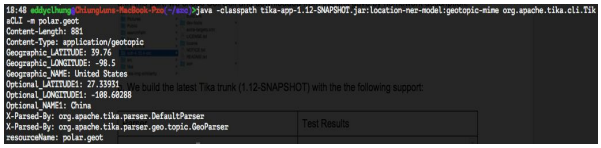
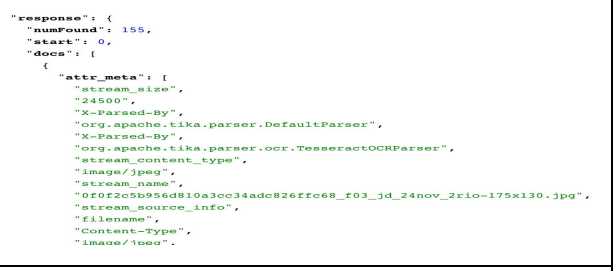
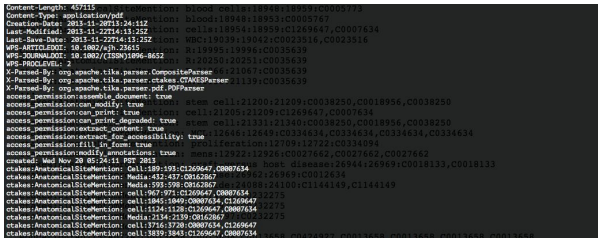
# CSCI572 Hw2 Report Team17

**1. Develop an indexing system using Apache Solr and its ExtractingRequestHandler ("SolrCell") or using Elastic Search and Tika-Python.**

a. In this part, we chose SolrCell and downloaded from the link gave by homework description. In order to support the feature of GeoParser, OCR and cTAKES, we need to make sure that we have upgraded SolrCell. For upgrade SolrCell, we need to edit file from the path solr-4.10.4/lucene/ivy-versions.properties which enable you to build and upgraded to the latest version. From the path solr-4.10.4/solr/contrib/extraction/lib, we can check the upgraded new jars with the following:

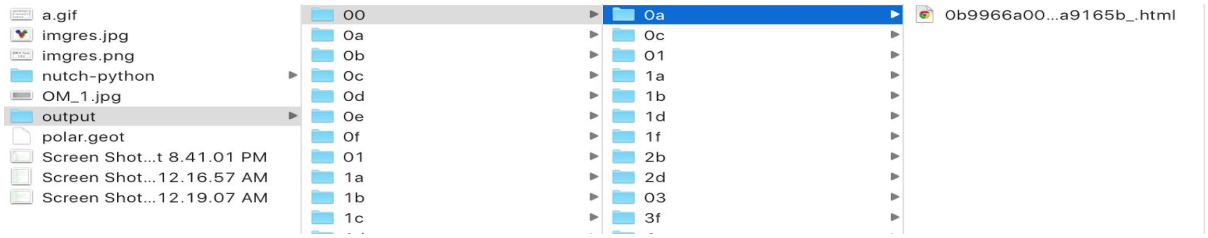
| Old version | New Version |
|---|---|
| metadata-extractor-2.6.2.jar | metadata-extractor-2.8.1.jar |
| commons-compress-1.7.jar | commons-compress-1.10.jar |
| tika-core-1.5.jar | tika-core-1.11.jar |
| tika-parsers-1.5.jar | tika-parsers-1.11.jar |
| tika-xmp-1.5.jar | tika-xmp-1.11.jar |

b. We build the latest Tika trunk (1.11-SNAPSHOT) with the the following support:

| Parser | Test Results |
|---|---|
| GeoTopic parsing |  |
| OCR |  |
| cTAKES / UIMA |  ..... |

| | |
|---|---|
| |  |

c. We execute command “bin/nutch dump” to dump data as input source for solr from the crawled data of assignment #1. Then, we write a bash file named “mypost.sh” for import dump data to solr.




**2.Leverage the Nutch indexing system to build up an Apache Solr index.**

a. Compare the metadata extracted from using Tika in Nutch during crawling upstream compared to your SolrCell based Tika run generated in Task #2.

