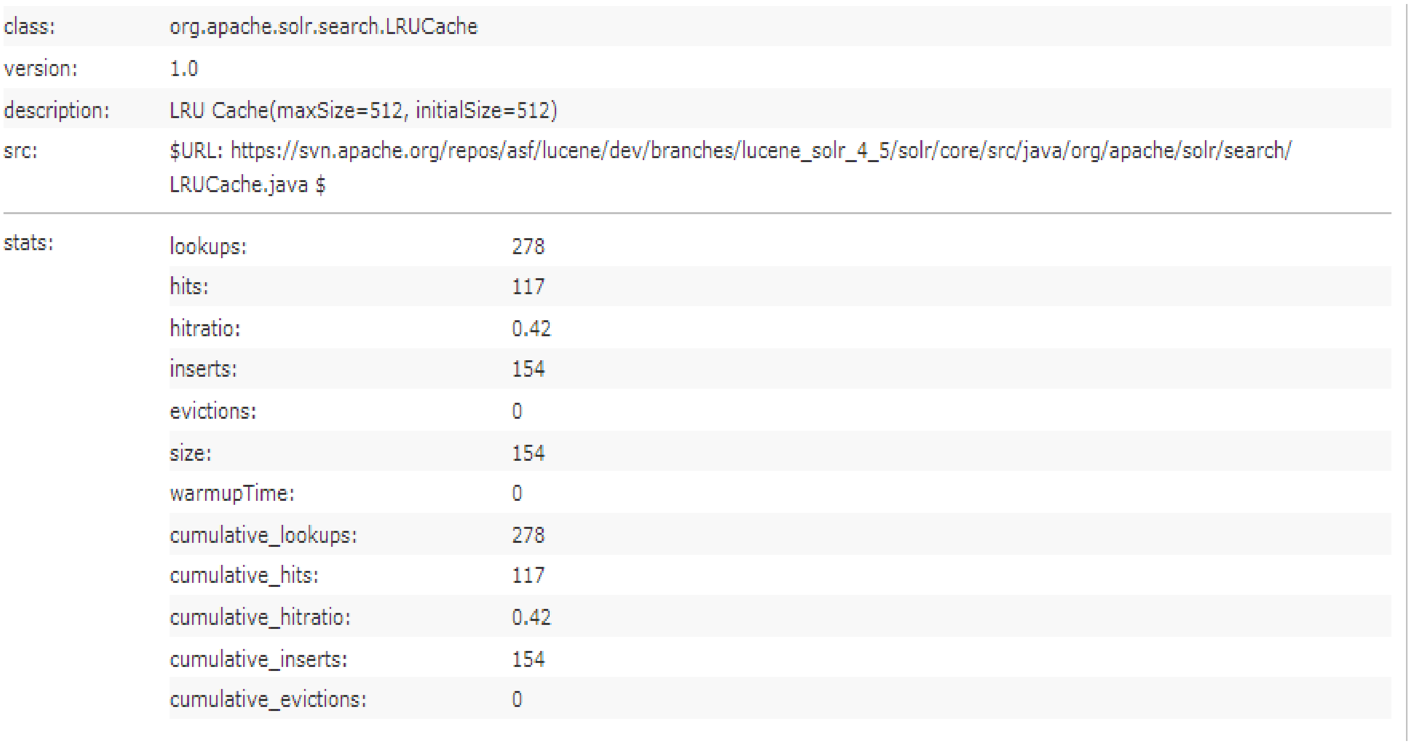
class:	org.apache.solr.search.FastLRUCache	
version:	1.0	
description:	Concurrent LRU Cache(maxSize=10000, initialSize=10, minSize=9000, acceptableSize=9500, cleanupThread=false)	
src:	$URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/search/FastLRUCache.java $	
stats:	lookups:	12
	hits:	10
	hitratio:	0.83
	inserts:	4
	evictions:	0
	size:	2
	warmupTime:	0
	cumulative_lookups:	12
	cumulative_hits:	10
	cumulative_hitratio:	0.83
	cumulative_inserts:	4
	cumulative_evictions:	0
	item_org_name_show:	{field=org_name_show,memSize=43602,tindexSize=82,time=3,phase1=2,nTerms=169,bigTerms
	item_fil_name_show:	{field=fil_name_show,memSize=49301,tindexSize=521,time=36,phase1=34,nTerms=2339,bigTer

Filter cache



class:	org.apache.solr.search.FastLRUCache	
version:	1.0	
description:	Concurrent LRU Cache(maxSize=512, initialSize=512, minSize=460, acceptableSize=486, cleanupThread=false)	
src:	$URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/search/FastLRUCache.java $	
stats:	lookups:	50
	hits:	18
	hitratio:	0.36
	inserts:	32
	evictions:	0
	size:	32
	warmupTime:	0
	cumulative_lookups:	50
	cumulative_hits:	18
	cumulative_hitratio:	0.36
	cumulative_inserts:	32
	cumulative_evictions:	0

Query Result Cache

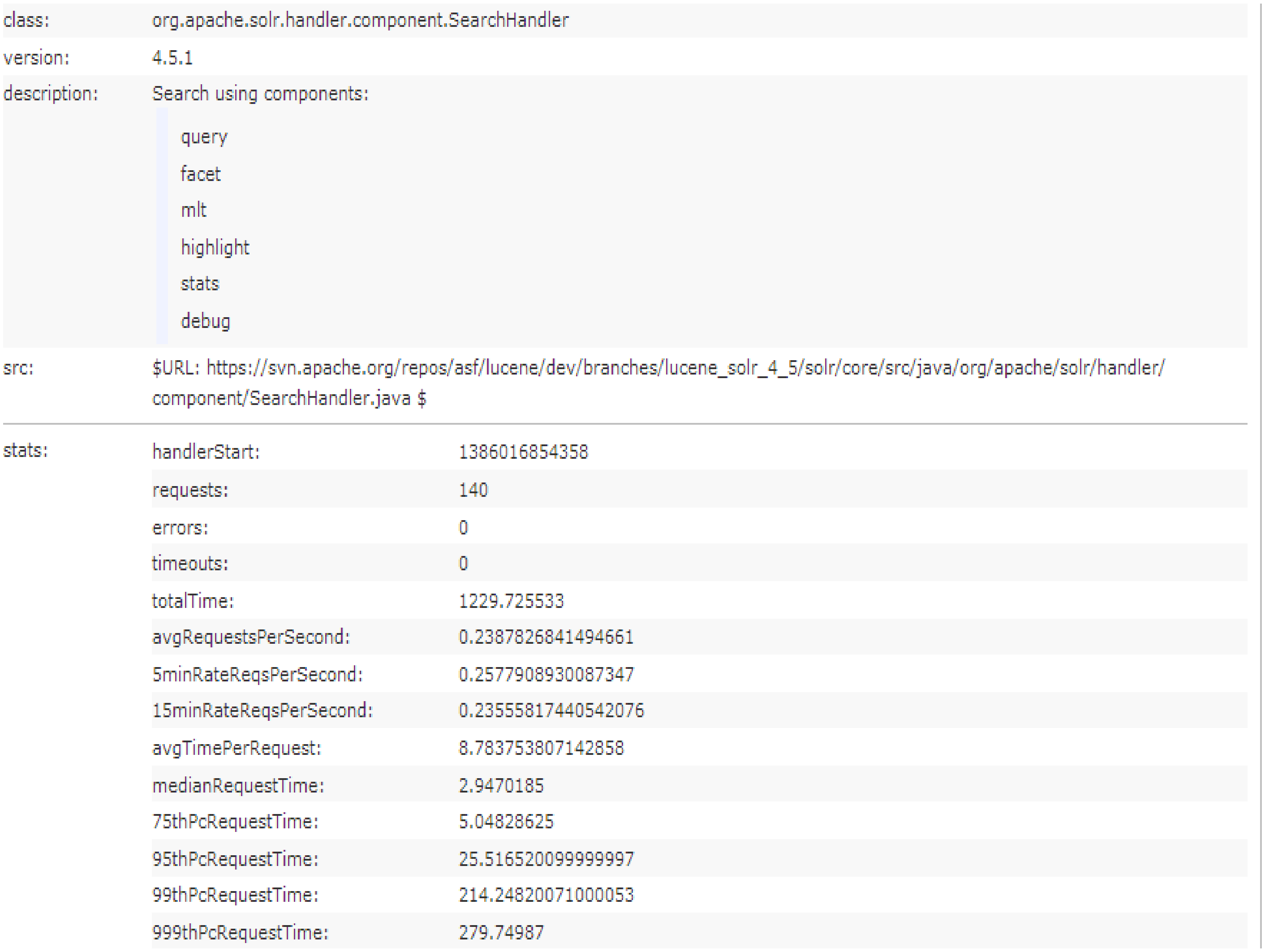


class:	org.apache.solr.search.LRUCache	
version:	1.0	
description:	LRU Cache(maxSize=512, initialSize=512)	
src:	$URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/search/LRUCache.java $	
stats:	lookups:	278
	hits:	117
	hitratio:	0.42
	inserts:	154
	evictions:	0
	size:	154
	warmupTime:	0
	cumulative_lookups:	278
	cumulative_hits:	117
	cumulative_hitratio:	0.42
	cumulative_inserts:	154
	cumulative_evictions:	0

