

RECALL

Increment 4

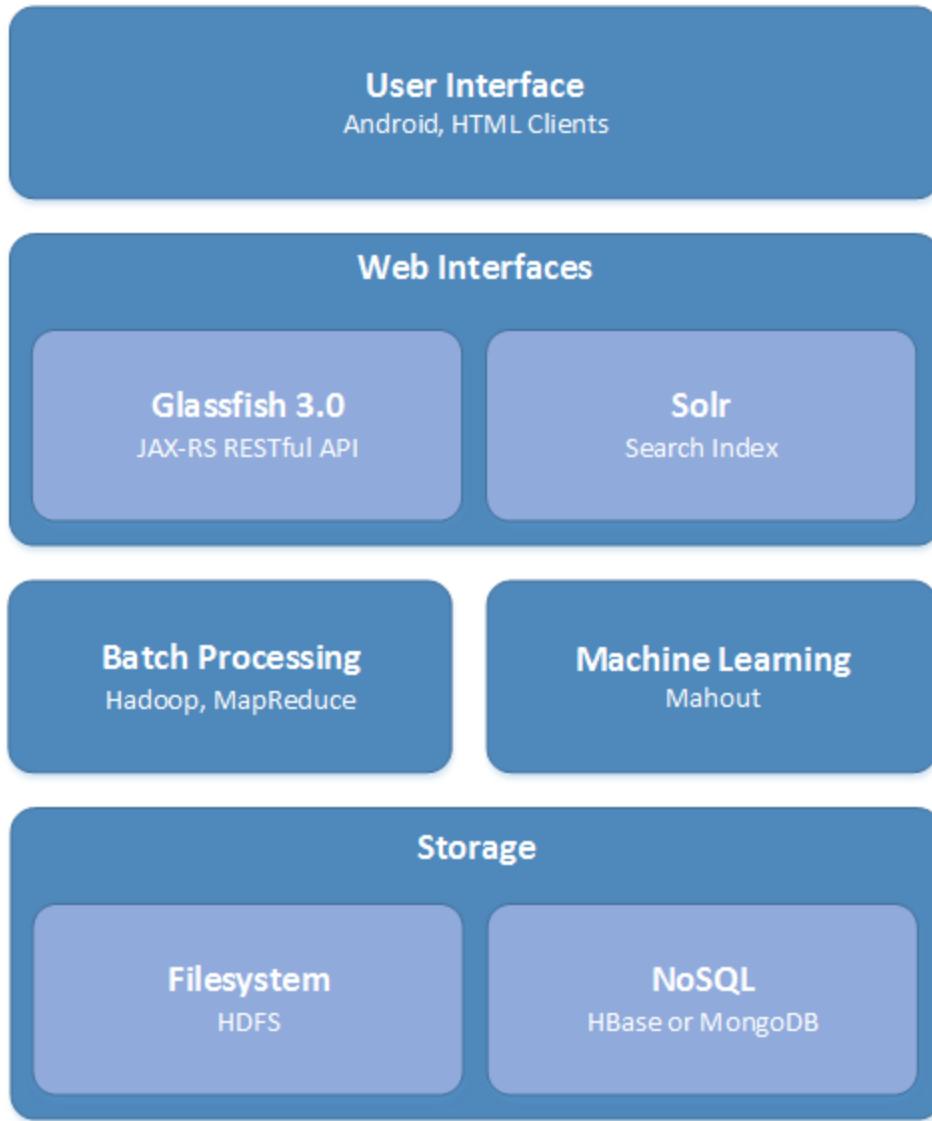
Group 6

James Clark
Tirumala Konireddy
Jagadish Tirumalasetty
Wang Zhang

University of Missouri - Kansas City

CS 560: Knowledge Discovery and Management
Dr. Yugyung Lee
May 2, 2014

System Architecture Diagram



Overall Process

1. supervised learning
2. manual data inspection and classification
3. most general classification possible
4. apply classification to entire organization
5. reinspect
6. reclassify, ex: (mislabeled, contamination) > (salmonella, child, pet)

Training the data

Data is extracted from digital gov. It's collected and pre-processed to TSV. Now, using Excel, I can copy rows into a table and manually categorize them. The benefit of Excel is that copy and paste results in an automatically tab separated values format, immediately compatible for training Mahout.

This is called assisted learning where a human interacts with computer, monitoring the application of algorithms, accuracy of results, and correspondingly approves or declines before arranging the data for production release.

Tutorial 7 covers classification of twitter data. The same TwitterTSVToSeq class can classify our files for loading into Hadoop.

Generally, this is an outline of the process. The actual source code for this outline is coded in the appendix.

1. Convert training set to Hadoop sequence file format
2. Upload sequence file to HDFS
3. Transform training sets to vectors with Mahout (using term frequency by document frequency weights)
4. Split data into training set and testing set
5. Use training set to train classifier
6. Test the classifier on the training set
7. Test the classifier on the testing set
8. To use classifier and classify new documents, copy several files from HDFS
9. Classify the new file (optional section in bold red writes to text file)

The final output document is printed in a uniform set of 2-line data objects following this format. Note the items in bold are relevant to our purpose of classification:

Tweet: 3d13ad94d4 Nutriom LLC, a Lacey, Wash. establishment, is recalling an additional 82,884 pounds of processed egg products that may be contaminated with Salmonella Washington Firm Recalls Dried Egg Products Due To Possible Salmonella Contamination contamination: -273.64178143939307 mislabeled: -325.0381960862568 => **contamination**

Classification Results

FDA

2 labels: contamination, mislabeled

5 labels: allergen, animal, debris, illness, medecine

Test the classifier on the training set

Correctly Classified Instances : 149 98.6755%

Incorrectly Classified Instances : 2 1.3245%

Test the classifier on the testing set

Correctly Classified Instances : 100 83.3333%

Incorrectly Classified Instances : 20 16.6667%

Testing the training data is between 92% and 100% accurate

Testing the testing data is approximately 85% accurate

Observed Data Anomalies

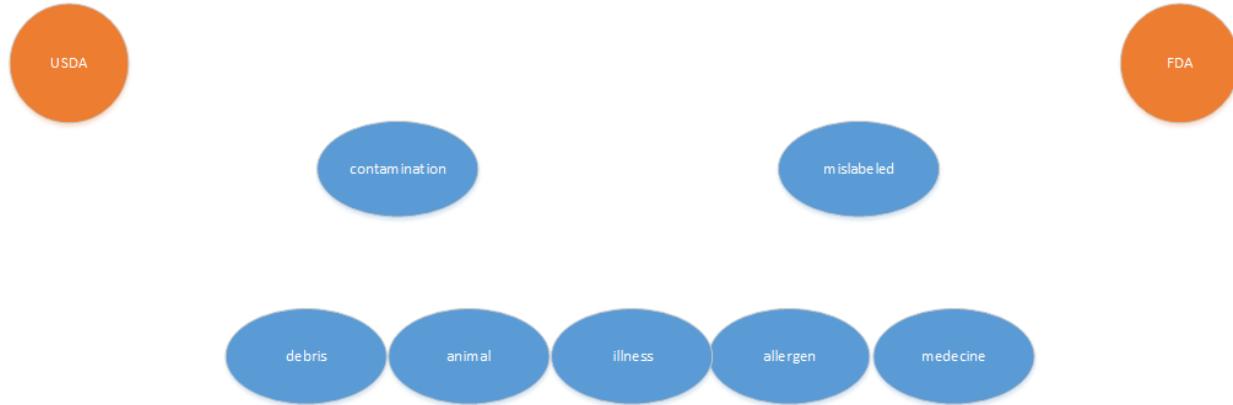
FDA recall_id cb7b35d997, 6266a8f724, 5f2fc28def, 8758f480f8 are examples of voluntary recall with no explanation as to why. Categorized as contamination.

Moving Data into Solr

The challenge with the UMKC managed Solr instance is a lack of flexibility to update fields for our application. While technically feasible, it would be high-risk to assume any changes on the production server this close to final presentations. A critical mistake would be a massive takedown for all the groups relying on Solr for their presentation and general app functionality. Therefore, we are left mapping our data to the available fields in Solr. This is a list of fields that we can use in our application.