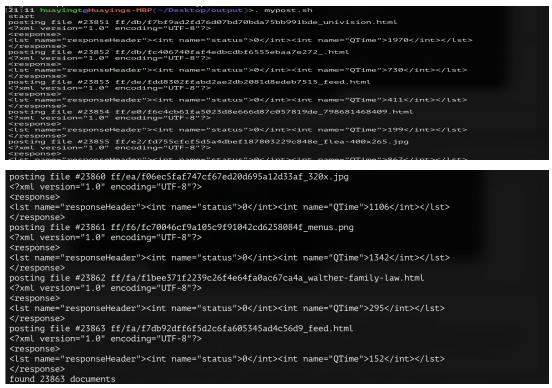
b. What did Tika extract in Nutch compared to what SolrCell extracts?

c. Describe the indexing process – what was easier – Nutch/Tika + SolrIndexing; or SolrCell?

The following results show more details for above questions.

| | Nutch/Tika + Solrindexing | SolrCell |
|---|---|---|
| Extracted Metadata | Metadata: Transparency Alpha=nonpremultipled PLTE PLTEEntry=index=0, red=255, green=255, blue=255 PLTE PLTEEntry=index=1, red=0, green=0, blue=0 tiff:ImageLength=1 Compression CompressionTypeName=deflate Chroma Palette PaletteEntry=index=0, red=255, green=255, blue=255, alpha=0 Chroma Palette PaletteEntry=index=1, red=0, green=0, blue=0, alpha=255 Data BitsPerSample=1 1 1 1 Data PlanarConfiguration=PixelInterleaved IHDR=width=1, height=1, bitDepth=1, colorType=Palette, compressionMethod=deflate, filterMethod=adaptive, interlaceMethod=none Chroma ColorSpaceType=RGB tRNS tRNS_Palette tRNS_PaletteEntry=index=0, alpha=0 tiff:BitsPerSample=1 1 1 1 Content-Type=image/png height=1 Dimension PixelAspectRatio=1.0 Compression NumProgressiveScans=1 Chroma BlackIsZero=true Compression Lossless=true width=1 Dimension ImageOrientation=Normal tiff:ImageWidth=1 Chroma NumChannels=4 Data SampleFormat=Index | "attr_meta": [<br>"stream_size",<br>"92793",<br>"X-Parsed-By",<br>"org.apache.tika.parser.DefaultParser",<br>"X-Parsed-By",<br>"org.apache.tika.parser.html.HtmlParser",<br>"stream_content_type",<br>"application/octet-stream",<br>"stream_name",<br>"0e575f0163d06ab211ea2a73eeb56756_jquery.min.js",<br>"stream_source_info",<br>"filename",<br>"Content-Encoding",<br>"UTF-8",<br>"Content-Type",<br>"text/html; charset=UTF-8" |
| Indexing Process | 1.Take the crawl data that we crawled from assignment #1 as input.<br>2.Before indexing we need to invert all of the links first, so that we may index incoming anchor text with the pages.<br>3.Then, we need to make sure solr in integrate with our nutch.<br>4.Finally, we use command “bin/nutch solrindex <solr url> <crawldb> [-linkdb <linkdb>][-params k1=v1&k2=v2...] ( ...| -dir <segments>) [-noCommit] [-deleteGone] [-filter] [-normalize]” to send our data to solr. | 1.Take the crawl data that we crawled from assignment #1 and use command bin/nutch dump to dump those data such as html or image files as input.<br>2.Run a bash file with code “curl "$URL?literal.id=doc-$RANDOM$RANDOM&uprefix=attr_&fmap.content=attr_content&commit=true" -F filename=@$f<br>” for importing data into solr. |

Snippet

From our opinion, we consider that the process of indexing, Nutch/Tika + Solrindexing is much easier than SolrCell. Since using SolrCell, we need to dump out those files from the original crawled data first. Second, we need to run the bash file for finding html and image files with command curl to import all data into solr. However, using Nutch/Tika +Solrindexing, we can import all crawled data into solr for indexing without additional handling for crawled data. Though SolrCell gives users to have more freedom on operating data for importing data into Solr. We, therefore, consider that Nutch/Tika + Solrindexing is much more easier than SolrCell.