3) Stats on Performance

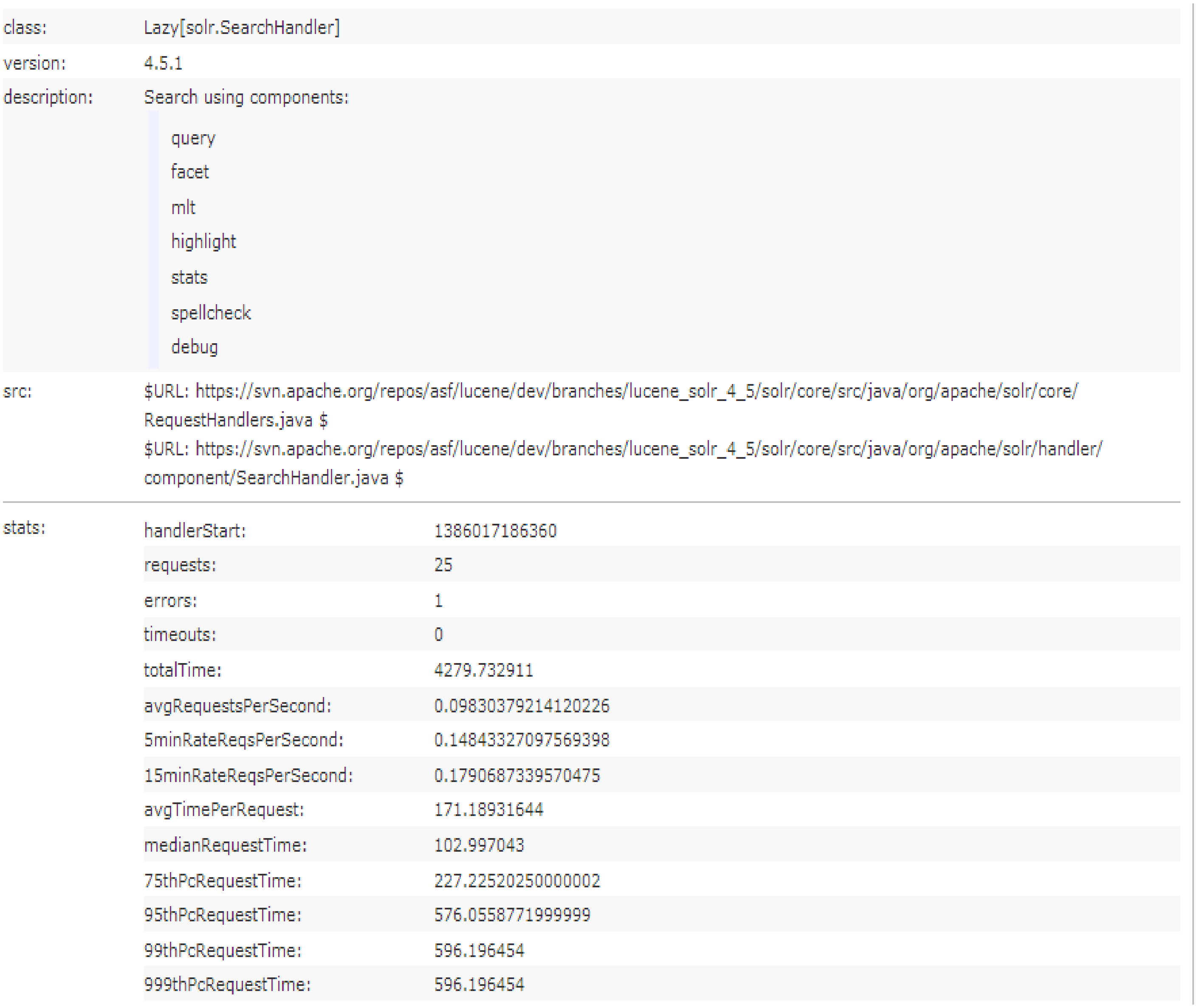
Response Time:

a) Default query handler



class:	org.apache.solr.handler.component.SearchHandler	
version:	4.5.1	
description:	Search using components: query facet mlt highlight stats debug	
src:	$URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/handler/component/SearchHandler.java $	
stats:	handlerStart:	1386016854358
	requests:	140
	errors:	0
	timeouts:	0
	totalTime:	1229.725533
	avgRequestsPerSecond:	0.2387826841494661
	5minRateReqsPerSecond:	0.2577908930087347
	15minRateReqsPerSecond:	0.23555817440542076
	avgTimePerRequest:	8.783753807142858
	medianRequestTime:	2.9470185
	75thPcRequestTime:	5.04828625
	95thPcRequestTime:	25.516520099999997
	99thPcRequestTime:	214.24820071000053
	999thPcRequestTime:	279.74987