FDA & USDA

Solr Field	Class Field
id	recall_number
name	organization
features[]	
title[]	
subject	
description	description
comments	summary
author	
keywords	classification2
category	organization
resourcename	classification1
url	recall_url
last_modified	recall_date

**NHTSA**

Solr Field	Class Field
id	recall_number
name	organization
features[]	
title[]	
subject	recall_subject
description	consequence_summary, defect_summary, corrective_summary
comments	notes
author	component_description
keywords	manufacturer
category	organization
resourcename	classification1
url	recall_url
last_modified	recall_date
popularity	potential_units_affected

Now, These results are available to retrieve from solr server through android application.

Classification For Nhtsa Data

Raw Data

The data collected from National Highway Transport Safety Administration API is as below.

C:\Users\tirumala\Desktop\Assignments and Homeworks\kdm\project\final project\data\nhtsa.tsv - Notepad++

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

prev_users.txt users_recommendations.txt prev_users_bt.txt users_choice.txt join.java spect.txt nhtsa.tsv finalcategories.txt

organization	recall_number	recall_date	recall_url	records	component_description	manufacturer	code	potential_units_affected	initiator	report_date
NHTSA	14V134000	2014-04-16	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V134000	EQUIPMENT:RECFC					
NHTSA	14V136000	2014-04-16	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V136000	EQUIPMENT:RECFC					
NHTSA	14V137000	2014-04-16	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V137000	ENGINE AND ENG					
NHTSA	14V139000	2014-04-16	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V139000	ENGINE AND ENG					
NHTSA	14V182000	2014-04-16	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V182000	POWER TRAIN:AC					
NHTSA	14V173000	2014-04-16	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V173000	ELECTRICAL SYS					
NHTSA	14C002000	2014-04-14	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14C002000	CHILD SEAT:HAF					
NHTSA	14V165000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V165000	STRUCTURE:FRONT					
NHTSA	14V148000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V148000	SEATS, TOYOTA, I					
NHTSA	14V130000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V130000	ENGINE AND ENG					
NHTSA	14V131000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V131000	ENGINE AND ENG					
NHTSA	14V132000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V132000	ENGINE AND ENG					
NHTSA	14V128000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V128000	EQUIPMENT:RECFC					
NHTSA	14V164000	2014-04-11	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V164000	SEATS:FRONT AS					
NHTSA	14V171000	2014-04-10	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V171000	ELECTRICAL SYS					
NHTSA	14V169000	2014-04-09	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V169000	SEATS, TOYOTA, I					
NHTSA	14V168000	2014-04-09	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V168000	AIR BAGS:FRONT					
NHTSA	14V129000	2014-04-09	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V129000	EQUIPMENT:OTHE					
NHTSA	14V127000	2014-04-09	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V127000	EQUIPMENT:RECFC					
NHTSA	14V125000	2014-04-08	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V125000	EXTERIOR LIGHT					
NHTSA	14V123000	2014-04-07	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V123000	EQUIPMENT:OTHE					
NHTSA	14V124000	2014-04-07	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V124000	ENGINE AND ENG					
NHTSA	14V121000	2014-04-07	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V121000	VEHICLE SPEED					
NHTSA	14V122000	2014-04-07	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V122000	STRUCTURE, TIRE					
NHTSA	14V120000	2014-04-04	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V120000	SUSPENSION:RECFC					
NHTSA	14V115000	2014-04-04	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V115000	TIRES:TEMPORAR					
NHTSA	14V119000	2014-04-04	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V119000	ENGINE AND ENG					
NHTSA	14V149000	2014-04-04	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V149000	ELECTRICAL SYS					
NHTSA	14C003000	2014-04-04	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14C003000	CHILD SEAT:HAF					
NHTSA	14V114000	2014-04-03	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V114000	FUEL SYSTEM, C					
NHTSA	14E007000	2014-04-03	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14E007000	ENGINE AND ENG					
NHTSA	14V113000	2014-04-03	http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=1	DtargetCategory=R&searchCriteria.nhtsa_ids=14V113000	STRUCTURE:FRONT					

The nhtsa data file contains 2856 recall categories regarding Cars, Buses, trailers, childcare etc.,

		C:\Users\t			C:\Users\t		
		File	Edit	Search	View	Encoding	Language
1	0	manufacturer					
2	1	Crossroads RV					
3	2	Keystone RV Company					
4	3	Mack Trucks, Inc.					
5	4	Volvo Trucks North America					
6	5	Volkswagen of America, Inc.					
7	6	Mazda North American Operations					
8	7	BABY TREND INC.,					
9	8	Ford Motor Company					
10	9	Toyota Motor Engineering & Manufacturing					
11	10	Autocar Industries, LLC					
12	11	Autocar, LLC					
13	12	Heartland Recreational Vehicles, LLC					
14	13	General Motors LLC					
15	14	Thor Motor Coach					
16	15	Nissan North America, Inc.					
17	16	Jaguar Land Rover North America, LLC					
18	17	Winnebago Industries, Inc					
19	18	Crane Carrier Company					
20	19	Terex Aerial Work Platforms					
21	20	Zero Motorcycles Inc.					
22	21	Triple E Recreational Vehicles					
23	22	Evenflo Company, Inc.					
24	23	Tiffany Coach Builders					
25	24	The Gates Corporation					
26	25	Chrysler Group LLC					
27	26	Daimler Buses North America					
28	27	Great Dane Trailers					
29	28	Blue Bird Body Company					
30	29	Contract Manufacturer, LLC					
31	30	Navistar, Inc.					
32	31	Eldorado National- California, Inc.					
33	32	Honda (American Honda Motor Co.)					
		2826	2825	compton truck & equip. co	Auto		
		2827	2826	pointer williamette trlr.	Others		
		2828	2827	dura	Others		
		2829	2828	boles-aero incorporated	Corporation		
		2830	2829	g.h. hicks & sons, inc.	Business		
		2831	2830	winter welding & machine	Others		
		2832	2831	gil-mar welding corp.	Corporation		
		2833	2832	skytop rig company	Company		
		2834	2833	rite-way, inc. of indiana	Business		
		2835	2834	truck trailer equip. co.	Auto		
		2836	2835	durham mfg. company	Company		
		2837	2836	motac, incorporated	Corporation		
		2838	2837	vista international corp.	Corporation		
		2839	2838	copco steel & eng. co.	Others		
		2840	2839	tuttle mfg. co.	Manufacturing company		
		2841	2840	nelson manufacturing co.	Manufacturing company		
		2842	2841	lofgren mfg. company	Company		
		2843	2842	midwest srvc & supply co.	Others		
		2844	2843	kershaw	Others		
		2845	2844	aluminum body corporation	Corporation		
		2846	2845	cox trailers, inc.	Trailer		
		2847	2846	transport trailers, inc.	Trailer		
		2848	2847	di giorgio leisure prod.	Others		
		2849	2848	traveleze industries,inc.	Industries		
		2850	2849	hill mfg. company	Company		
		2851	2850	barth, incorporated	Corporation		
		2852	2851	cavalier mfg. co.	Manufacturing company		
		2853	2852	banner homes corporation	Corporation		
		2854	2853	flare mfg. inc.	Manufacturing company		
		2855	2854	ford motor company test adv as	Company		
		2856	2855	renault, incorporated	Corporation		
		2857					