**3. Design and implement two ranking algorithms for your Weapons data documents**
**Describe in detail and formally both of your ranking algorithms. You should describe the input, what your algorithms do to compute a rank, how to test them (and prove that they are working as expected). Do NOT simply provide advantages and disadvantages from a quick Google search.**

In content-based, the program asks solr for fields ‘tf’ and ‘idf’ of every document for the query text. Tf means the term frequency of the query term in certain document. Idf is the inverse of having query term in how many document. The higher the idf score means more important when this query term showing. And each ‘tf’ multiply ‘idf’ stands for how the document is relevant to the query text among all documents. At last, we sort the sum of each ‘tf’ multiply ‘idf’ in a descending order.

```
23:47 kage-/Dropbox/572/hw2)>python content.py
Enter your query: texas rifle
Enter your limit # of rows: 5
http://www.shooting.org/Guns/Ruger
43.8975
http://www.shooting.org/Guns/Mossberg
39.790997
http://www.shooting.org/Guns/Walther
36.408417
http://www.shooting.org/Guns/Chiappa_Firearms
30.137402
http://www.shooting.org/Guns/Keystone_Sporting_Arms
30.137402
```

We ask user for the query and the lines of data going to show first. For example if the user query for “texas rifle” and ‘3 lines for showing’, then the program will get the ‘tf’ of texas and rifle, and ‘idf’ of texas and rifle. Then showing the top 3 result after sorting.

The tf and idf fields are calculated by solr. No matter how many query we input, there’s no change of those two fields of certain document of certain input query. For example, if input 1 query for “rifle 2015 california” and input 2 query only for “rifle”; those two query will have the same ‘rifle’ idf field among all document, and will have same ‘rifle’ tf field for same document. It proves that tf and idf in solr is stable. So that this method must work.

( idf of rifle are all the same ) (tf of rifle are the same among same document)

```
00:15 kage-/Dropbox/572/hw2)>python content.py
Enter your query: rifle
Enter your limit # of rows: 5
http://www.shooting.org/Guns/Ruger
12.529964
3.50340175629
http://www.shooting.org/Guns/Mossberg
11.357817
3.50340175629
```

```
00:16 kage-/Dropbox/572/hw2)>python content.py
Enter your query: rifle 2015 california
Enter your limit # of rows: 5
http://www.shooting.org/Guns/Ruger
12.529964
3.50340175629
http://www.shooting.org/Guns/A_Zoom
8.0
3.50340175629
```

In link-based, we simply build a graph and link all document which share same features. After linking the documents, we use pagerank algorithm to calculate the score of each document based on the query. The score stands for the linking relevancy to the query input. The linking relevancy of document A is defined as LR(A) = (1-0.85)*0.85*LR(B)/out_link_num(B). It has been proven that the value of all linking relevancies in a graph will become stable after an amount of iterations.
We ask the user the input the query just as previous. Then we try to find the features of the query term through metadata, geo-data and solr-data. Each feature will form a graph, and a document will be in several groups if it is in multiple features. At last, we apply the pagerank algorithm among all graphs, and assume that if the total variation of each iteration is less than 0.01 means that the pagerank score is becoming stable and just stop iterating.
To prove the link-based algo is not easy, because the linking relevance is not easy to compute by intuition. Since the most important part is the graph, I think the easiest way is to show it on D3. D3 can simply show the graph of our linking result and also showing the score of the link-based score by the size of each node. In problem 8 there's a graph example of applying D3,which We mostly use to evaluate our graph.

```
23:49 kage~/Dropbox/572/hw2)>python link.py
Enter your query: texas AK47 online
Enter your limit # of rows: 5
http://assets1.lionseek.com/item/guns/thumb_107748253-new-vepr-7-62x39-ak-47-ak47-arcadia-no-s
hipping.jpg
0.167283543295
----------------------
http://www.lionseek.com/guns/brand/pointer-brand/new-vepr-7-62x39-ak-47-ak47-arcadia-no-shippi
ng-2b31e8
0.165763693795
----------------------
http://www.gunlistings.org/safetransactions.php
0.16380742041
----------------------
http://www.gunlistings.org/advertise.php
0.16380742041
----------------------
http://assets1.lionseek.com/item/guns/medium_149535768-new-vepr-7-62x39-ak47-arcadia-ak-47.jpg
0.162723991795
```

Please answer how effective the link-based algorithm was compared to the content-based ranking algorithm in light of these weapons challenge questions?

