Increment 2

Group 6: LeeStudents
James Clark
Tirumala Konireddy
Jagadish Tirumalasetty
Wang Zhang

Dr. Yugyung Lee
CS 560: KDM
April 1, 2014

Data

It turns out there's an ethical approach to consider when crawling sites. Two things to immediately consider: 1) robots.txt usually found at domain.com/robots.txt and 2) the robots meta tag such as <meta name="robots" content="index, follow" />. In the case of this site, it is acceptable to crawl for more data. I will conduct such a crawl at a metered pace, something like one request per 15 seconds. Results should be stored to prevent the need for repeated crawling of the same pages. MongoDB is probably ideal for this. More info is available at
http://www.metatags.info/meta_name_robots
http://www.robotstxt.org/robotstxt.html

Discovery

- For different algorithms, data format is different[1], therefore we need a data format processor.

Suggested workflow:
Store data set in DB such as HBase or MongoDB.
Get data from DB
Convert data to format
Process with Mahout

- For the classifier to work properly, this set (Naive Bayes classifier) must have at least 50 tweets messages in each category.[2]

The CPSC has an API:
http://www.cpsc.gov/en/Recalls/CPSC-Recalls-Application-Program-Interface-API-Information/
Using this API, I found 5000+ links to recall data in date range 1972-01-01 to 2015-01-01

USA.gov has an API:
http://search.digitalgov.gov/developer/recalls.html

Algorithms

The CPSC provides an ASMX web service endpoint, and they serve data from 1973 to current. A one-time process is needed to gather all recall data. After that, a periodic check is needed to look for the latest updates.

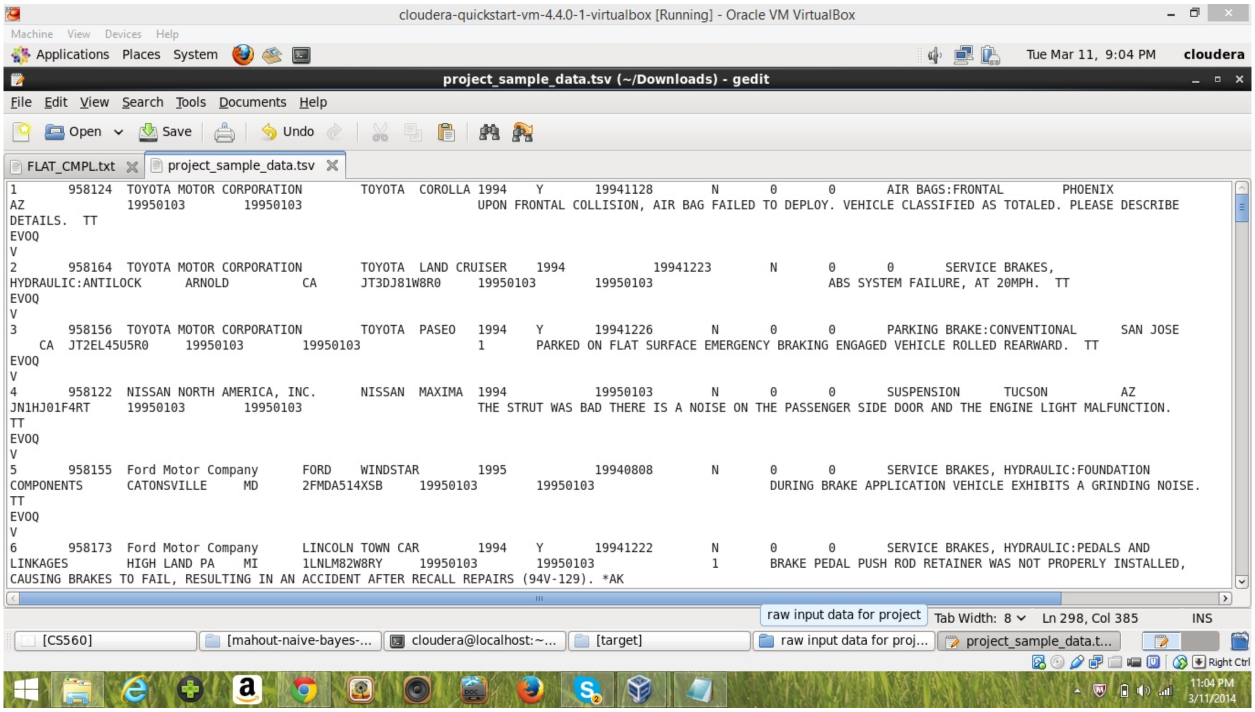
Connect to CPSC web service, http://www.cpsc.gov/cgibin/CPSCUpcWS/CPSCUpcSvc.asmx
Invoke getRecallByDate method
foreach result in list do