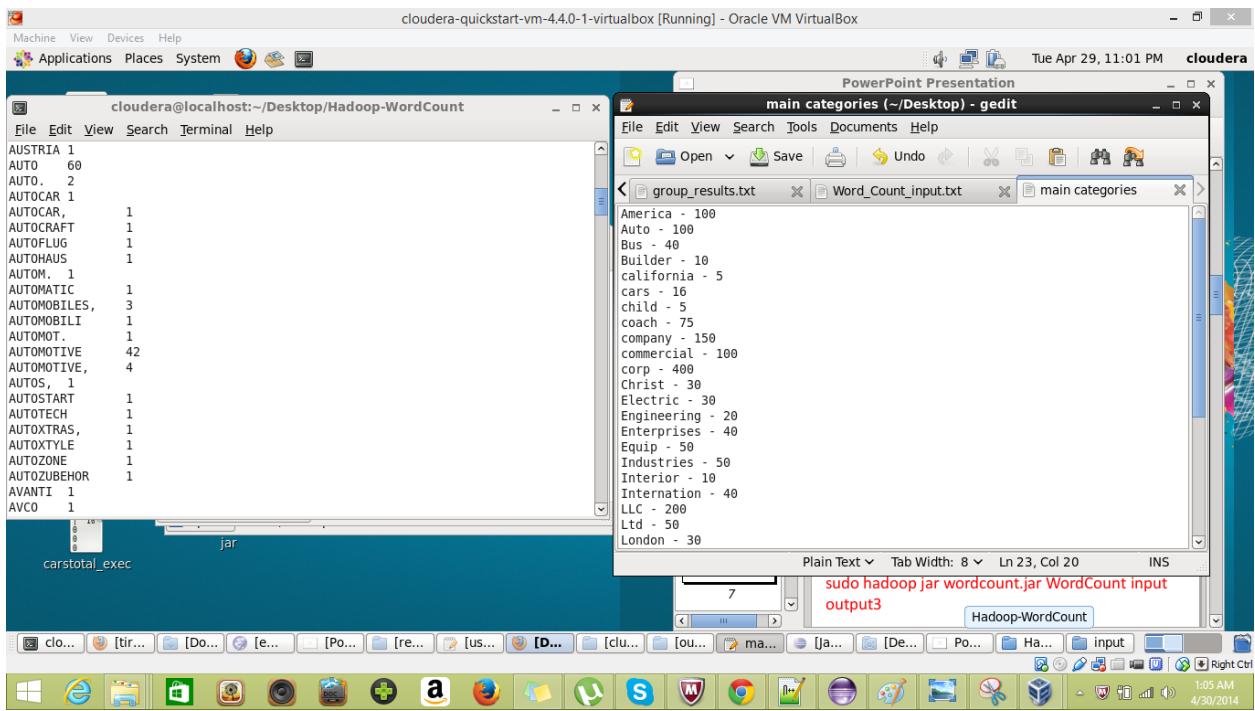
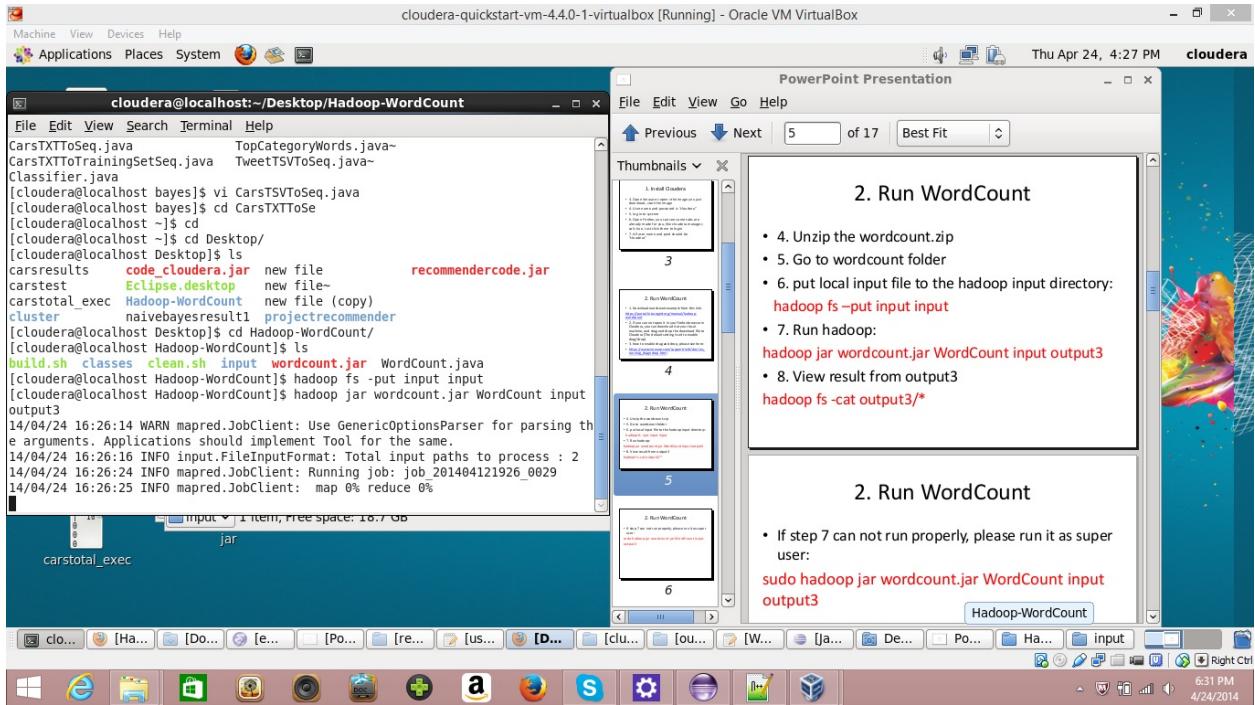
It is difficult for a user to choose a category from a list of 2856 very specific categories like above.

Then, I had grouped or classified the relevant specific categories of 2856 into 24 general categories like Auto, Childcare, Industry etc ., like below



```
1 Auto
2 USA
3 Children
4 Company
5 Engineering
6 Limited liability company
7 Coaches
8 Industries
9 Others
10 Buisness
11 Buildings
12 Corporation
13 Trailer
14 California
15 Manufacturing company
16 Private Ltd company
17 Equipment
18 National
19 International
20 Interiors
21 Commercial
22 Technology
23 Electronics
24 Men
25
```

To do the classification for above results i had used a word count program to find the key words or peculiar words in the above 2856 specific categories like below to classify a specific category to a generalized category.



I got the word count of all words in the specific categories list.

```

1 America - 100
2 Auto - 100
3 Bus - 40
4 Builder - 10
5 California - 5
6 cars - 16
7 chid - 5
8 coach - 75
9 company - 150
10 commercial - 100
11 CORP - 400
12 Christ - 30
13 Electric - 30
14 Engineering - 20
15 Enterprises - 40
16 Equip - 50
17 Industries - 50
18 Interior - 10
19 Internation - 40
20 LLC - 200
21 Ltd - 50
22 London - 30
23 Manufacturing - 100
24 MFG. - 100
25 Motor - 56
26 Miriam - 139
27 Morel - 258
28 Mr. - 51
29 Mrs. - 342
30 National - 19
31 Tech - 20
32 Trailer - 63
33 vehicle - 30
34

```

Then, i had identified some particular words those can be used to classify the specific categories into generalised categories as below. With the above words i found that i am able to classify upto 2500 specific categories into one of the generalized category. I had introduced a new category called “Others” category for those 280+ unclassified categories to limit the generalized categories to “Number 24” and those 24 categories are as below.