After testing these two methods, we find that content-based one is more sensitive of the query string and the 'content' of the document. Since the tf and idf value is calculated by solr, it's not that difficult and expensive to get these fields. Otherwise the cost is high to use tfidf score. The link-based method is based on links and query features and apply for pagerank algorithm. The pagerank algo is a great method to evaluate the linking score in graph. The issue about link-based method is that how precise and how complete the feature graphs are. If the feature graphs are perfectly designed, then linked-based algo is quicker, cheaper and more efficient, since we only need to build the graphs in the beginning. However, a great searching must be use a combination of methods. They are both great method, but I think link-based should have more weight in the searching engine.
In the weapons questions, I think tfidf perform better since solr has already get the tf idf fields. And also we are not able to build such a completed graph. However, if we have more time and more sources available, I would say that building the link-based is a better direction.

What questions were more appropriate for the link based algorithm compared to the content one?

Searching for a specific key words from the content of is the main feature of content-based algo. Thus, if we try to search for certain specific content, it's more efficient to use content-based. For example, if I like to find 'AK47 with bullet', it's better to use content-based and input 'AK47 AND bullet'. Linke-based do better like 'buying guns online in us ', which has some features we might able to extract from, and it's not precise if doing with content-based.

**4. Develop a suite of queries that demonstrate answers to the relevant weapons related questions below.**

a. Identify the Mexican unauthorized purchase of gun in the past 5 years.
   - Query: Mexico^4 AND gun^4 AND content: (illegal OR prohibit OR ban OR (official AND lack)) AND (sell OR buy OR sale) AND [2010 TO NOW] AND url: *html
   - Explanation:
     > keywords for "unauthorized": illegal, prohibit, ban, (lack, official)
     > keywords for "purchase": sell, buy, sale

b. 1) Identify all rifles which are sold in Mexico from 2012 to 2015.
   - Query: content: rifle AND Mexico AND [2012 TO 2015] AND (sale OR buy OR sell ) AND url:*html
   - Explanation:
     > keywords for "sold": sell, buy, sale

   2) Determine whether rifles that is queried via b(1) are stolen goods or not.
   - Query: content: rifle AND Mexico AND [2009 TO 2012] AND (sale OR buy OR sell ) AND url:*html
   - Explanation:
     > In order to know whether the rifles that we queried in b(1) are stolen goods or not, we have to compare the number of rifles that are sold from 2012 to 2015 with the number of rifles that are sold earlier than 2012 (In our case, the time window we selected is from 2009 to 2012). If the former is larger than the later, then we conclude that all rifles that we retrieve are related to stolen goods.
     > Based on our queries for b(1) and b(2), both the number of queries that we found in b(1) and b(2) are 1. Thus, all rifles that we get in b(1) are NOT related to stolen goods.

c. Identify whether the rifles which are sold in Mexico from 2012 to 2015 are stolen.
   - Query: content: rifle AND Mexico AND [2012 TO 2015] AND (sale OR buy OR sell ) AND url:*html
   - Explanation:
     > keywords for "sold": sell, buy, sale
     To decide if the rifles which are sold from 2012 to 2015 are related to stolen goods, we have to know the number of rifles that is sold earlier than 2012 (In our case, the time window we selected is from 2009 to 2012).

