**Brief Overview**

Apache Nutch is used to crawl links obtained from DMOZ data. The crawled data is pre-processed using Apache Lucene tokenizer and filters, resulting terms are used to build vector-space index. The application is setup as a Java web application using Apache Maven and packaged into a WAR file. Upon deployment, it is accessible at <http://localhost:8080/irproject/>. The application upon receiving a search query does the necessary pre-processing, then uses JWNL (Java WordNet library) to access the WordNet data for incorporating new terms for query expansion, then calculates the cosine similarity against the relevant documents and returns the sorted results to the user. The user is also provided the opportunity to submit relevance feedback for each of the relevant documents, which the application makes uses of in calculating both positive and negative feedback using Rocchio method, updates the query accordingly and generates new results to the user. There are two separate projects, the ‘IRProject’ which is the actual project that accepts queries and feedback against a collection of crawled web pages, and the second project ‘IRProjectEval’ which is cloned from the original project and edited to do the evaluation over test data. Both projects are accessible at following github links:

<https://github.com/blouke/IRProject>

<https://github.com/blouke/IRProjectEval>

**Note:** See deployment section for steps to download and deploy WAR files for these projects.

The IRProject is also hosted on an Amazon EC2 instance, accessible at

<http://52.88.63.246/irproject/>

The path needs to be specified completely including the trailing slash.

**Application Description**

The application is build using Apache Maven build utility. The pom.xml file contains all the dependencies. Description of different sections of the application follows.

**Parsing**

Source code package ir.parser contains two sets of classes meant for parsing data, both are invoked as command line java classes. One set of classes is used for parsing DMOZ data to get the links of interest in a file. This links file is fed to Apache Nutch which produces a Hadoop SequenceFile containing content of all the web pages crawled. The second set of parsing classes are used to parse the Hadoop SequenceFile into a text file containing the content of all web pages crawled along with their URLs. Parsing until now is done using command line java programs contained in the ir.parser package. Final parsing is done while building the index using a custom regex and using Lucene tokenizer and filter classes.

**Indexing**

Source code package ir.indexer contains classes for building the index. Lucene tokenizer and filter classes are used for removing stop words and stemming the terms using Porter stemmer. The document vector length is calculated, and the IDF for each term is also created. Along with the document vector length, its URL and a snippet of first 200 characters are also stored. The index does not store TFxIDF weights for term occurrences, which are actually calculated at query processing time. The index contains raw count frequencies. All classes in this package implements the Serializable interface so that the index can be serialized to the disk.

**Query Processing**

Source code package ir.query contains classes for processing a query against the index along with processing of relevance feedback if provided. Each search query is processed in the same way as the web page documents were processed during index creation. Each query is parsed, stop words removed, and stemmed using Porter stemmer using Apache Lucene Tokenizer and StopFilter classes. Then JWNL (Java WordNet library) is used along with WordNet data to expand on query terms. Word sense disambiguation is not accommodated for, so words corresponding to SynSets of all word senses obtained from WordNet are incorporated into the query vector. The IDF for each query term is set to 1. At this point, the TFxIDF weights for terms in the relevant documents are calculated. The query vector is then used to calculate the cosine similarity against the set of relevant documents. The set of relevant documents is then sorted based on cosine similarity score, the ranked documents are presented to the user along with their URL, description and the similarity score. The interface allows the users to submit relevance feedback to the system which then employs Rocchio method with beta value of 0.5 for positive feedback and gamma value of 0.1 for negative feedback. The original query vector is then augmented with the results of the Rocchio method. The new query is then used to calculate cosine similarity against a new set of relevant documents, which are then sorted and presented to the user.

**Evaluation**

The test data used for system evaluation is accessible at <http://web.eecs.utk.edu/research/lsi/corpa.html>. The test data contains 423 Times Magazine Articles (from 1963). The evaluation is done in a separate project named ‘IRProjectEval’. This project can be run to generate evaluation results, but an html document generated by the system ‘evaluation\_results.html’ containing the evaluation results is also included in the submission. The evaluation is done by retrieving relevant documents by running each test data query against the system. Multiple similarity measures are used to measure system performance namely Cosine, Jaccard and Dice similarity measures. A threshold of 0.5 is used for all similarities and only documents having similarity score more than 0.5 are deemed similar by the system. The evaluation page result.jsp displays a table containing precision, recall and f1-measure for each query in the test data along with the number of documents retrieved by the system and the actual test data judgments. The table displays only the results using the cosine similarity measure because it produced the better results among all the similarity measures used. The evaluation page also displays Mean Absolute Error for all three similarity measures which are:

**Mean Absoulte Error**

Cosine similarity :2.892

Dice similarity :5.361

Jaccard similarity :8.904

**Application Front Controller that serves HTTP requests**

Source code package ir.controller contains only one file SearchController.java, a servlet which intercepts all HTTP requests and runs the query against the index by passing the query to the QueryProcessor.java in package ir.query. SearchController is responsible for creating the index and serializing/deserializing it along with other initializations and also for routing the HTTP requests to the appropriate view, the JSP pages.

**Instructions for Application deployment.**

The main application is created as a java web application and packaged into a WAR file.