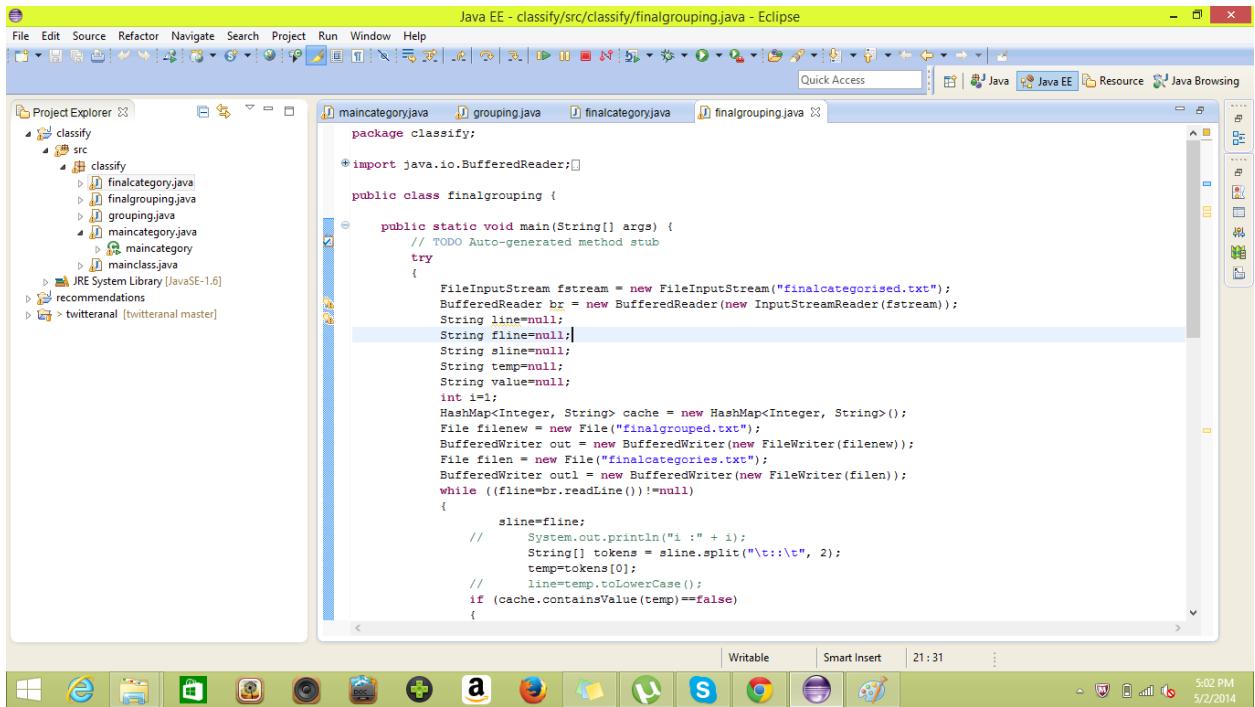


The screenshot shows a Notepad window with the following content:

```
1 Auto
2 USA
3 Children
4 Company
5 Engineering
6 Limited liability company
7 Coaches
8 Industries
9 Others
10 Buisiness
11 Buildings
12 Corporation
13 Trailer
14 California
15 Manufacturing company
16 Private Ltd company
17 Equipment
18 National
19 International
20 Interiors
21 Commercial
22 Technology
23 Electronics
24 Men
25
```

The word "Buisiness" is underlined in red. Line 14 ("California") is highlighted with a light purple background.

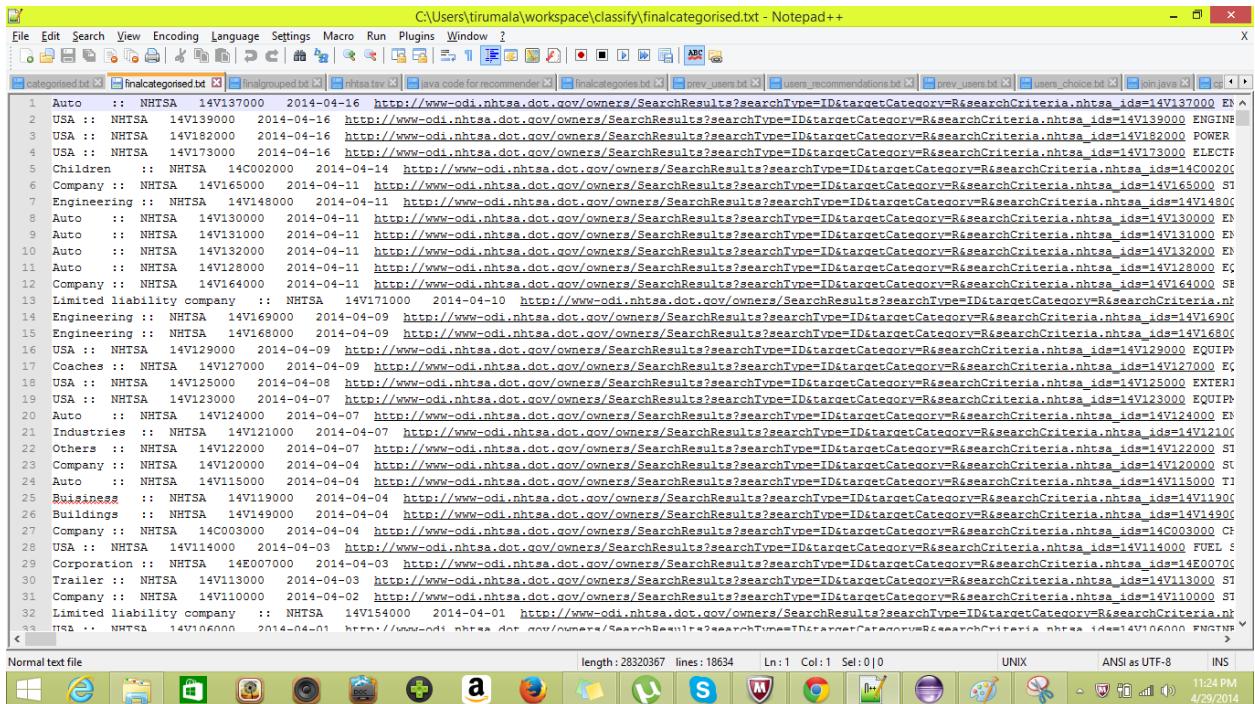
Then, I had developed a java program to classify the data entries containing specific categories into generalised categories.



The screenshot shows the Eclipse IDE interface with the following details:

- Project Explorer:** Shows a project named "classify" with several source files: finalcategory.java, finalgrouping.java, grouping.java, maincategory.java, and mainclass.java.
- Java EE - classify/src/classify/finalgrouping.java - Eclipse:** The code for the `finalgrouping` class is displayed. It reads from a file named `finalcategorised.txt`, processes lines, and writes to files `finalgrouped.txt`, `finalcategories.txt`, and `finalgrouped.txt`.
- File Bar:** Contains options like File, Edit, Source, Refactor, Navigate, Search, Project, Run, Window, Help.
- Toolbar:** Includes standard Eclipse icons for file operations, search, and navigation.
- Bottom Status Bar:** Shows the date and time (5:02 PM 5/2/2014).

I had written a new file with appending new generalized category in first column as below.



The screenshot shows a Notepad++ window displaying a text file named `finalcategorised.txt`. The file contains approximately 40 lines of data, each consisting of a category identifier followed by a colon, a space, and a value. The categories include Auto, USA, Children, Company, Engineering, Limited liability company, Engineering, Engineering, USA, Coaches, USA, USA, Company, Buildings, Company, Industries, Others, Company, Auto, Business, Buildings, Company, Company, USA, Corporation, Trailer, Company, and Limited liability company. The values are mostly NHTSA followed by a series of digits and letters. The file is a tab-delimited text file.