   - Query: content: rifle AND Mexico AND [2009 TO 2012] AND (sale OR buy OR sell ) AND url:*html

   Since both the number of rifles that we quiry are 1, we conclude that all rifles which are sold in Mexico from 2012 to 2015 are NOT stolen.

d. Identify gun and weapon ads that are posted by person whom are underage.
   1) If underage means the age under 17:
   - Query: (gun OR weapon)^4 content: underage OR ((under OR below) AND 17) AND url: *html
   - Explanation:
     > keywords for "underage": underage, under 17, below 17

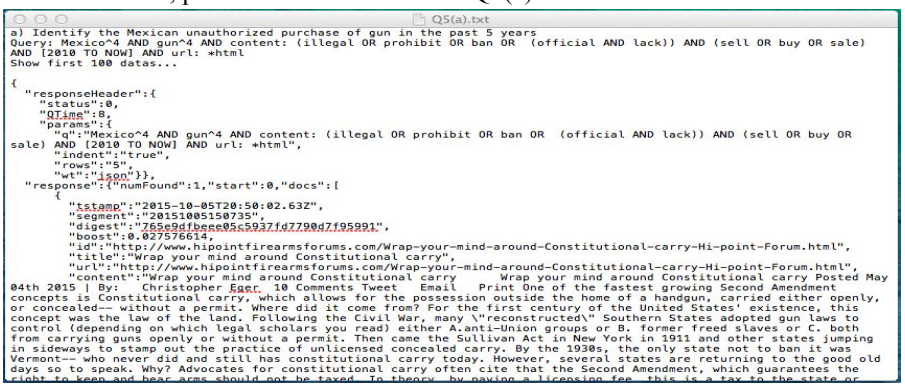
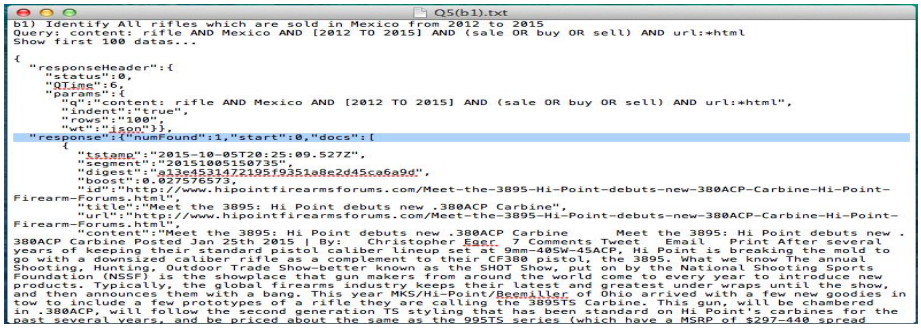
   2) If underage means the age under 21:
   - Query: (gun OR weapon)^4 content: underage OR ((under OR below) AND 21) AND url: *html
   - Explanation:
     > keywords for "underage": underage, under 21, below 21