[1] Feichen, Tutorial 7, Page 18
[2] Feichen, Tutorial 7, Page 19

```
start json object
store result attributes to json object
visit result url
parse web page at url
store fields from html into json object
return json to caller
```

Analytical Tasks

1. Collection of data : We get data from http://www-odi.nhtsa.gov/downloads/flatfiles.cfm about all the product recalls that had done in past till now. Now, we are manually monitor and download the updated recall data flat .txt file.



cloudera-quickstart-vm-4.4.0-1-virtualbox [Running] - Oracle VM VirtualBox
Machine View Devices Help
Applications Places System
Tue Mar 11, 9:04 PM
cloudera
project_sample_data.tsv (~/Downloads) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
FLAT_CMPL.txt
project_sample_data.tsv
raw input data for project
Tab Width: 8
Ln 298, Col 385
INS
[CS560]
[mahout-naive-bayes-...
cloudera@localhost:~...
[target]
raw input data for proj...
project_sample_data.t...
Right Ctrl
11:04 PM
3/11/2014

```
1       958124  TOYOTA MOTOR CORPORATION        TOYOTA  COROLLA 1994    Y       19941128        N       0       0       AIR BAGS:FRONTAL        PHOENIX
AZ              19950103        19950103                        UPON FRONTAL COLLISION, AIR BAG FAILED TO DEPLOY. VEHICLE CLASSIFIED AS TOTALED. PLEASE DESCRIBE
DETAILS.  TT
EVOQ
V
2       958164  TOYOTA MOTOR CORPORATION        TOYOTA  LAND CRUISER    1994            19941223        N       0       0       SERVICE BRAKES,
HYDRAULIC:ANTILOCK      ARNOLD          CA      JT3DJ81W8R0     19950103        19950103                        ABS SYSTEM FAILURE, AT 20MPH.  TT
EVOQ
V
3       958156  TOYOTA MOTOR CORPORATION        TOYOTA  PASEO   1994    Y       19941226        N       0       0       PARKING BRAKE:CONVENTIONAL      SAN JOSE
    CA  JT2EL45U5R0     19950103        19950103                1       PARKED ON FLAT SURFACE EMERGENCY BRAKING ENGAGED VEHICLE ROLLED REARWARD.  TT
EVOQ
V
4       958122  NISSAN NORTH AMERICA, INC.      NISSAN  MAXIMA  1994            19950103        N       0       0       SUSPENSION      TUCSON          AZ
JN1HJ01F4RT     19950103        19950103                        THE STRUT WAS BAD THERE IS A NOISE ON THE PASSENGER SIDE DOOR AND THE ENGINE LIGHT MALFUNCTION.
TT
EVOQ
V
5       958155  Ford Motor Company      FORD    WINDSTAR        1995            19940808        N       0       0       SERVICE BRAKES, HYDRAULIC:FOUNDATION
COMPONENTS      CATONSVILLE     MD      2FMDA514XSB     19950103        19950103                DURING BRAKE APPLICATION VEHICLE EXHIBITS A GRINDING NOISE.
TT
EVOQ
V
6       958173  Ford Motor Company      LINCOLN TOWN CAR        1994    Y       19941222        N       0       0       SERVICE BRAKES, HYDRAULIC:PEDALS AND
LINKAGES        HIGH LAND PA    MI      1LNLM82W8RY     19950103        19950103        1       BRAKE PEDAL PUSH ROD RETAINER WAS NOT PROPERLY INSTALLED,
CAUSING BRAKES TO FAIL, RESULTING IN AN ACCIDENT AFTER RECALL REPAIRS (94V-129). *AK
```

2. Classification of data :

We are using naive bayes algorithm to classify user comments to a recall category. We used mahout tool for naive bayes classification. The following steps are followed to classify data using mahout.

(i) Converting data into sequential files:

Mahout works directly only on sequential files, So to convert this data which had a tab separator between every two columns in a line, we had developed a java program “CarsTSVToSeq”.

(ii) Upload data :

After converting data into sequential files, we have to upload the sequential file to hadoop on which mahout directly works on.