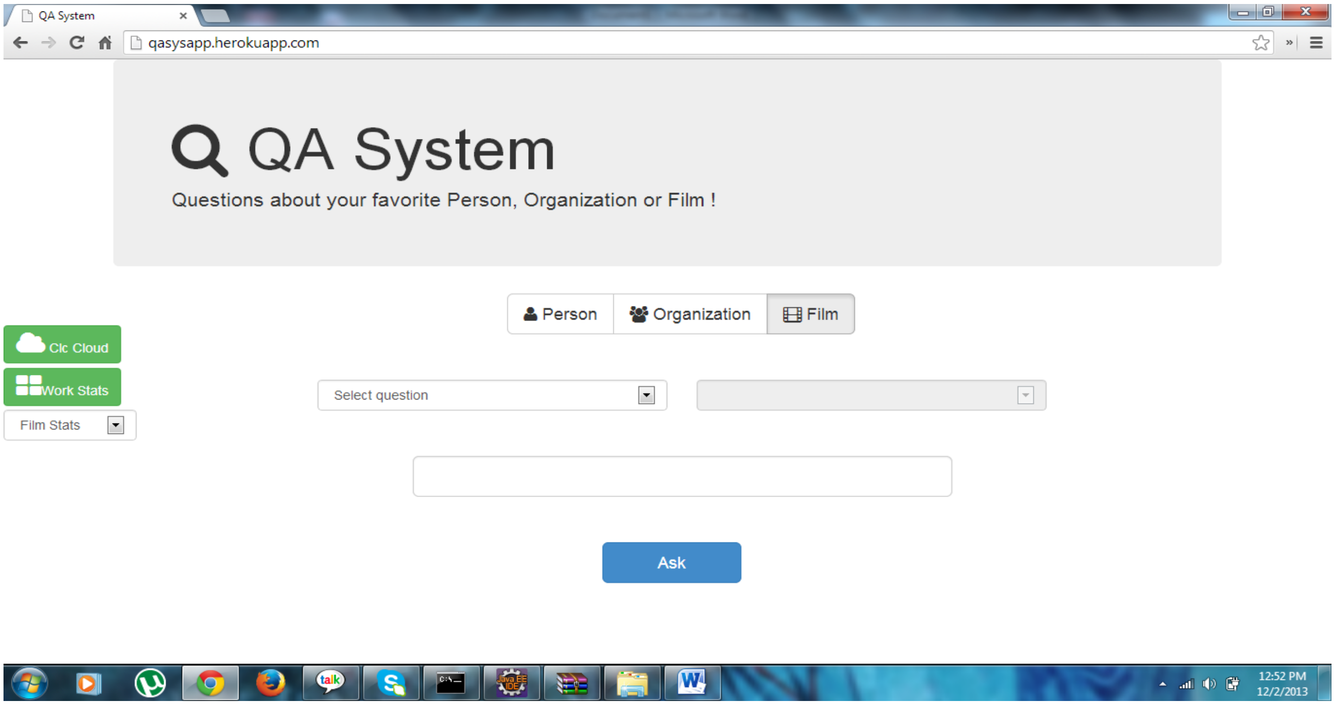
b) Spell Check handler



class:	Lazy[solr.SearchHandler]	
version:	4.5.1	
description:	Search using components: query facet mlt highlight stats spellcheck debug	
src:	$URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/core/RequestHandlers.java $ $URL: https://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_5/solr/core/src/java/org/apache/solr/handler/component/SearchHandler.java $	
stats:	handlerStart:	1386017186360
	requests:	25
	errors:	1
	timeouts:	0
	totalTime:	4279.732911
	avgRequestsPerSecond:	0.09830379214120226
	5minRateReqsPerSecond:	0.14843327097569398
	15minRateReqsPerSecond:	0.1790687339570475
	avgTimePerRequest:	171.18931644
	medianRequestTime:	102.997043
	75thPcRequestTime:	227.22520250000002
	95thPcRequestTime:	576.0558771999999
	99thPcRequestTime:	596.196454
	999thPcRequestTime:	596.196454

UI SCREENSHOTS:

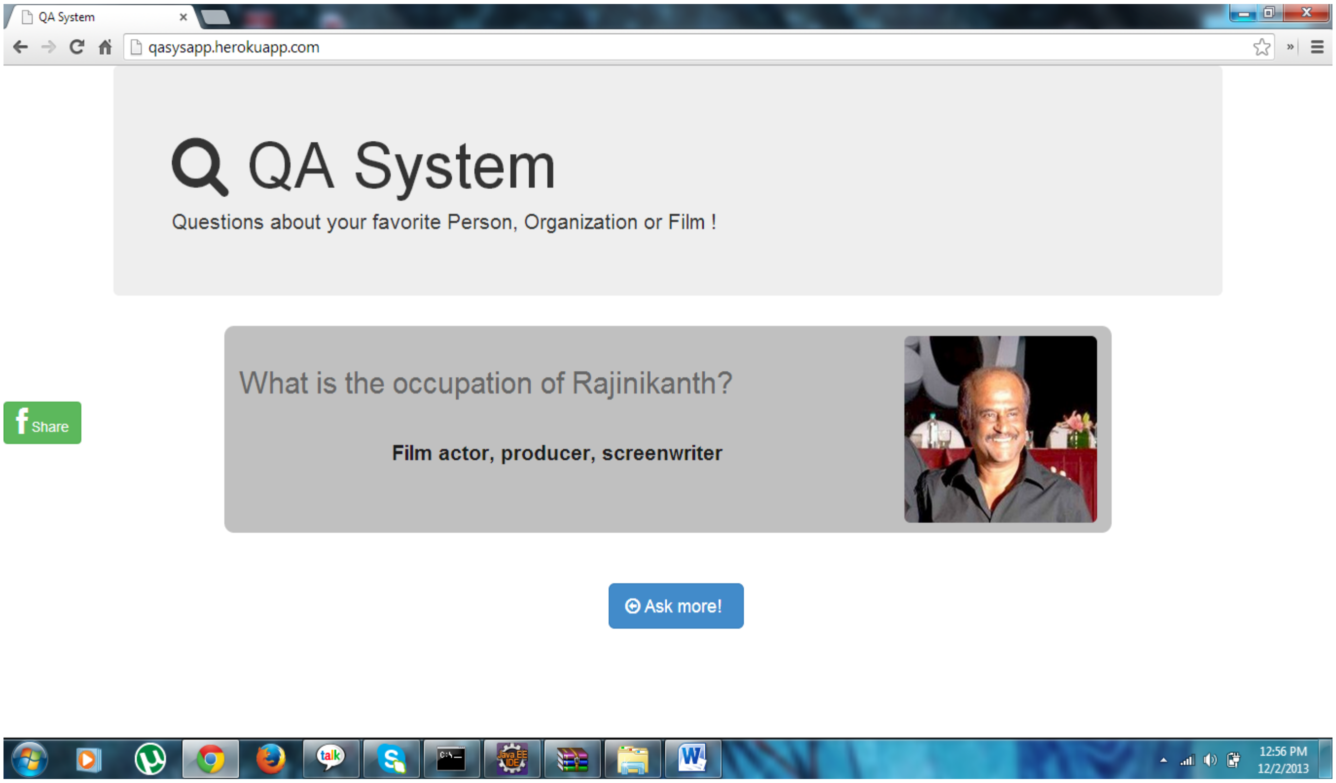
Basic System:



QA System

Questions about your favorite Person, Organization or Film !

Query regarding a Personality:



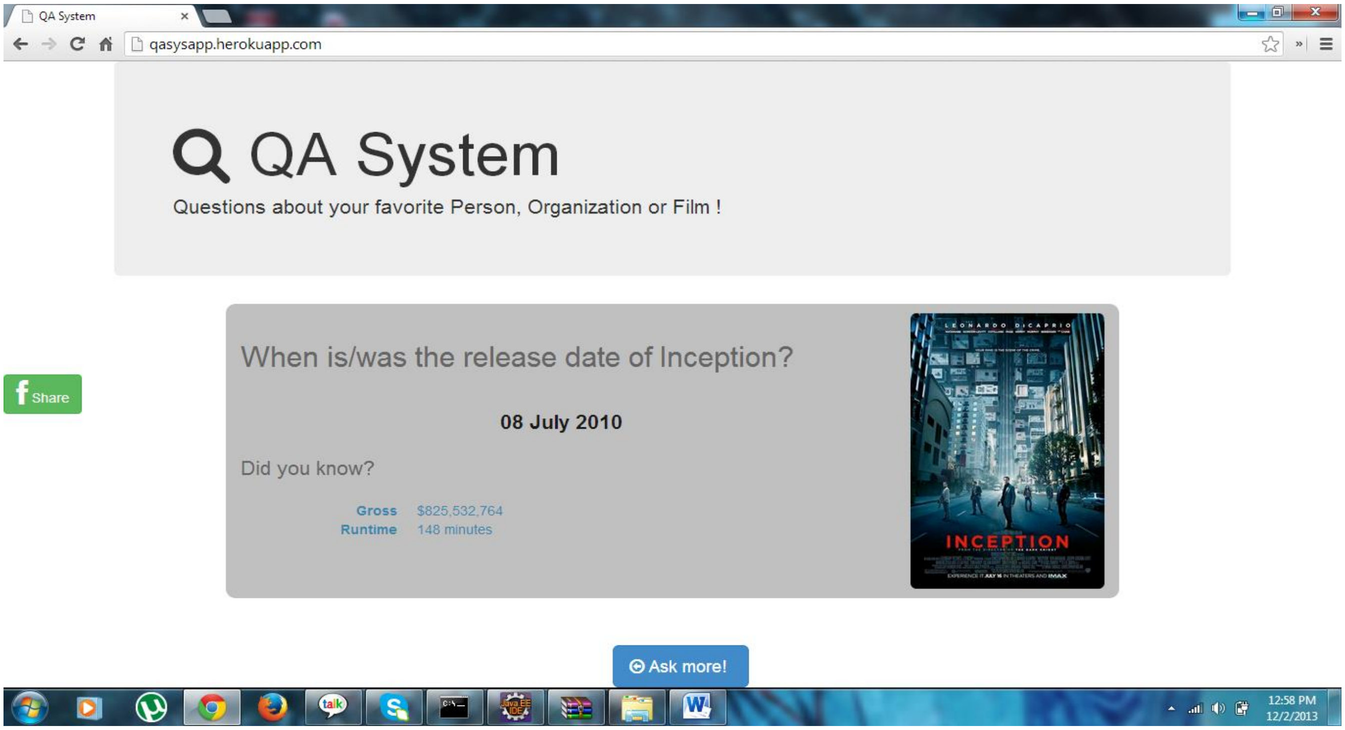
QA System

Questions about your favorite Person, Organization or Film !