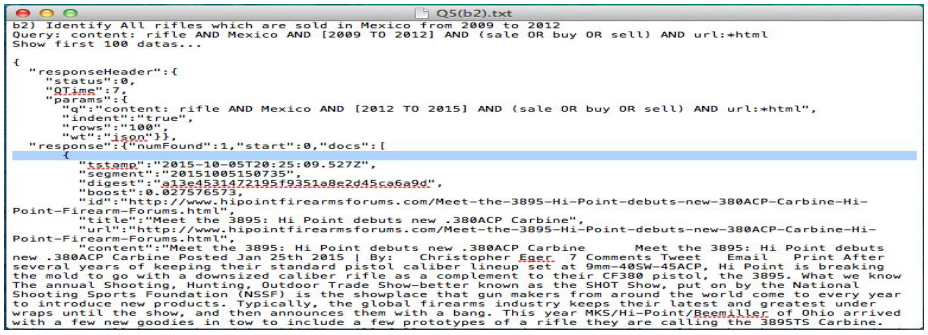
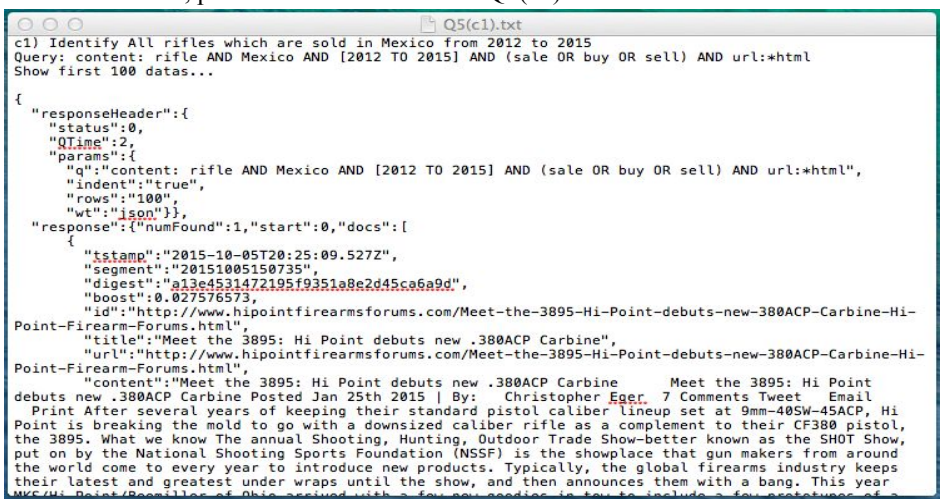
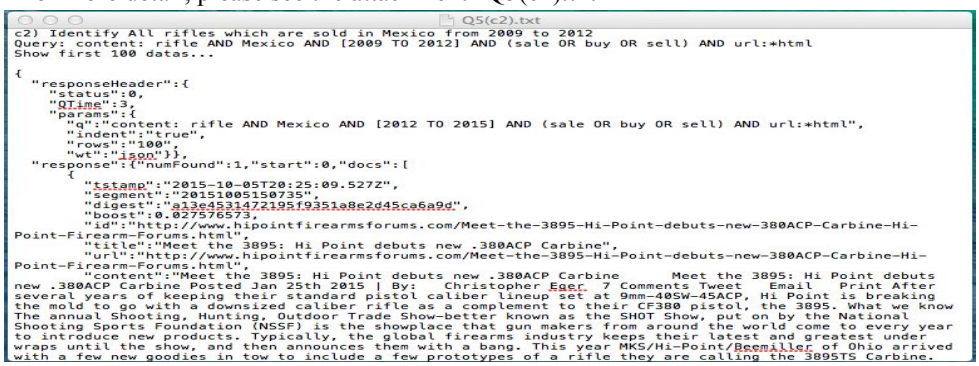
e. Identify the unlawful transfer, sale, possession of explosives, and WMD devices

- Query: (violate OR violation OR illegal OR prohibit OR ban OR criminal)^4 AND content: (explore OR explosive OR nuclear OR chemical OR transfer OR (sale OR buy OR sell)) AND url: *html
- Explanation:
  > keywords for “unlawful”: violate, violation, illegal, prohibit, ban, criminal
  > keywords for “sale”: sell, buy, sale
  > keywords for “possession of explosives”: explore, explosive
  > keywords for “WMD devices”: nuclear, chemical

**5. Develop a program in Python, Java, and/or Bash that runs your queries against your Solr or ElasticSearch index and outputs the results in an easy to read list of results demonstrating your relevancy algorithms and answers to your challenge questions from Task #5.**

* The program that we developed is the attached file “Q5.py”

| | |
|---|---|
| a.Identify the Mexican unauthorized purchase of gun in the past 5 years. | * For more detail, please see the attachment “Q5(a).txt”<br>  |
| b(1). Identify all rifles which are sold in Mexico from 2012 to 2015. | * For more detail, please see the attachment “Q5(b1).txt”<br>  |
| b(2). Determine whether rifles that is queried via b(1) are stolen goods or not. | As the explanation in Q4(b2), we have to retrieve all rifles which are sold in Mexico from 2009 to 2012. Both the number of rifles in b(1) and b(2) are 1. Thus, all rifles that we get in b(1) are NOT related to stolen goods.<br>* For more detail, please see the attachment “Q5(b2).txt” |