(iii) Training and Testing:

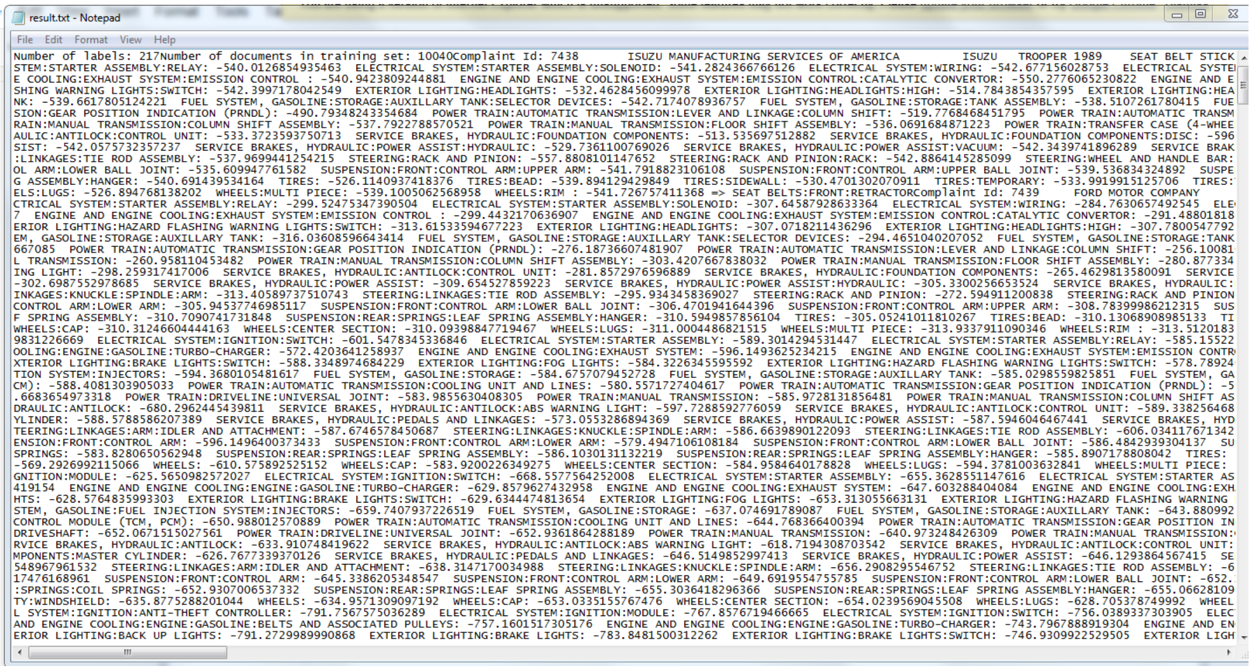
We had developed a Classifier, trainer java files using bayes classes in java to do train the classifier with the given seq files.

We had trained the classifier with varying number of input data records.

We got 90.67% accuracy in testing the trained data entries when we trained 10040 lines of entries, whereas we got only 77.4%, 55.08% accuracy when we trained with 100,000 and 967,487 data records.

(iv) Classification :