What is the occupation of Rajinikanth?

Film actor, producer, screenwriter

Query about a Film:



QA System

qasysapp.herokuapp.com

QA System

Questions about your favorite Person, Organization or Film !

f Share

When is/was the release date of Inception?

08 July 2010

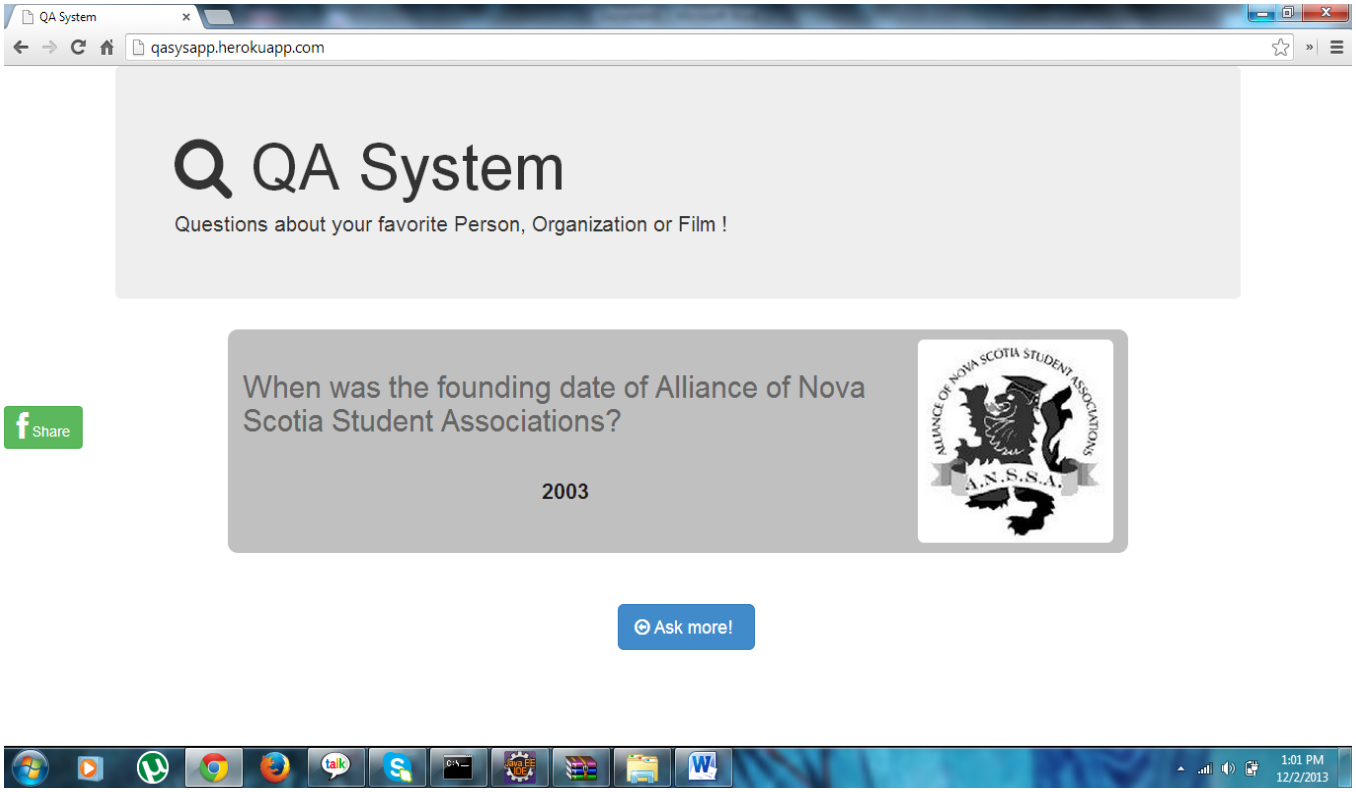
Did you know?

Gross	$825,532,764
Runtime	148 minutes

Ask more!

12:58 PM
12/2/2013

Query about an Organization:



QA System

qasysapp.herokuapp.com

QA System

Questions about your favorite Person, Organization or Film !

f Share

When was the founding date of Alliance of Nova Scotia Student Associations?

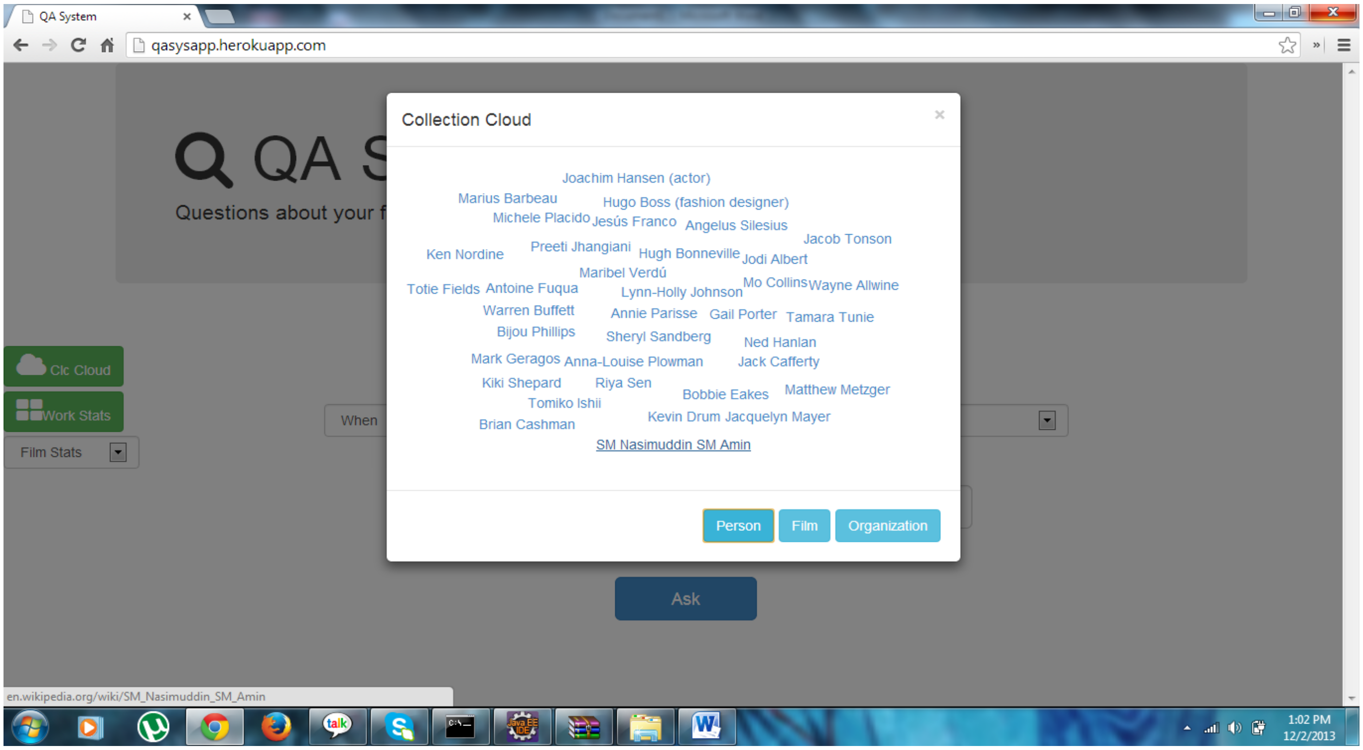
2003

Ask more!