I had rewritten this file again by grouping all the entries based on the main category as below.

```

C:\Users\tirumala\workspace\classify\finalgrouped.txt - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
catevised.txt finalcatevised.txt finalgrouped.txt nhtsa.txt java code for recommender finalcategories.txt prev_users.txt users_recommendations.txt prev_users_be.txt users_choices.txt car.java
1 Auto :: NHTSA 14V137000 2014-04-16 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V137000 EI ^
2 Auto :: NHTSA 14V130000 2014-04-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V130000 EI
3 Auto :: NHTSA 14V131000 2014-04-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V131000 EI
4 Auto :: NHTSA 14V132000 2014-04-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V132000 EI
5 Auto :: NHTSA 14V128000 2014-04-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V128000 EI
6 Auto :: NHTSA 14V124000 2014-04-07 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V124000 EI
7 Auto :: NHTSA 14V115000 2014-04-04 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V115000 EI
8 Auto :: NHTSA 14E006000 2014-03-26 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14E006000 EI
9 Auto :: NHTSA 14V095000 2014-03-25 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V095000 EI
10 Auto :: NHTSA 14V091000 2014-03-24 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V091000 EI
11 Auto :: NHTSA 14V096000 2014-03-24 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V096000 EI
12 Auto :: NHTSA 14V073000 2014-03-20 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V073000 EI
13 Auto :: NHTSA 14V088000 2014-03-20 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V088000 EI
14 Auto :: NHTSA 14V076000 2014-03-20 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V076000 EI
15 Auto :: NHTSA 14V078000 2014-03-20 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V078000 EI
16 Auto :: NHTSA 14V102000 2014-03-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V102000 EI
17 Auto :: NHTSA 14V081000 2014-03-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V081000 EI
18 Auto :: NHTSA 14V059000 2014-03-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V059000 EI
19 Auto :: NHTSA 14V086000 2014-03-10 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V086000 EI
20 Auto :: NHTSA 14V070000 2014-03-07 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V070000 EI
21 Auto :: NHTSA 14V058000 2014-03-07 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V058000 EI
22 Auto :: NHTSA 14V057000 2014-03-06 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V057000 EI
23 Auto :: NHTSA 14V036000 2014-02-28 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V036000 EI
24 Auto :: NHTSA 14V038000 2014-02-27 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V038000 EI
25 Auto :: NHTSA 14V033000 2014-02-24 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V033000 EI
26 Auto :: NHTSA 14V037000 2014-02-24 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V037000 EI
27 Auto :: NHTSA 14T001000 2014-02-20 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14T001000 EI
28 Auto :: NHTSA 14V027000 2014-02-18 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V027000 EI
29 Auto :: NHTSA 14V026000 2014-02-18 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V026000 EI
30 Auto :: NHTSA 14V102000 2014-03-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V102000 EI
31 Auto :: NHTSA 14V081000 2014-03-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V081000 EI
32 Auto :: NHTSA 14V059000 2014-03-11 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V059000 EI
33 Auto :: NHTSA 14V086000 2014-03-10 http://www-odi.nhtsa.dot.gov/owners/SearchResults?searchType=IDstargetCategory=ResearchCriteria.nhtsa_ids=14V086000 EI

```

Normal text file length: 28320367 lines: 18634 Ln: 1 Col: 148 Sel: 0|0 UNIX ANSI as UTF-8 INS 11:24 PM 4/29/2014

Then, I had uploaded this file to our group 6 solr server group6_shard1_replica1 collection with the following field names.

Recommendation Algorithm

Recommending a category is different than recommending a product. A particular product may bought more than once, whereas in the case of subscriptions for a category, subscription to a particular category is done only once. Also, we need very little complexity and less number of calculations in run time to produce instant dynamic recommendations. Keeping the above criteria in mind, I had developed a new algorithm to work efficiently for our data by considering the likeliness of both interest and disinterest on a particular category with present subscriptions of user based on previous user's subscriptions.

First, I am collecting all the categories present in the previous user's subscriptions.

Then, I am calculating the number of occurrences of every item in the whole user's data base. This number itself denotes the importance or priority of a particular item in recommending it.

Next, I am providing a index for all individual items with all other items as below.

Index for a recommending item B to user who subscribed to A = (sum of number of occurrences of B along with A in previous user's subscriptions list) - (sum of number of disappearances of B when A is present in previous user's subscriptions list)

Then, We recommend a particular category in the entire list to a user subscribed for A and C which has the highest sum of indexes in A and C other than A and C.

Index of recommending B for user subscribed to A and C categories is

$$\text{Index B (A and C)} = \text{Index B(A)} + \text{Index B(C)}$$

Similarly, If i had A,B,C,D,E categories in my data base, then

$$\text{Index D (A and C)} = \text{Index D(A)} + \text{Index D(C)}$$

$$\text{Index E (A and C)} = \text{Index E(A)} + \text{Index E(C)}$$

Then, I recommend the category in B, D, E which has highest Index.

Using this algorithm, we can eliminate the unlikely combinations in providing recommendations.

For Example :

In amazon.com

if, user 1 had bought a Sony Laptop and a back pack bag

user 2 had bought a Samsung Laptop and a back pack bag too

Here, recommending both Sony Laptop and Samsung Laptop is somewhat good for user who bought a back pack bag.

Recommender algorithms which uses only the appearance of a object with the common object regardless the object type. This will provide a recommendation of a Samsung laptop to a user who had bought a Sony Laptop as it is bought by a user who had bought a back pack bag.

But, it is useless to recommend a Samsung Laptop for a user immediately after buying a Sony Laptop as it is atleast 2 years investment. This type of recommendations can be avoided by calculating the number disappearances along with the number of appearances when a particular item is present.

We had developed a program in java, to implement the above algorithm in providing recommendations to user.

The screenshot shows the Eclipse IDE interface with the title "Java EE - recommendations/src/recommendations/mainclass.java - Eclipse". The Project Explorer view on the left shows a package named "recommendations" containing several Java files: classify, finalcategory.java, finalgrouping.java, grouping.java, maincategory.java, mainclass.java, prev_users.java, and values.java. The main editor window displays the content of the "mainclass.java" file:

```

package recommendations;

import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.File;
import java.io.FileInputStream;
import java.io.FileWriter;
import java.io.InputStreamReader;
import java.util.StringTokenizer;

public class mainclass {
    public static void main(String[] args){
        try{
            // Open the file that is the first
            // command line parameter
            FileInputStream fstream = new FileInputStream("prev_users.txt");
            BufferedReader br = new BufferedReader(new InputStreamReader(fstream));
            // String[] strLine = br.readLine();
            int count=0;
            int j=0;
            int old_j=0;
            // int[] testing = new int[5000];
            int[][] input_data = new int[5000][5000];
            String store = null;
            String unique=null;
            int check =0;
            // String[] temp_data = null;
            String strLine = null;
            while (( strLine = br.readLine()) != null) {
                StringTokenizer st = new StringTokenizer(strLine);
                while (st.hasMoreTokens()){
                    String token = st.nextToken();
                    if(token.equals("Business")){ // Check if token is "Business"
                        if(count==0){ // If count is 0, set old_j to j
                            old_j=j;
                        }
                        if(j==old_j){ // If j is equal to old_j, increment count
                            count++;
                        }
                        if(count==5000){ // If count reaches 5000, break the loop
                            break;
                        }
                    }
                    else{ // If token is not "Business", set j to i
                        j=i;
                    }
                    i++;
                }
            }
            br.close();
        } catch (Exception e){
            System.out.println("Error");
        }
    }
}

```

My previous user database

The screenshot shows the Notepad++ application window with the title "C:\Users\tirumala\workspace\recommendations\prev_users.txt - Notepad++". The file contains a list of user categories separated by commas. The lines are numbered from 1 to 34. The text content is as follows:

```

1 Interiors,International,Industries,Trailer
2 Private Ltd company,International,Interiors,Electronics
3 Children,Limited liability company
4 Interiors,Children,Limited liability company,California,Corporation
5 Auto,Coaches,Engineering,USA,Equipment,International
6 California,Equipment,Corporation,Interiors,Children
7 International,Equipment
8 California,Commercial,Engineering,Limited liability company,USA
9 Electronics
10 Commercial,Company,Buildings
11 Company,Interiors,Coaches
12 National,California,Buildings,Limited liability company,Commercial,Engineering,Private Ltd company
13 Electronics,Trailer,Men,Business,International,Company,Coaches,Limited liability company
14 Manufacturing company,Others,Company,Children,Commercial,Interiors
15 International,Coaches,Trailer
16 USA,Private Ltd company,Company,California,Trailer,Industries
17 Engineering,Others,Corporation,Electronics,Business
18 International,Industries,Equipment,Private Ltd company,Children,Manufacturing company,Engineering
19 Industries
20 Interiors,Private Ltd company,Industries
21 Private Ltd company
22 Commercial,California
23 Others,Limited liability company,International,Commercial,Corporation,Electronics,Business,Trailer
24 Buildings,Children,Equipment,Coaches
25 Engineering,National,Electronics,USA,Company,Coaches,Manufacturing company
26 Limited liability company,National,Industries,Coaches,Manufacturing company,California,Trailer
27 USA,Coaches,Others,Company,Manufacturing company,Auto
28 Limited liability company,Industries,Business,Engineering,California
29 Others,Men,Manufacturing company,Electronics,Coaches,Limited liability company,Industries
30 International,California,Auto,Others,Corporation,Buildings,National,Trailer,Equipment
31 Corporation,Buildings,Technology,Commercial,Limited liability company
32 Equipment,Technology,Private Ltd company
33 Industries,Equipment,Trailer,Coaches,International,Technology,Manufacturing company
34 National,International,Corporation,Company

```

Recommendation Index table generated by our program for providing recommendations for any user subscribed with any combination of categories.