We had used 217 data entries for classification and we found good results by cross-verification.



```
Number of labels: 217Number of documents in training set: 10040Complaint Id: 7438       ISUZU MANUFACTURING SERVICES OF AMERICA       ISUZU   TROOPER 1989    SEAT BELT STICK
STEM:STARTER ASSEMBLY:RELAY: -540.0126854935463  ELECTRICAL SYSTEM:STARTER ASSEMBLY:SOLENOID: -541.2824366766126  ELECTRICAL SYSTEM:WIRING: -542.677156028753  ELECTRICAL SYSTE
E COOLING:EXHAUST SYSTEM:EMISSION CONTROL : -540.9423809244881  ENGINE AND ENGINE COOLING:EXHAUST SYSTEM:EMISSION CONTROL:CATALYTIC CONVERTOR: -550.2776065230822  ENGINE AND E
SHING WARNING LIGHTS:SWITCH: -542.3997178042549  EXTERIOR LIGHTING:HEADLIGHTS: -532.4628456099978  EXTERIOR LIGHTING:HEADLIGHTS:HIGH: -514.7843854357595  EXTERIOR LIGHTING:HEA
NK: -539.6617805124221  FUEL SYSTEM, GASOLINE:STORAGE:AUXILLARY TANK:SELECTOR DEVICES: -542.7174078936757  FUEL SYSTEM, GASOLINE:STORAGE:TANK ASSEMBLY: -538.5107261780415  FUE
SION:GEAR POSITION INDICATION (PRNDL): -490.79348243354684  POWER TRAIN:AUTOMATIC TRANSMISSION:LEVER AND LINKAGE:COLUMN SHIFT: -519.7768468451795  POWER TRAIN:AUTOMATIC TRANSM
RAIN:MANUAL TRANSMISSION:COLUMN SHIFT ASSEMBLY: -537.7922788570521  POWER TRAIN:MANUAL TRANSMISSION:FLOOR SHIFT ASSEMBLY: -536.0691684871223  POWER TRAIN:TRANSFER CASE (4-WHEE
AULIC:ANTILOCK:CONTROL UNIT: -533.3723593750713  SERVICE BRAKES, HYDRAULIC:FOUNDATION COMPONENTS: -513.535697512882  SERVICE BRAKES, HYDRAULIC:FOUNDATION COMPONENTS:DISC: -596
SIST: -542.0575732357237  SERVICE BRAKES, HYDRAULIC:POWER ASSIST:HYDRAULIC: -529.7361100769026  SERVICE BRAKES, HYDRAULIC:POWER ASSIST:VACUUM: -542.3439741896289  SERVICE BRAK
:LINKAGES:TIE ROD ASSEMBLY: -537.9699441254215  STEERING:RACK AND PINION: -557.8808101147652  STEERING:RACK AND PINION:RACK: -542.8864145285099  STEERING:WHEEL AND HANDLE BAR:
OL ARM:LOWER BALL JOINT: -535.609947761582  SUSPENSION:FRONT:CONTROL ARM:UPPER ARM: -541.7918823106108  SUSPENSION:FRONT:CONTROL ARM:UPPER BALL JOINT: -539.536834324892  SUSPE
G ASSEMBLY:HANGER: -540.691439534164  TIRES: -526.1140937418376  TIRES:BEAD: -539.894129429849  TIRES:SIDEWALL: -530.4701302070911  TIRES:TEMPORARY: -533.9919915125706  TIRES:
ELS:LUGS: -526.894768138202  WHEELS:MULTI PIECE: -539.1005062568958  WHEELS:RIM : -541.726757411368 => SEAT BELTS:FRONT:RETRACTORComplaint Id: 7439      FORD MOTOR COMPANY
CTRICAL SYSTEM:STARTER ASSEMBLY:RELAY: -299.52475347390504  ELECTRICAL SYSTEM:STARTER ASSEMBLY:SOLENOID: -307.64587928633364  ELECTRICAL SYSTEM:WIRING: -284.7630657492545  ELE
7  ENGINE AND ENGINE COOLING:EXHAUST SYSTEM:EMISSION CONTROL : -299.4432170636907  ENGINE AND ENGINE COOLING:EXHAUST SYSTEM:EMISSION CONTROL:CATALYTIC CONVERTOR: -291.48801818
ERIOR LIGHTING:HAZARD FLASHING WARNING LIGHTS:SWITCH: -313.61533594677223  EXTERIOR LIGHTING:HEADLIGHTS: -307.0718211436296  EXTERIOR LIGHTING:HEADLIGHTS:HIGH: -307.7800547792
EM, GASOLINE:STORAGE:AUXILLARY TANK: -316.03608596643414  FUEL SYSTEM, GASOLINE:STORAGE:AUXILLARY TANK:SELECTOR DEVICES: -294.4651040207052  FUEL SYSTEM, GASOLINE:STORAGE:TANK