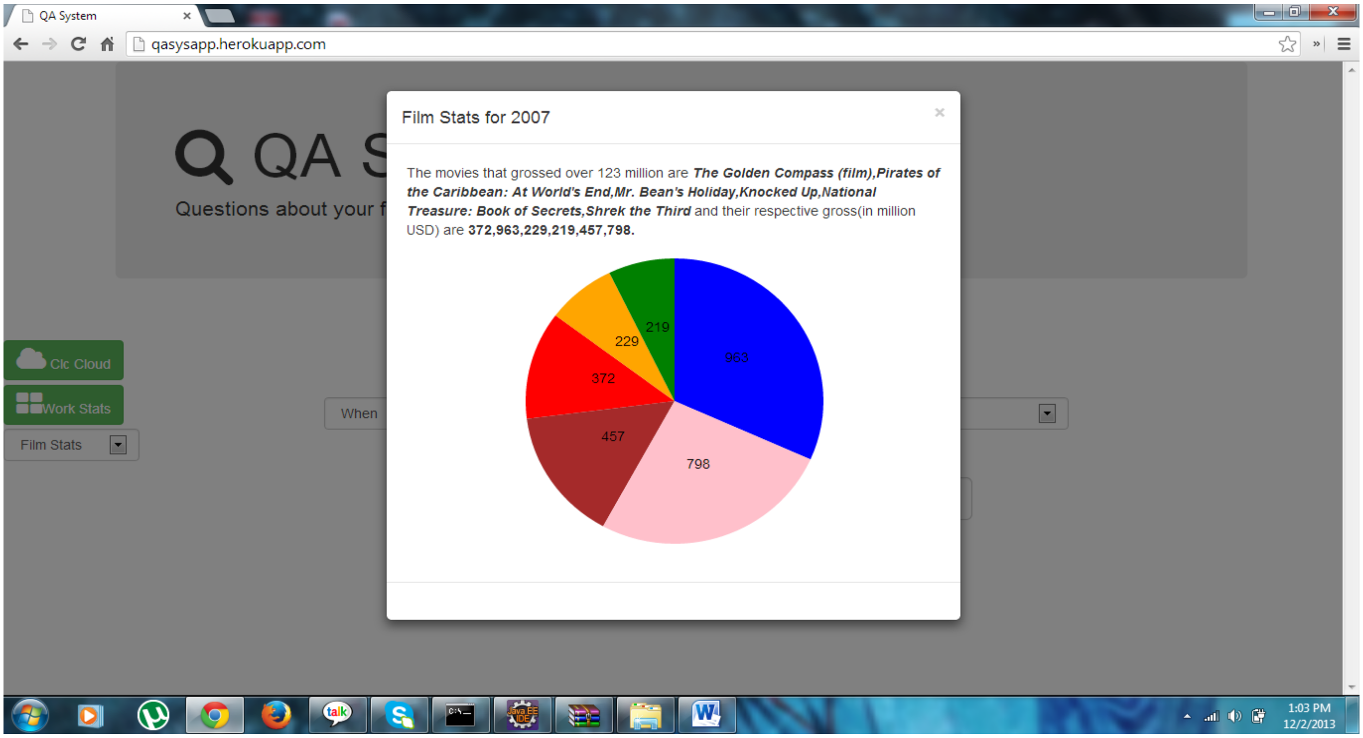
1:01 PM
12/2/2013

ADDITIONAL FEATURES (USP's):

Tag Cloud for easy navigation:



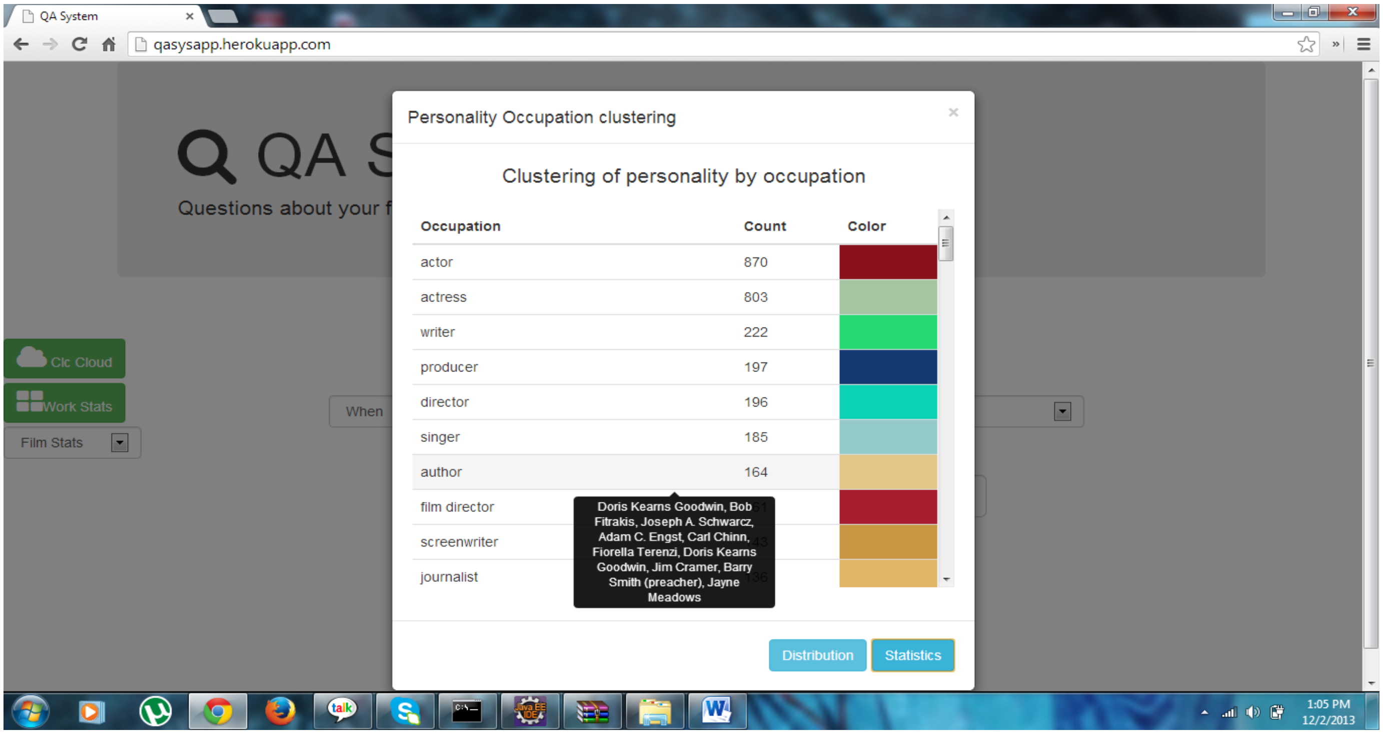
Dynamic Film Statistics represented graphically based on the year selected by the user:



Film Stats for 2007

The movies that grossed over 123 million are ***The Golden Compass (film),Pirates of the Caribbean: At World's End,Mr. Bean's Holiday,Knocked Up,National Treasure: Book of Secrets,Shrek the Third*** and their respective gross(in million USD) are **372,963,229,219,457,798.**

Clustering of Data:



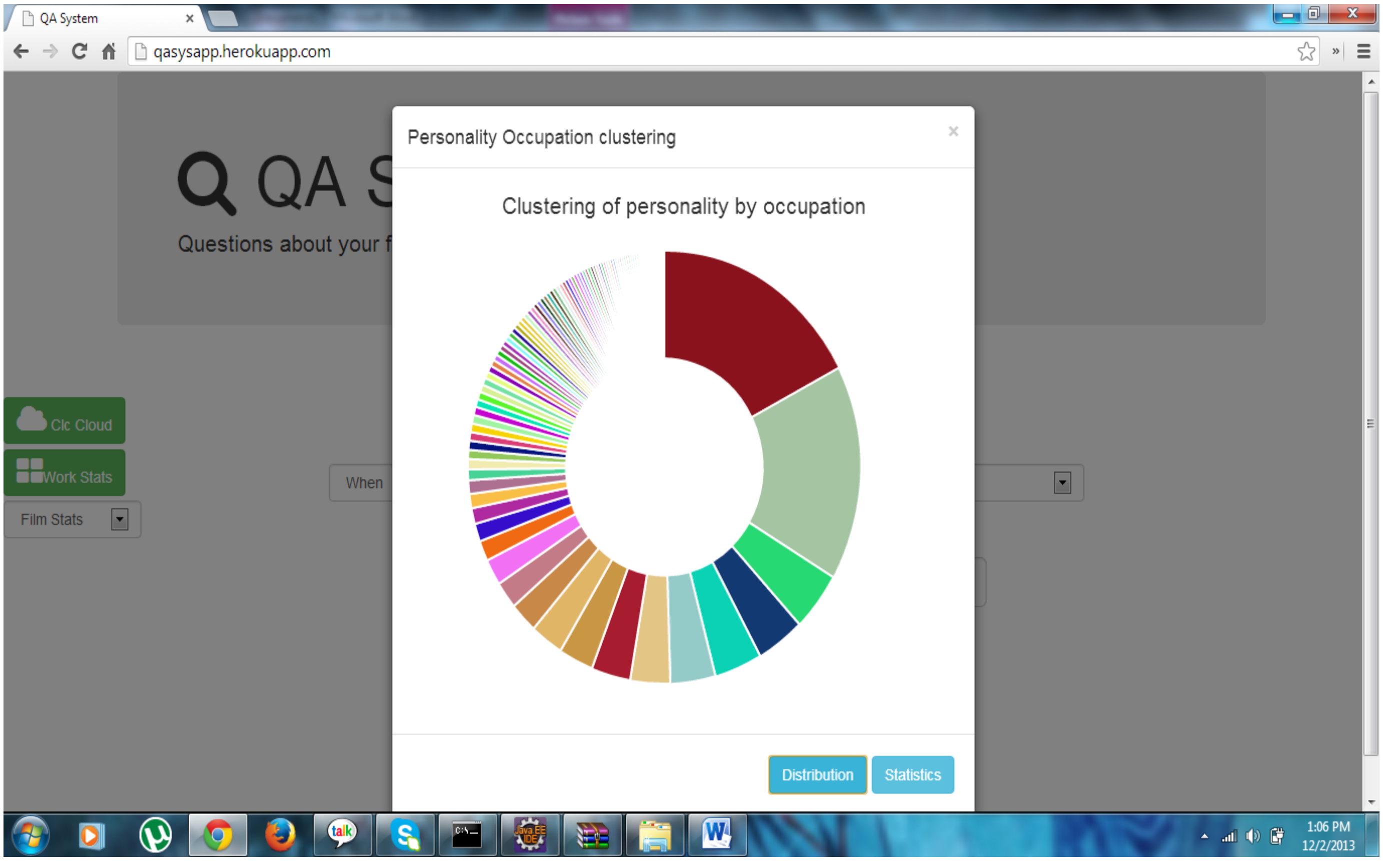
Personality Occupation clustering

Clustering of personality by occupation

Occupation	Count	Color
actor	870	🟥
actress	803	🟩
writer	222	🟩
producer	197	🟦
director	196	🟦
singer	185	🟦
author	164	🟨
film director	161	🟥
screenwriter	143	🟧
journalist	136	🟨

Doris Kearns Goodwin, Bob Fitrakis, Joseph A. Schwarcz, Adam C. Engst, Carl Chinn, Fiorella Terenzi, Doris Kearns Goodwin, Jim Cramer, Barry Smith (preacher), Jayne Meadows

Distribution Statistics

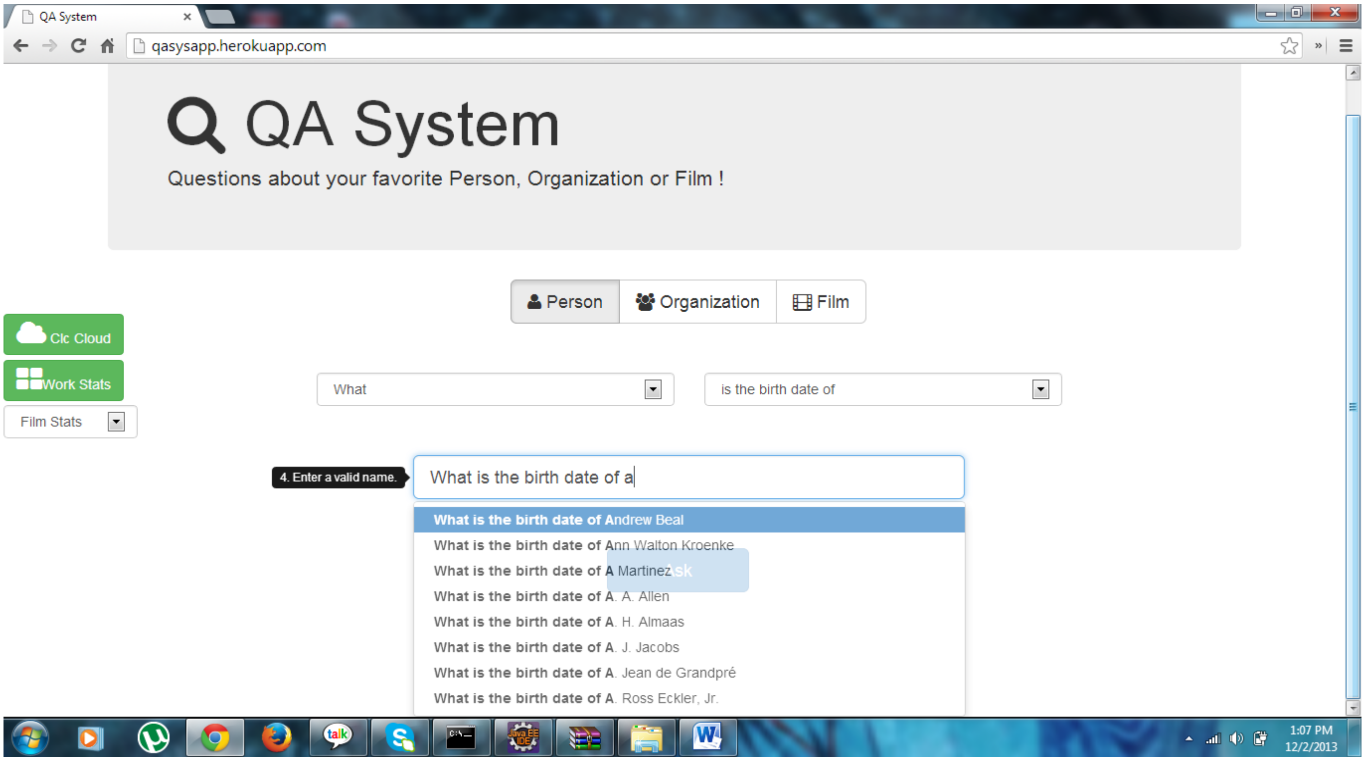


Personality Occupation clustering

Clustering of personality by occupation

Distribution Statistics

Auto Suggest:



QA System

Questions about your favorite Person, Organization or Film !