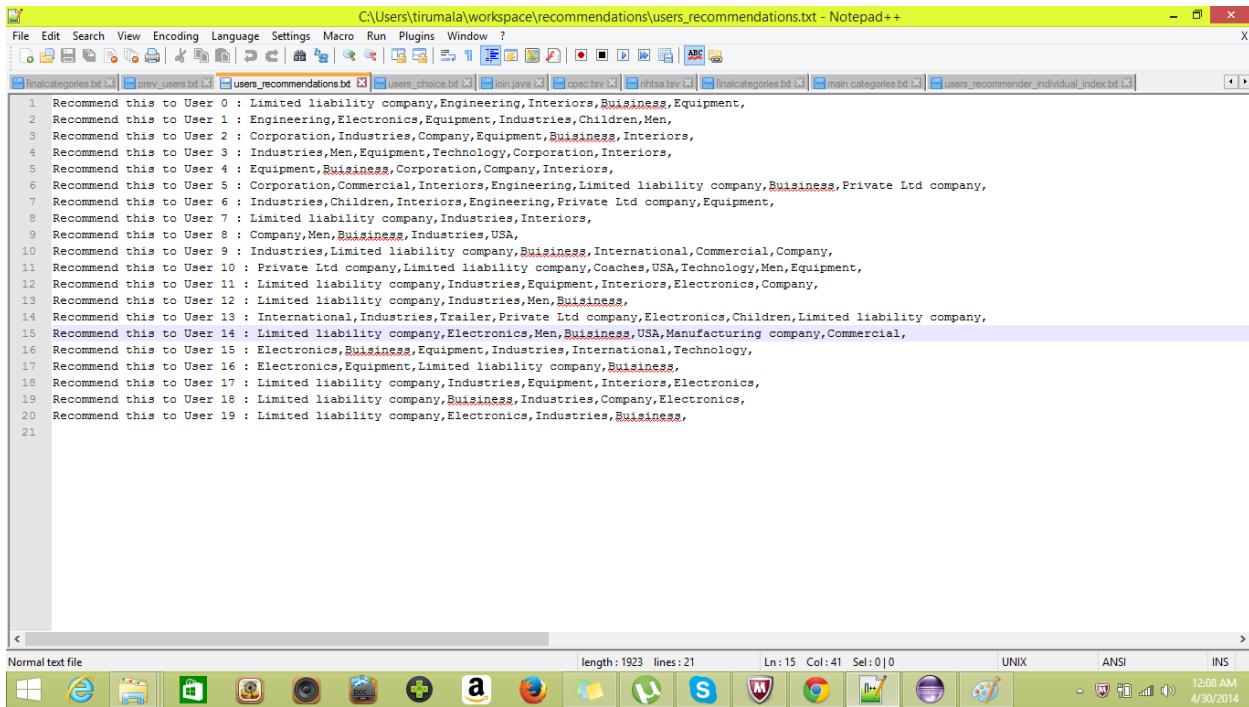
```
1 Interiors::0,Interiors::438,International::-254,Industries::-238,Trailer::-260,Private Ltd company::-226,Electronics::-270,Children::-276,Limited liability company
2 International::0,Interiors::-242,International::426,Industries::-238,Trailer::-246,Private Ltd company::-274,Electronics::-236,Children::-260,Limited liability comp
3 Industries::0,Interiors::-259,International::-271,Industries::459,Trailer::-263,Private Ltd company::-261,Electronics::-283,Children::-287,Limited liability comp
4 Trailer::10,Interiors::-255,International::-253,Industries::-237,Trailer::433,Private Ltd company::-275,Electronics::-247,Children::-271,Limited liability company::-
5 Private Ltd company::0,Interiors::-216,International::-276,Industries::-230,Trailer::-270,Private Ltd company::428,Electronics::-260,Children::-238,Limited liability
6 Electronics::0,Interiors::-269,International::-247,Industries::-261,Trailer::-251,Private Ltd company::-269,Electronics::437,Children::-285,Limited liability comp
7 Children::0,Interiors::-264,International::-260,Industries::-254,Trailer::-264,Private Ltd company::-236,Electronics::-274,Children::426,Limited liability company::-
8 Limited liability company::0,Interiors::-269,International::-287,Industries::-269,Trailer::-301,Private Ltd company::-277,Electronics::-279,Children::-277,Limited li
9 California::0,Interiors::-246,International::-280,Industries::-222,Trailer::-234,Private Ltd company::-266,Electronics::-220,Children::-262,Limited liability comp
10 Corporation::0,Interiors::-264,International::-246,Industries::-232,Trailer::-262,Private Ltd company::-294,Electronics::-254,Children::-290,Limited liability comp
11 Auto::0,Interiors::-220,International::-240,Industries::-232,Trailer::-238,Private Ltd company::-228,Electronics::-252,Children::-236,Limited liability company
12 Coaches::0,Interiors::-240,International::-282,Industries::-250,Trailer::-266,Private Ltd company::-282,Electronics::-242,Children::-240,Limited liability company
13 Engineering::0,Interiors::-280,International::-258,Industries::-270,Trailer::-282,Private Ltd company::-274,Electronics::-230,Children::-270,Limited liability comp
14 USA::0,Interiors::-242,International::-258,Industries::-226,Trailer::-234,Private Ltd company::-256,Electronics::-212,Children::-224,Limited liability company::-260
15 Equipment::0,Interiors::-288,International::-256,Industries::-246,Trailer::-258,Private Ltd company::-248,Electronics::-280,Children::-274,Limited liability comp
16 Commercial::0,Interiors::-241,International::-277,Industries::-251,Trailer::-283,Private Ltd company::-259,Electronics::-267,Children::-267,Limited liability comp
17 Company::0,Interiors::-274,International::-298,Industries::-252,Trailer::-254,Private Ltd company::-276,Electronics::-298,Children::-284,Limited liability company::-
18 Buildings::0,Interiors::-252,International::-240,Industries::-234,Trailer::-252,Private Ltd company::-264,Electronics::-244,Children::-244,Limited liability company
19 National::0,Interiors::-230,International::-244,Industries::-230,Trailer::-232,Private Ltd company::-270,Electronics::-216,Children::-268,Limited liability company
20 Men::0,Interiors::-282,International::-254,Industries::-260,Trailer::-268,Private Ltd company::-278,Electronics::-240,Children::-258,Limited liability company::-226
21 Business::0,Interiors::-255,International::-285,Industries::-259,Trailer::-263,Private Ltd company::-287,Electronics::-253,Children::-267,Limited liability comp
22 Manufacturing company::0,Interiors::-229,International::-271,Industries::-233,Trailer::-261,Private Ltd company::-225,Electronics::-289,Children::-245,Limited liability
23 Others::0,Interiors::-246,International::-240,Industries::-256,Trailer::-252,Private Ltd company::-258,Electronics::-218,Children::-268,Limited liability company::-
24 Technology::0,Interiors::-264,International::-250,Industries::-254,Trailer::-248,Private Ltd company::-262,Electronics::-216,Children::-240,Limited liability comp
25
```

My present user's subscriptions:(to whom we need to provide recommendations now)

```
1 Auto,Children,Coaches,  
2 USA,Business,  
3 Auto,Trailer,  
4 Company,Limited liability company,  
5 Auto,Engineering,Industries,  
6 Auto,  
7 USA,Manufacturing company,  
8 Others,California,Manufacturing company,International,Coaches,  
9 Others,Trailer,Equipment,  
10 Corporation,Interiors,  
11 Children,  
12 California,Private Ltd company,  
13 Others,Corporation,Manufacturing company,National,  
14  
15 Others,  
16 Corporation,Engineering,  
17 Others,Trailer,California,International,  
18 Others,Corporation,Private Ltd company,  
19 Others,Trailer,Interiors,  
20 Corporation,California,National,Coaches,  
21
```

Recommendations provided by my algorithm for above users

This is just a file of reference to show recommendations list provided for user.