667085  POWER TRAIN:AUTOMATIC TRANSMISSION:GEAR POSITION INDICATION (PRNDL): -276.18736607481907  POWER TRAIN:AUTOMATIC TRANSMISSION:LEVER AND LINKAGE:COLUMN SHIFT: -256.10081
L TRANSMISSION: -260.958110453482  POWER TRAIN:MANUAL TRANSMISSION:COLUMN SHIFT ASSEMBLY: -303.4207667838032  POWER TRAIN:MANUAL TRANSMISSION:FLOOR SHIFT ASSEMBLY: -280.877334
ING LIGHT: -298.259317417006  SERVICE BRAKES, HYDRAULIC:ANTILOCK:CONTROL UNIT: -281.8572976596889  SERVICE BRAKES, HYDRAULIC:FOUNDATION COMPONENTS: -265.4629813580091  SERVICE
-302.6987552978685  SERVICE BRAKES, HYDRAULIC:POWER ASSIST: -309.654527859223  SERVICE BRAKES, HYDRAULIC:POWER ASSIST:HYDRAULIC: -305.3300256653524  SERVICE BRAKES, HYDRAULIC:
INKAGES:KNUCKLE:SPINDLE:ARM: -313.40589737510743  STEERING:LINKAGES:TIE ROD ASSEMBLY: -295.9343458369027  STEERING:RACK AND PINION: -272.594911200838  STEERING:RACK AND PINION
CONTROL ARM:LOWER ARM: -305.94537746985117  SUSPENSION:FRONT:CONTROL ARM:LOWER BALL JOINT: -306.4701941644396  SUSPENSION:FRONT:CONTROL ARM:UPPER ARM: -308.78399986212315  SUS
F SPRING ASSEMBLY: -310.7090741731848  SUSPENSION:REAR:SPRINGS:LEAF SPRING ASSEMBLY:HANGER: -310.5949857856104  TIRES: -305.05241011810267  TIRES:BEAD: -310.13068908985133  TI
WHEELS:CAP: -310.31246604444163  WHEELS:CENTER SECTION: -310.09398847719467  WHEELS:LUGS: -311.0004486821515  WHEELS:MULTI PIECE: -313.9337911090346  WHEELS:RIM : -313.5120183
9831226669  ELECTRICAL SYSTEM:IGNITION:SWITCH: -601.5478345336846  ELECTRICAL SYSTEM:STARTER ASSEMBLY: -589.3014294531447  ELECTRICAL SYSTEM:STARTER ASSEMBLY:RELAY: -585.15522
OOLING:ENGINE:GASOLINE:TURBO-CHARGER: -572.4203641258937  ENGINE AND ENGINE COOLING:EXHAUST SYSTEM: -596.1493625234215  ENGINE AND ENGINE COOLING:EXHAUST SYSTEM:EMISSION CONTR
XTERIOR LIGHTING:BRAKE LIGHTS:SWITCH: -588.3348974684229  EXTERIOR LIGHTING:FOG LIGHTS: -584.3226345595592  EXTERIOR LIGHTING:HAZARD FLASHING WARNING LIGHTS:SWITCH: -578.78924
TION SYSTEM:INJECTORS: -594.3680105481617  FUEL SYSTEM, GASOLINE:STORAGE: -584.6757079452728  FUEL SYSTEM, GASOLINE:STORAGE:AUXILLARY TANK: -585.0298559825851  FUEL SYSTEM, GA
CM): -588.4081303905033  POWER TRAIN:AUTOMATIC TRANSMISSION:COOLING UNIT AND LINES: -580.5571727404617  POWER TRAIN:AUTOMATIC TRANSMISSION:GEAR POSITION INDICATION (PRNDL): -5
.6683654973318  POWER TRAIN:DRIVELINE:UNIVERSAL JOINT: -583.9855630408305  POWER TRAIN:MANUAL TRANSMISSION: -585.9728131856481  POWER TRAIN:MANUAL TRANSMISSION:COLUMN SHIFT AS
DRAULIC:ANTILOCK: -680.2962445439811  SERVICE BRAKES, HYDRAULIC:ANTILOCK:ABS WARNING LIGHT: -597.7288592776059  SERVICE BRAKES, HYDRAULIC:ANTILOCK:CONTROL UNIT: -589.338256468
YLINDER: -588.5788586207389  SERVICE BRAKES, HYDRAULIC:PEDALS AND LINKAGES: -573.0553286894369  SERVICE BRAKES, HYDRAULIC:POWER ASSIST: -587.5946046467441  SERVICE BRAKES, HYD
TEERING:LINKAGES:ARM:IDLER AND ATTACHMENT: -587.6746578450687  STEERING:LINKAGES:KNUCKLE:SPINDLE:ARM: -586.6639890122093  STEERING:LINKAGES:TIE ROD ASSEMBLY: -606.034117671342