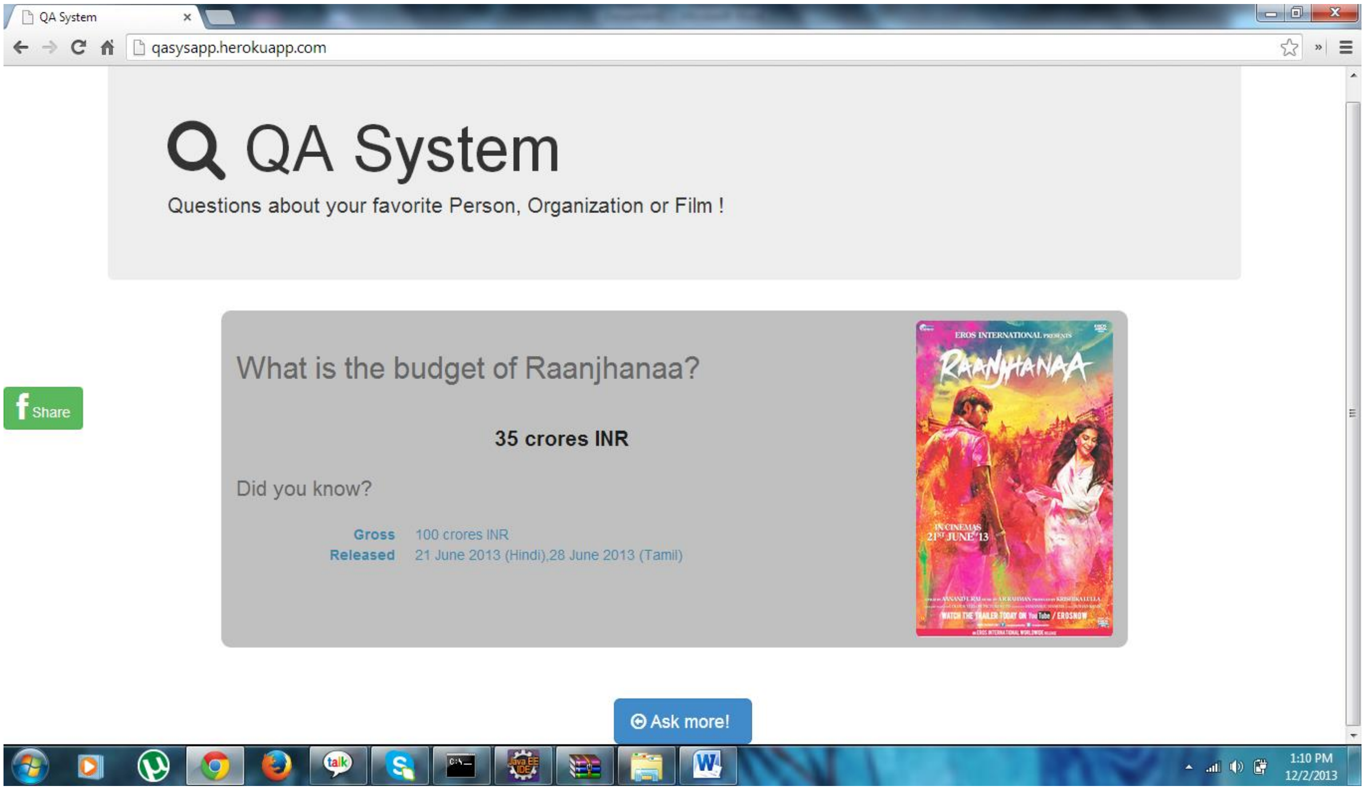
What

is the birth date of

What is the birth date of a

- What is the birth date of Andrew Beal
- What is the birth date of Ann Walton Kroenke
- What is the birth date of A Martinez
- What is the birth date of A. A. Allen
- What is the birth date of A. H. Almaas
- What is the birth date of A. J. Jacobs
- What is the birth date of A. Jean de Grandpré
- What is the birth date of A. Ross Eckler, Jr.

Image Retrieval:



QA System

Questions about your favorite Person, Organization or Film !

What is the budget of Raanjhanaa?

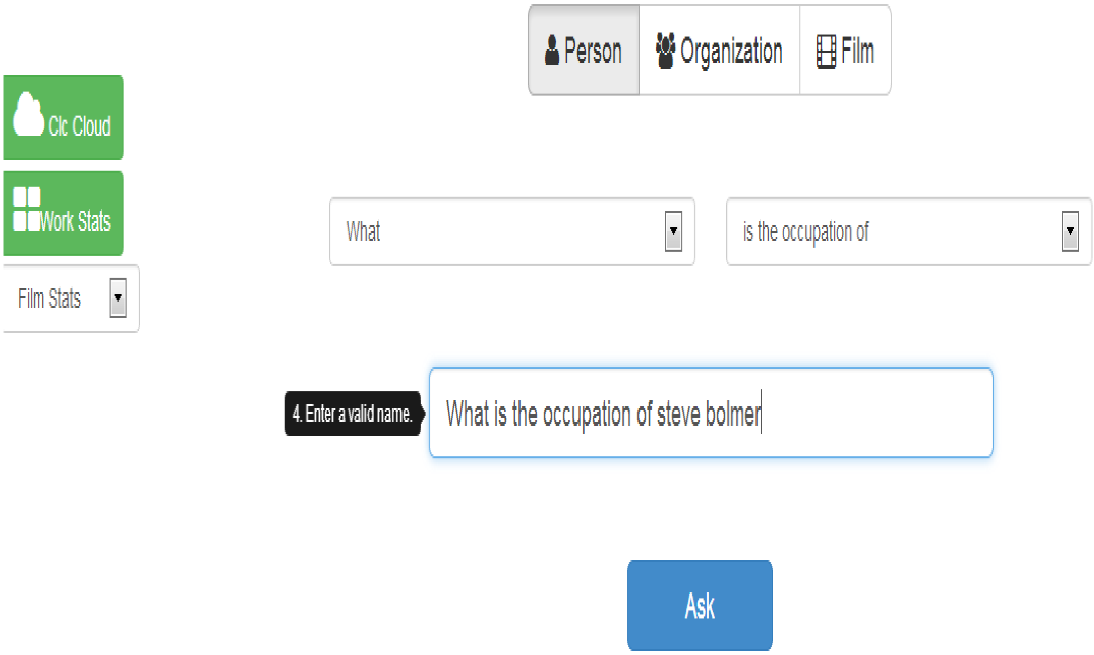
35 crores INR

Did you know?

Gross	100 crores INR
Released	21 June 2013 (Hindi),28 June 2013 (Tamil)

Automatic Spell Correction:

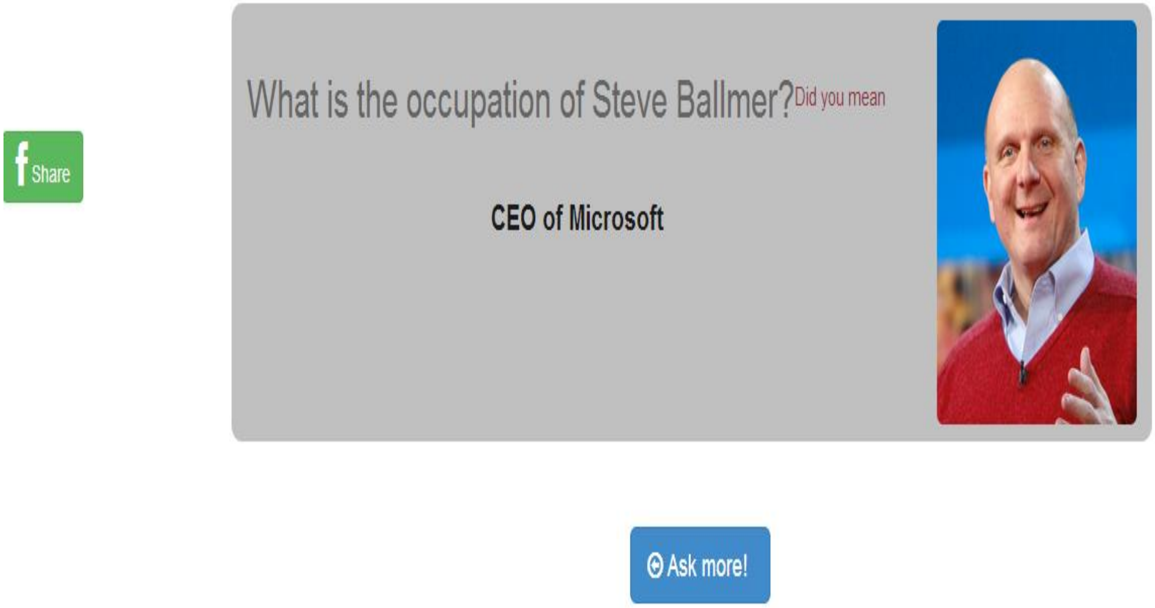
Spelling Mistake:



4. Enter a valid name.

What is the occupation of steve bolmer

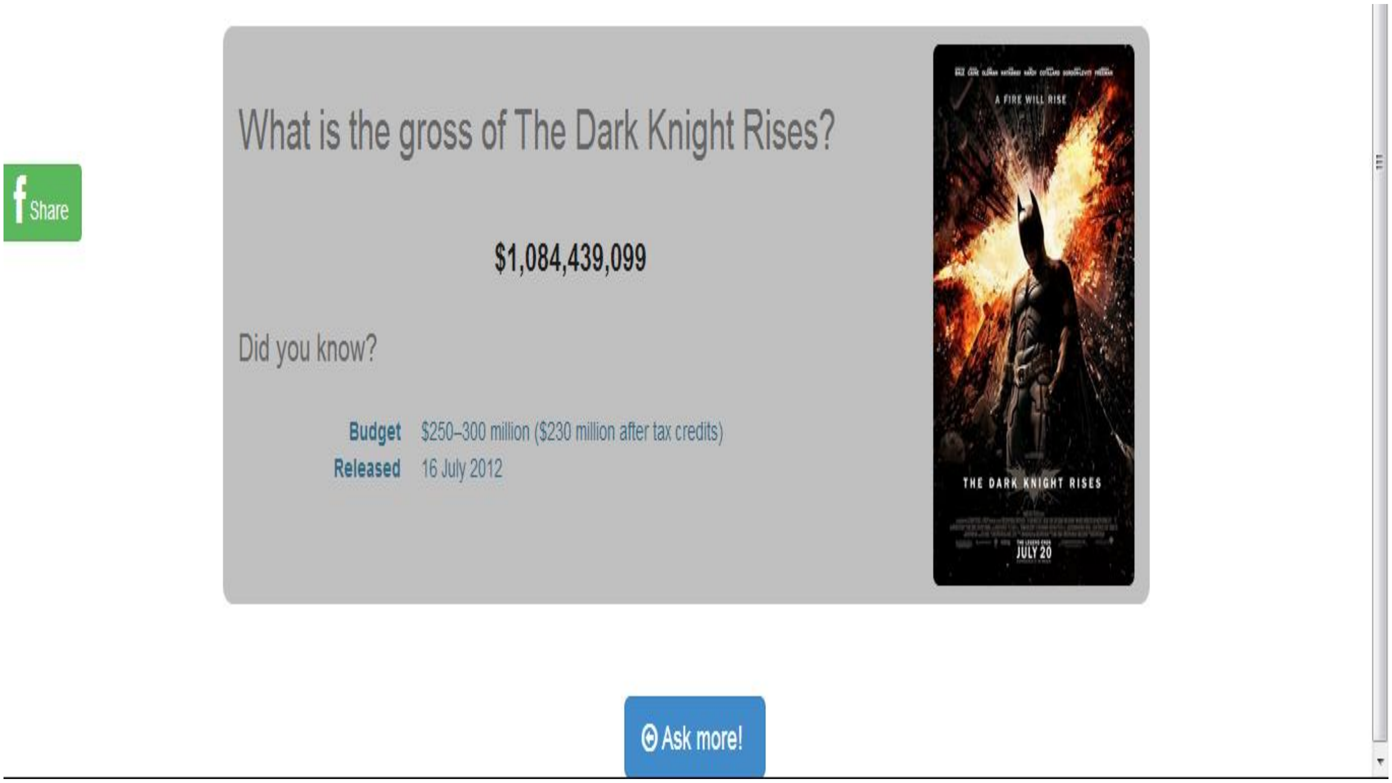
Corrected Version:



What is the occupation of Steve Ballmer?Did you mean

CEO of Microsoft

Related Trivia that varies depending upon the query:



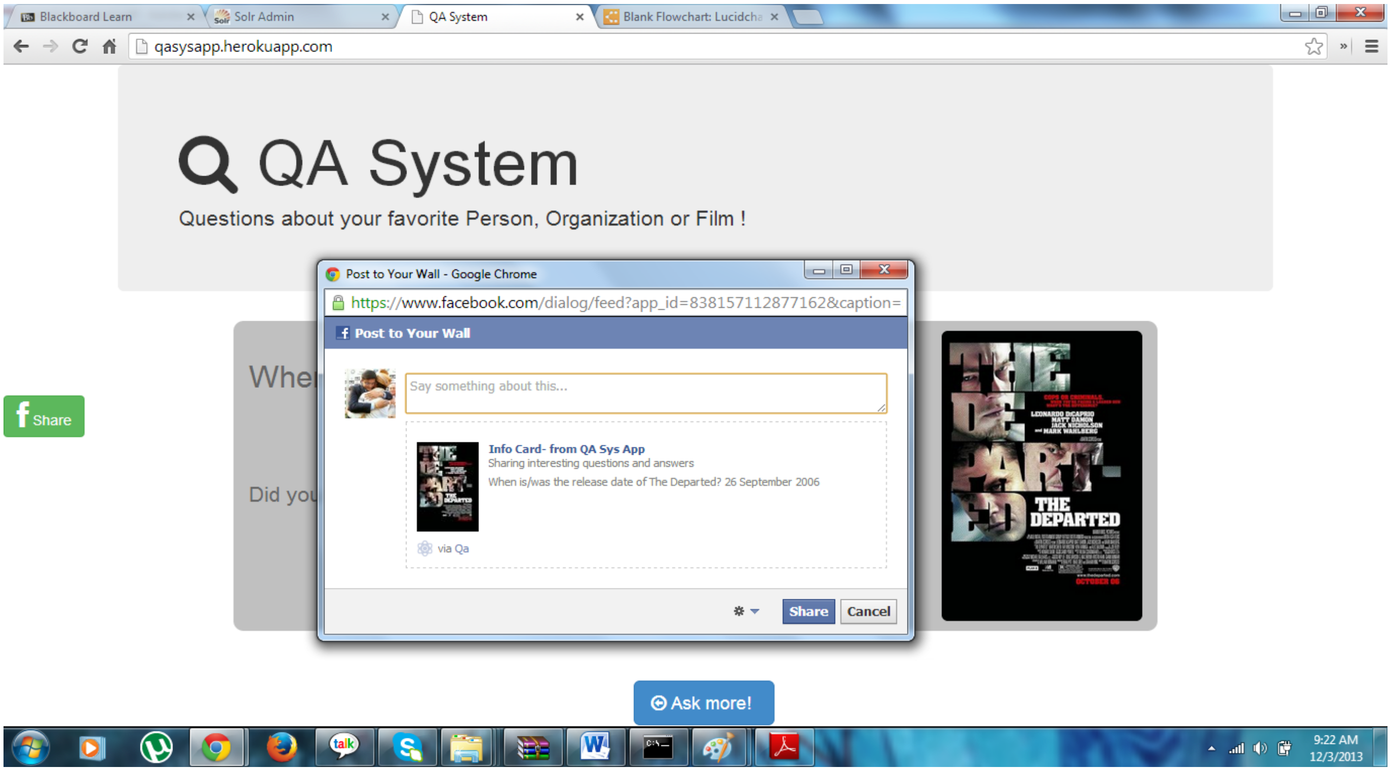
What is the gross of The Dark Knight Rises?

$1,084,439,099

Did you know?

Budget	$250–300 million ($230 million after tax credits)
Released	16 July 2012

Option to share on Facebook:



QA System

Questions about your favorite Person, Organization or Film !

https://www.facebook.com/dialog/feed?app_id=838157112877162&caption=

Info Card- from QA Sys App

Sharing interesting questions and answers

When is/was the release date of The Departed? 26 September 2006

FUTURE SCOPE:

- The system could be augmented to show the birthplace and residence of a personality on a map by using the latitude and longitude coordinates.
- The auto suggestion feature for current personalities could be ranked based on the number of followers they have on twitter. This would serve as a measure of their popularity.
- Natural language Processing (NLP) can be implemented in order to support free text queries.
- The system could be enhanced to support speech recognition based responses.
- The large amount of data available could be rendered more effectively on the front end.
- As of now we have indexed around 10000 documents that can be categorized broadly into three topics. The system is scalable and more topics and its related documents can be included to better the user experience.

INDIVIDUAL MEMBER CONTRIBUTION:

User Interface:

- *Aswin Bharadwaj Ramesh*
- *Deepak Ravishankar Ramkumar*

Solr Configuration:

- *Vivekanandh Vel Rathinam*

Parsing of data:

- *Aravindhan Thanigachalam*