C:\Users\tirumala\workspace\recommendations\users_recommendations.txt - Notepad++

```

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
finalCategories.txt privy_users.txt users_recommendations.txt main.java cpsc.txt nhtsa.txt finalCategories.txt main_categories.txt users_recommender_individual_index.txt

1 Recommend this to User 0 : Limited liability company,Engineering,Interiors,Business,Equipment,
2 Recommend this to User 1 : Engineering,Electronics,Equipment,Industries,Children,Men,
3 Recommend this to User 2 : Corporation,Industries,Company,Equipment,Business,Interiors,
4 Recommend this to User 3 : Industries,Men,Equipment,Technology,Corporation,Interiors,
5 Recommend this to User 4 : Equipment,Business,Corporation,Company,Interiors,
6 Recommend this to User 5 : Corporation,Commercial,Interiors,Engineering,Limited liability company,Business,Private Ltd company,
7 Recommend this to User 6 : Industries,Children,Interiors,Engineering,Private Ltd company,Equipment,
8 Recommend this to User 7 : Limited liability company,Industries,Interiors,
9 Recommend this to User 8 : Company,Men,Business,Industries,USA,
10 Recommend this to User 9 : Industries,Limited liability company,Business,International,Commercial,Company,
11 Recommend this to User 10 : Private Ltd company,Limited liability company,Coaches,USA,Technology,Men,Equipment,
12 Recommend this to User 11 : Limited liability company,Industries,Equipment,Interiors,Electronics,Company,
13 Recommend this to User 12 : Limited liability company,Industries,Men,Business,
14 Recommend this to User 13 : International,Industries,Trailer,Private Ltd company,Electronics,Children,Limited liability company,
15 Recommend this to User 14 : Limited liability company,Electronics,Men,Business,USA,Manufacturing company,Commercial,
16 Recommend this to User 15 : Electronics,Business,Equipment,Industries,International,Technology,
17 Recommend this to User 16 : Electronics,Equipment,Limited liability company,Business,
18 Recommend this to User 17 : Limited liability company,Industries,Equipment,Interiors,Electronics,
19 Recommend this to User 18 : Limited liability company,Business,Industries,Company,Electronics,
20 Recommend this to User 19 : Limited liability company,Electronics,Industries,Business,
21