| | |
|---|---|
| |  |
| c. Identify whether the rifles which are sold in Mexico from 2012 to 2015 are stolen. | Following is the output for “all rifles which are sold in Mexico from 2012 to 2015.”<br>* For more detail, please see the attachment “Q5(c1).txt”<br>  <br>In order to know whether the rifles which are sold from 2012 to 2015 are related to stolen goods, we have to know the number of rifles that is sold earlier than 2012 (in our case, the time window we selected is from 2009 to 2012). Following is the output that “all rifles which are sold in Mexico from 2009 to 2012.”<br>* For more detail, please see the attachment “Q5(c2).txt”<br>  <br>Since both the number of rifles that we quiry are 1, we conclude that all rifles which are sold in Mexico from 2012 to 2015 are NOT stolen. |
| d. Identify gun and weapon ads that are posted by | If underage means the age under 17:<br>* For more detail, please see the attachment “Q5(d1).txt” |

| | |
|---|---|
| person whom are underage. | Q5(d1).txt<br>d) Identify gun and weapon ads that are posted by person whom are underage.<br>1) If underage means the age under 17...Query: (gun OR weapon)^4 content: underage OR ((under OR below) AND 17) AND url: *html<br>Show first 100 datas...<br><br>{<br>"responseHeader":{<br>"status":0,<br>"QTime":17,<br>"params":{<br>"q":"(gun OR weapon)^4 content: underage OR ((under OR below) AND 17) AND url: *html",<br>"indent":"true",<br>"rows":"100",<br>"wt":"json"}},<br>"response":{"numFound":22,"start":0,"docs":[<br>{<br>"tstamp":"2015-10-05T20:50:02.63Z",<br>"segment":"20151005150735",<br>"digest":"765e9dfbeee05c5937fd7790d7f95991",<br>"boost":0.027576614,<br>"id":"http://www.hipointfirearmsforums.com/Wrap-your-mind-around-Constitutional-carry-Hi-point-Forum.html",<br>"title":"Wrap your mind around Constitutional carry",<br>"url":"http://www.hipointfirearmsforums.com/Wrap-your-mind-around-Constitutional-carry-Hi-point-Forum.html",<br>"content":"Wrap your mind around Constitutional carry Wrap your mind around Constitutional carry Posted May 04th 2015 \| By: Christopher Eger 10 Comments Tweet Email Print One of the fastest growing Second Amendment concepts is Constitutional carry, which allows for the possession outside the home of a handgun, carried either<br><br>If underage means the age under 21:<br>* For more detail, please see the attachment "Q5(d2).txt"<br>Q5(d2).txt<br>d) Identify gun and weapon ads that are posted by person whom are underage.<br>2) If underage means the age under 21...Query: (gun OR weapon)^4 content: underage OR ((under OR below) AND 21) AND url: *html<br>Show first 100 datas...<br><br>{<br>"responseHeader":{<br>"status":0,<br>"QTime":7,<br>"params":{<br>"q":"(gun OR weapon)^4 content: underage OR ((under OR below) AND 21) AND url: *html",<br>"indent":"true",<br>"rows":"100",<br>"wt":"json"}},<br>"response":{"numFound":32,"start":0,"docs":[<br>{<br>"tstamp":"2015-10-05T20:02:22.368Z",<br>"segment":"20151005150735",<br>"digest":"0486d47abc91c0e454134fba818d9045",<br>"boost":0.02758735,<br>"id":"http://www.hipointfirearmsforums.com/Do-you-carry-in-church-hipoint-firearms-forum.html",<br>"title":"Do you carry in church?",<br>"url":"http://www.hipointfirearmsforums.com/Do-you-carry-in-church-hipoint-firearms-forum.html",<br>"content":"Do you carry in church? Do you carry in church? Posted Apr 06th 2015 \| By: Christopher Eger 7 Comments Tweet Email Print If you are able under your local laws to carry concealed, it is advocated by self-defense professionals that you do. After all, it is better to have it and not need it than to need it and not have it. |
| e. Identify the unlawful transfer, sale, possession of explosives, and WMD devices. | * For more detail, please see the attachment "Q5(e).txt"<br>Q5(e).txt<br>e) Identify the unlawful transfer, sale, possession of explosives, and WMD devices<br>Query: (violate OR violation OR illegal OR prohibit OR ban OR criminal)^4 AND content: (explore OR explosive OR nuclear OR chemical OR transfer OR (sale OR buy OR sell)) AND url: *html<br>Show first 100 datas...<br><br>{<br>"responseHeader":{<br>"status":0,<br>"QTime":3,<br>"params":{<br>"q":"(violate OR violation OR illegal OR prohibit OR ban OR criminal)^4 AND content: (explore OR explosive OR nuclear OR chemical OR transfer OR (sale OR buy OR sell)) AND url: *html",<br>"indent":"true",<br>"rows":"100",<br>"wt":"json"}},<br>"response":{"numFound":1,"start":0,"docs":[<br>{<br>"tstamp":"2015-10-05T20:50:02.63Z",<br>"segment":"20151005150735",<br>"digest":"765e9dfbeee05c5937fd7790d7f95991",<br>"boost":0.027576614,<br>"id":"http://www.hipointfirearmsforums.com/Wrap-your-mind-around-Constitutional-carry-Hi-point-Forum.html",<br>"title":"Wrap your mind around Constitutional carry",<br>"url":"http://www.hipointfirearmsforums.com/Wrap-your-mind-around-Constitutional-carry-Hi-point-Forum.html",<br>"content":"Wrap your mind around Constitutional carry Wrap your mind around Constitutional carry Posted May 04th 2015 \| By: Christopher Eger 10 Comments Tweet Email Print One of the fastest growing Second Amendment concepts is Constitutional carry, which allows for the possession outside the home of a handgun, carried either openly, or concealed-- without a permit. Where did it come from? For the first century of the United States' existence, this concept was the law of the land. Following the Civil War, many \"reconstructed\" Southern States adopted gun laws to control (depending on which legal scholars you read) either A.anti-Union groups or B. former freed slaves or C. both from carrying guns openly or without a permit. Then came the Sullivan Act in New York in 1911 and other |