The size of the source code for each project is rather small for each project around only 2MB. The following links point to IRProject.war (original project serves queries against crawled web pages) and IRProjectEval.war (original project modified to do evaluation on test data) files for both projects ready to be deployed on Tomcat container.

<https://github.com/blouke/IRProject/blob/master/irproject.war>

<https://github.com/blouke/IRProjectEval/blob/master/irprojecteval.war>

**Required configuration before deployment**

The IRProject makes use of WordNet by using JWNL (Java WordNet Library) which expects a properties.xml file to point to the actual absolute path to the WordNet directory. The properties.xml file in IRProject.war file needs to be edited, and the value attribute of the following element needs to point to the actual WordNet directory which may be different based on Tomcat directory.

<param name="dictionary\_path" value=”specify wordnet directory path here, it should end with …tomcat\webapps\irproject\wordnet\dict”/>

This path is system dependent and will be different for Linux. This properties file can be edited before deploying it to Tomcat by editing it in the war file, in the packaged war file the properties.xml is located at /WEB-INF/classes/properties.xml. The war file is just an archive file, it could be opened and edited by 7zip program.

The application makes use of Java Servlets to serve HTTP requests for searching queries, which requires the application to be deployed on any Java Servlet container or an JEE Application server. Tomcat was chosen for testing the application during development. Following instructions pertain to Tomcat and $TOMCAT\_HOME refers to the Tomcat installation directory. In order to deploy the application to Tomcat, simply copy the WAR file into the $TOMCAT\_HOME/webapps/ folder and start Tomcat by invoking $TOMCAT\_HOME/bin/startup.bat (for Microsoft Windows). Point the browser to <http://localhost:8080/irproject/> to start using the application. Tomcat unpacks the WAR file into a directory by the same name resulting in $TOMCAT\_HOME/webapps/irproject/, this directory will contain the dictionary and the vector-space Index in a Java serialized binary format file named ‘index.ser’. Upon initial deployment of the application, the Controller servlet SearchController.java in package ir.controller builds the index and persist it to the disk as index.ser which may take about a minute. Any subsequent invocation of the application will make use of the index.ser rather than creating the index again from scratch. In order to force the application to re-create the index, stop Tomcat and delete the file $TOMCAT\_HOME/webapps/index.ser.

**Navigating the Application**

The application contains two publicly accessible JSP pages index.jsp and result.jsp along with a stylesheet at src/main/webapp/ folder. The index.jsp is the main landing page where search queries are to be submitted. The results of the search query are presented in result.jsp. The result.jsp presents the list of all relevant documents along with a checkbox to specify the relevance of particular documents in order to provide relevance feedback to the application which uses Rocchio method to accommodate both positive and negative relevance feedback.

**Preliminary steps**

1. DMOZ open directory project data dump file <http://rdf.dmoz.org/rdf/content.rdf.u8.gz> is downloaded for the purpose of collecting seed URLs. The format of the file is RDF-like and not strict RDF. A parsing program based on custom regex, consisting of files DMOZParser.java and RDFHandler.java contained in package ir.parser is used to parse DMOZ data. The output of the command line program is a file src/main/resources/links.txt that contains all the URLs from DMOZ data in relevance to the topic ‘Top/Shopping/Clothing/Casual’.
2. The file links.txt containing 296 seed URLs one per line is injected to Apache Nutch. File src/main/resources/nutch-commands.txt contains the list of commands and instructions employed to inject, fetch and parse URL content. Few decisions were made with respect to crawling in order to limit the collection of data set. Nutch property ‘db.ignore.external.links’ was set to ‘true’ in order to avoid accumulating URLs outside the original URLs domains. Only two rounds of Nutch fetch and parse commands were used after updating the crawl database which contains the links to crawl. A brief description follows.

Nutch 1.x Crawling Steps

# INJECT the seed URLs

bin/nutch inject crawl/crawldb urls

# FIRST ROUND

bin/nutch generate crawl/crawldb crawl/segments

export SEGMENT=`ls -d crawl/segments/2\* | tail -1`

bin/nutch fetch $SEGMENT

bin/nutch parse $SEGMENT

bin/nutch updatedb crawl/crawldb $SEGMENT

# SECOND ROUND

bin/nutch generate crawl/crawldb crawl/segments

export SEGMENT=`ls -d crawl/segments/2\* | tail -1`

bin/nutch fetch $SEGMENT

bin/nutch parse $SEGMENT

bin/nutch updatedb crawl/crawldb $SEGMENT

#merge the segments

bin/nutch mergesegs crawl/merge -dir crawl/segments/

#dump only parsed text of each url from all segments.

bin/nutch readseg -dump crawl/merge/\* crawl/dump -nocontent -nofetch -nogenerate -noparse –noparsedata

Note:

Set Nutch property fetcher.max.crawl.delay to 5 seconds to ignore pages beyond this much delay.

1. The execution of Nutch commands above results in a Hadoop SequenceFile which is fed into another command line parsing program NutchSegmentParser.java contained in package ir.parser. The output is a text file src/main/recources/dump which contains the URLs along with their content which still needs to be parsed to extract URLs and content in a meaningful format.