```

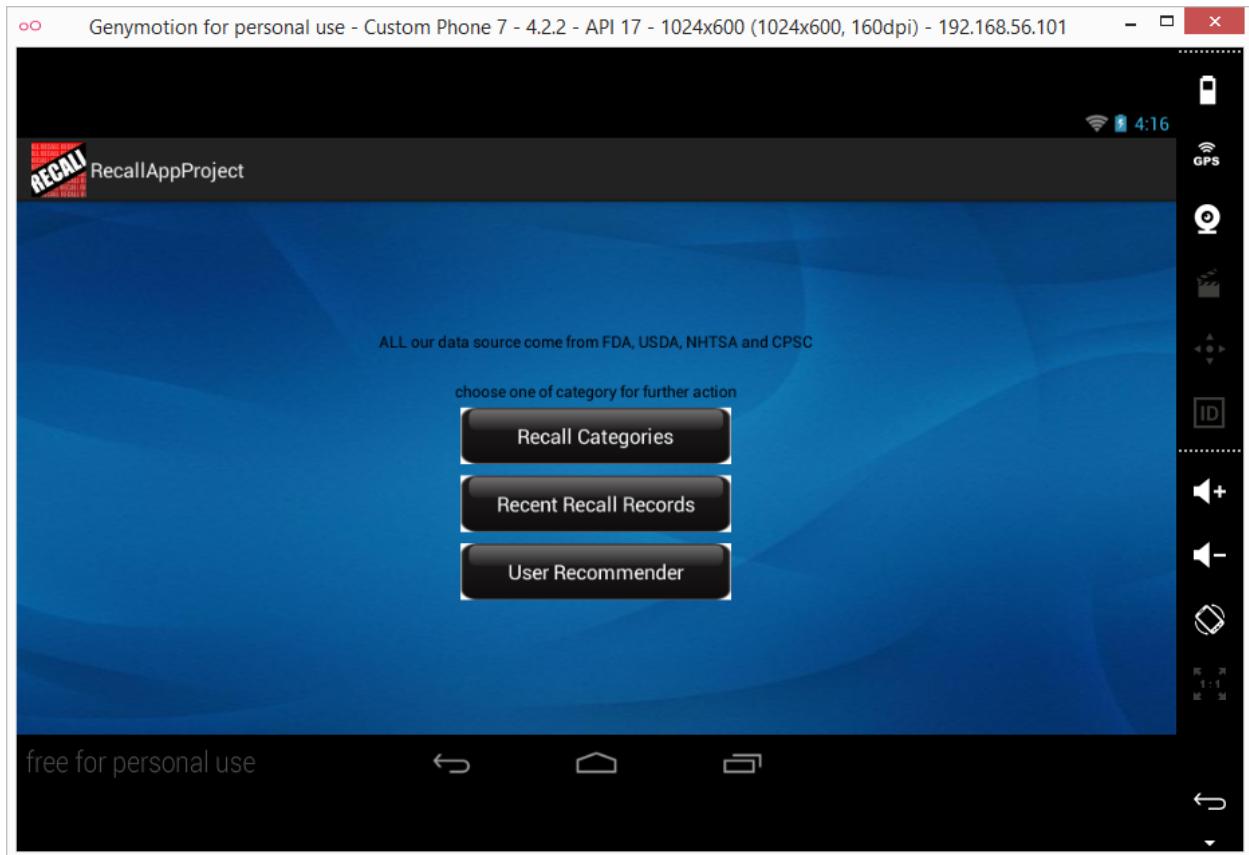
Normal text file

length : 1923 lines : 21 Ln : 15 Col : 41 Sel : 0 | 0 UNIX ANSI INS

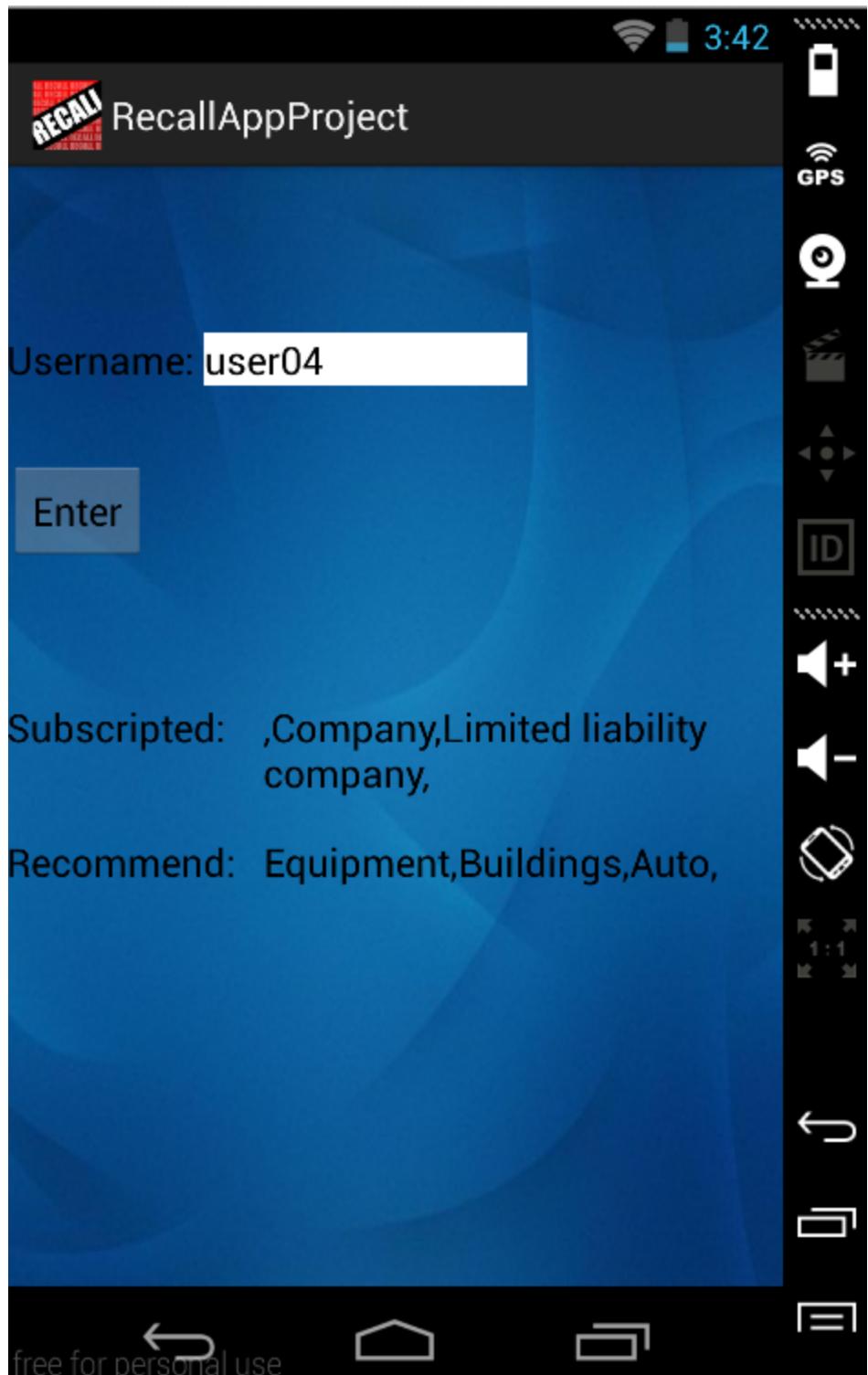
12:08 AM 4/30/2014

Implementation of Android Client

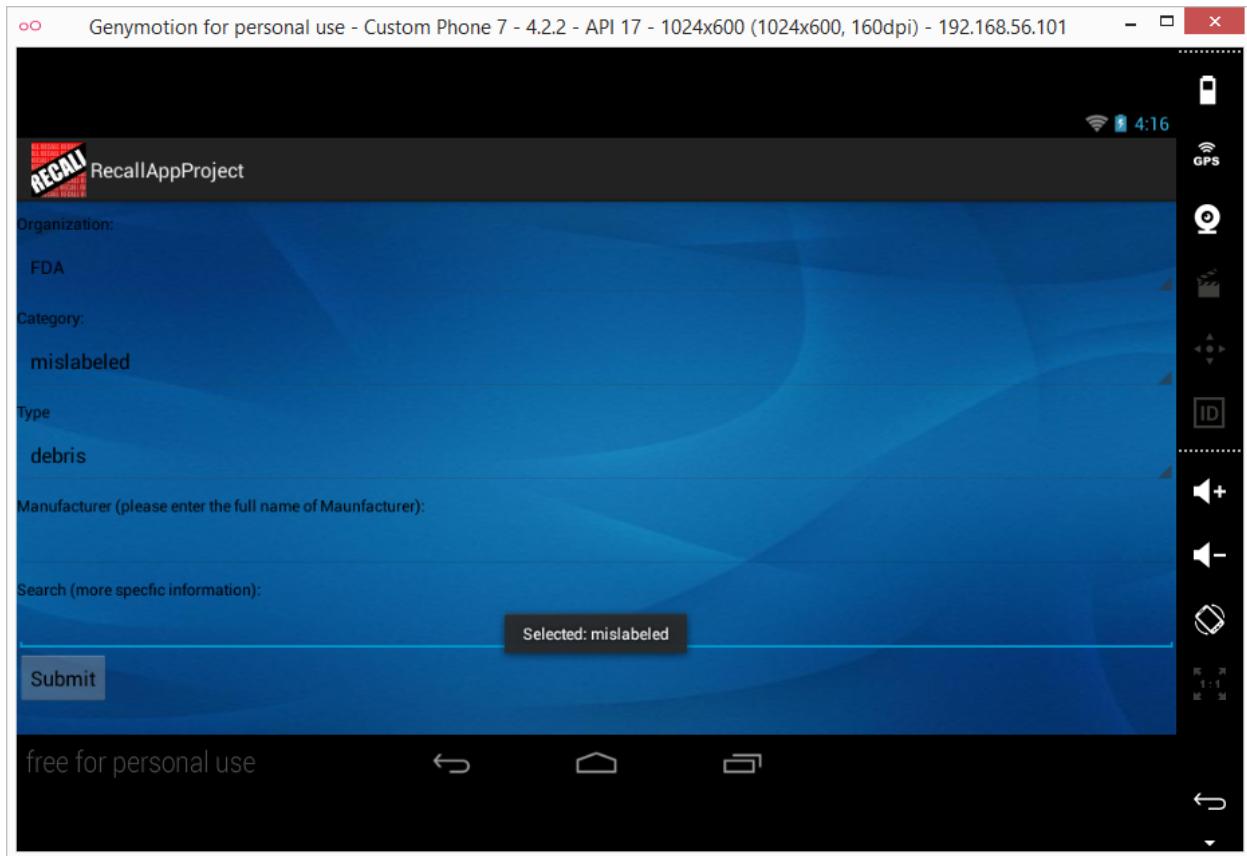
this is the main screen for our app now.



After we done the classification for data from 4 different organization and recommendation for user. I added a new view for the app, user need to enter their username, base on their subscription, out recommendation algorithm made 3 possible recommend categories for user.



I restructure our app UI. Now the user will choose the organization first , then for each organization, there are some classified categories. And also for each category, there are some different classified subcategories. Because there are too many classified subcategories for NHTSA recalls. So I made a editText view to let user type any manufacturer they wanna see. Also I made another editText view for user to search any additional information.



Appendix

Hadoop and Mahout linux shell commands for classification

This listing uses classes found at [fredang/mahout-naive-bayes-example](https://github.com/fredang/mahout-naive-bayes-example) project on Github

1. Convert training set to Hadoop sequence file format

```
java -cp target/tnb.jar com.chimpler.example.bayes.TweetTSVToSeq
data/fda-training.tsv fda-seq
```

2. Upload sequence file to HDFS

```
hadoop fs -put fda-seq fda-seq
```

3. Transform training sets to vectors with Mahout (using term frequency by document frequency weights)

```
mahout seq2sparse -i fda-seq -o fda-vectors
```

4. Split data into training set and testing set

```
mahout split -i fda-vectors/tfidf-vectors --trainingOutput
train-fda-vectors --testOutput test-fda-vectors
--randomSelectionPct 40 --overwrite --sequenceFiles -xm
sequential
```

5. Use training set to train classifier

```
mahout trainnb -i train-fda-vectors -el -li labelindex -o model
-ow -c
```

6. Test the classifier on the training set

```
mahout testnb -i train-fda-vectors -m model -l labelindex -ow -o
fda-testing -c
```

7. Test the classifier on the testing set

```
mahout testnb -i test-fda-vectors -m model -l labelindex -ow -o
fda-testing -c
```

8. To use classifier and classify new documents, copy several files from HDFS

```
hadoop fs -get labelindex labelindex
hadoop fs -get model model
hadoop fs -get fda-vectors/dictionary.file-0 dictionary.file-0
hadoop fs -getmerge fda-vectors/df-count df-count
```

9. Classify the new file (optional section in bold red writes to text file)

```
java -cp target/tnb.jar com.chimpler.example.bayes.Classifier
model labelindex dictionary.file-0 df-count data/fda.tsv >
fda-classified-result.txt
```