**6. (Extra Credit) Develop a Lucene-latent Dirichlet allocation (LDA) technique for topic modeling on your index and use it to rank and return documents.**

The main idea for Lucene-latent Dirichlet allocation (LDA) is that this algorithm uses unsupervised learning while function query in solr use merely TF/IDF algorithm.
First, using command “bin/indexDirectory [--help] <inDir> <outIndexDir> <outLDAIndex> [--fileCodes <fileCodes>] [--ldaConfig ldaConfig1,ldaConfig2,...,ldaConfigN ]” will generate the LDA index for our crawled data. The LDA will using unsupervised learning to separate those files into different groups according to each files similarity.
Second, using command “bin/queryWithLDA [--help] <indexDir> <LDAIndexDir> <queryDir> <resultsDir> [--K <K>] [--scoringCode <scoringCode>]”, we can get the output sets with respect to input queries and learned model. Take the query “What time-based trends exist in Gun ads?” in question 4.a for example. We perceive that keywords for “unauthorized” are “illegal, prohibit, ban, (lack, official)” and keywords for “purchase” are “sell, buy, sale”. In function query, the results only return data that match word “illegal, sell” in content field. However, using LDA, we get the result sets for keyword “illegal” are content field contains words “legal, banned, forbidden, restrict...” for the reason of LDA use unsupervised learning and consider those files are similar.

**7. (Extra Credit) Figure out how to integrate your relevancy algorithms into Nutch.**
The interface of ScoringFilter in Nutch provided by homework description contains method injectScore, initialScore and distributeScoreToOutlinks. We implement this interface and use method injectScore to store the original score from SolrCell and use method initialScore to store the score of relevant links that are produced by our link-based algorithm. Finally, we use the method of distributeScoreToOutlinks for keep updating the new links and scores. That’s the way how we figure out to integrate relevancy algorithm into Nutch.

**8. (Extra Credit) Create a D3-based visualization of your link-based relevancy.**
We write python to implement our link-based algorithm and generate json, and then present the json data by D3. We try to make a query with small amount of data because Plesase check getjson.py, getJson.json and d3visual.html.