ENSION:FRONT:CONTROL ARM: -596.1496400373433  SUSPENSION:FRONT:CONTROL ARM:LOWER ARM: -579.4947106108184  SUSPENSION:FRONT:CONTROL ARM:LOWER BALL JOINT: -586.4842939304137  SU
SPRINGS: -583.8280650562948  SUSPENSION:REAR:SPRINGS:LEAF SPRING ASSEMBLY: -586.1030131132219  SUSPENSION:REAR:SPRINGS:LEAF SPRING ASSEMBLY:HANGER: -585.8907178808042  TIRES:
-569.2926992115066  WHEELS: -610.575892525152  WHEELS:CAP: -583.9200226349275  WHEELS:CENTER SECTION: -584.9584640178828  WHEELS:LUGS: -594.3781003632841  WHEELS:MULTI PIECE:
GNITION:MODULE: -625.5650982572027  ELECTRICAL SYSTEM:IGNITION:SWITCH: -668.5577564252008  ELECTRICAL SYSTEM:STARTER ASSEMBLY: -655.3628551147616  ELECTRICAL SYSTEM:STARTER AS
419154  ENGINE AND ENGINE COOLING:ENGINE:GASOLINE:TURBO-CHARGER: -629.8579627432958  ENGINE AND ENGINE COOLING:EXHAUST SYSTEM: -647.603288404084  ENGINE AND ENGINE COOLING:EXH
HTS: -628.5764835993303  EXTERIOR LIGHTING:BRAKE LIGHTS:SWITCH: -629.6344474813654  EXTERIOR LIGHTING:FOG LIGHTS: -653.313055663131  EXTERIOR LIGHTING:HAZARD FLASHING WARNING
STEM, GASOLINE:FUEL INJECTION SYSTEM:INJECTORS: -659.7407937226519  FUEL SYSTEM, GASOLINE:STORAGE: -637.074691789087  FUEL SYSTEM, GASOLINE:STORAGE:AUXILLARY TANK: -643.880992
CONTROL MODULE (TCM, PCM): -650.988012570889  POWER TRAIN:AUTOMATIC TRANSMISSION:COOLING UNIT AND LINES: -644.768366400394  POWER TRAIN:AUTOMATIC TRANSMISSION:GEAR POSITION IN
DRIVESHAFT: -652.0671515027561  POWER TRAIN:DRIVELINE:UNIVERSAL JOINT: -652.9361864288189  POWER TRAIN:MANUAL TRANSMISSION: -640.973248426309  POWER TRAIN:MANUAL TRANSMISSION:
RVICE BRAKES, HYDRAULIC:ANTILOCK: -633.910748419622  SERVICE BRAKES, HYDRAULIC:ANTILOCK:ABS WARNING LIGHT: -618.7194308703542  SERVICE BRAKES, HYDRAULIC:ANTILOCK:CONTROL UNIT:
MPONENTS:MASTER CYLINDER: -626.7677339370126  SERVICE BRAKES, HYDRAULIC:PEDALS AND LINKAGES: -646.5149852997413  SERVICE BRAKES, HYDRAULIC:POWER ASSIST: -646.1293864567415  SE
548967961532  STEERING:LINKAGES:ARM:IDLER AND ATTACHMENT: -638.3147170034988  STEERING:LINKAGES:KNUCKLE:SPINDLE:ARM: -656.2908295546752  STEERING:LINKAGES:TIE ROD ASSEMBLY: -6
17476168961  SUSPENSION:FRONT:CONTROL ARM: -645.3386205348547  SUSPENSION:FRONT:CONTROL ARM:LOWER ARM: -649.6919554755785  SUSPENSION:FRONT:CONTROL ARM:LOWER BALL JOINT: -652.
:SPRINGS:COIL SPRINGS: -652.9307006537332  SUSPENSION:REAR:SPRINGS:LEAF SPRING ASSEMBLY: -655.3036418296366  SUSPENSION:REAR:SPRINGS:LEAF SPRING ASSEMBLY:HANGER: -655.06628109
TY:WINDSHIELD: -635.8775288201044  WHEELS: -634.9571309097192  WHEELS:CAP: -653.0335155767476  WHEELS:CENTER SECTION: -654.0239569045508  WHEELS:LUGS: -628.705378749992  WHEEL
L SYSTEM:IGNITION:ANTI-THEFT CONTROLLER: -791.7567575036289  ELECTRICAL SYSTEM:IGNITION:MODULE: -767.8576719466665  ELECTRICAL SYSTEM:IGNITION:SWITCH: -756.0389337303905  ELEC
AND ENGINE COOLING:ENGINE:GASOLINE:BELTS AND ASSOCIATED PULLEYS: -757.1601517305176  ENGINE AND ENGINE COOLING:ENGINE:GASOLINE:TURBO-CHARGER: -743.7967888919304  ENGINE AND EN
ERIOR LIGHTING:BACK UP LIGHTS: -791.2729989990868  EXTERIOR LIGHTING:BRAKE LIGHTS: -783.8481500312262  EXTERIOR LIGHTING:BRAKE LIGHTS:SWITCH: -746.9309922529505  EXTERIOR LIGH
```

(v) Uploading results to Solr:

Because of some issues to work with Glassfish server, we are unable to upload this results to Solr using web restful service.

Implementation

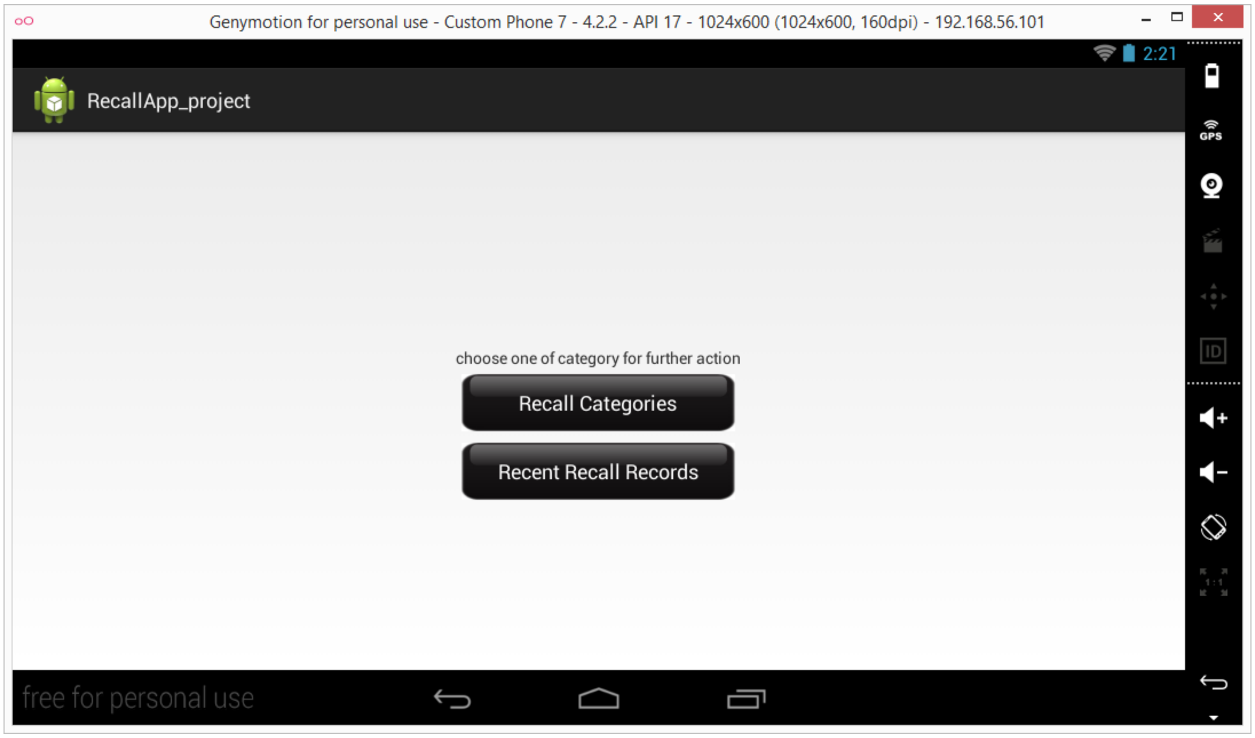
Implementation of data model and algorithms (Machine Learning)

We are using flat data tsv files which contains details of about complaint id, manufacturer, model, recalled component description, number of deaths involved in fail of such component, customer description of that component failure etc., in 47 columns using a tab separator between every two columns.

We are using naive bayes algorithm which uses bayes probability by assuming every occurrence of event as independent for training a classifier to classify user comments to a recalled component category. For implementing this naive bayes algorithm we used hadoop files system, mahout tool for classifying data. We had used Solr server to store the classified results so that they will be available for accessing by end user. We are developing a web service to convert the classified results into json format and upload them to Solr server using Glassfish server.

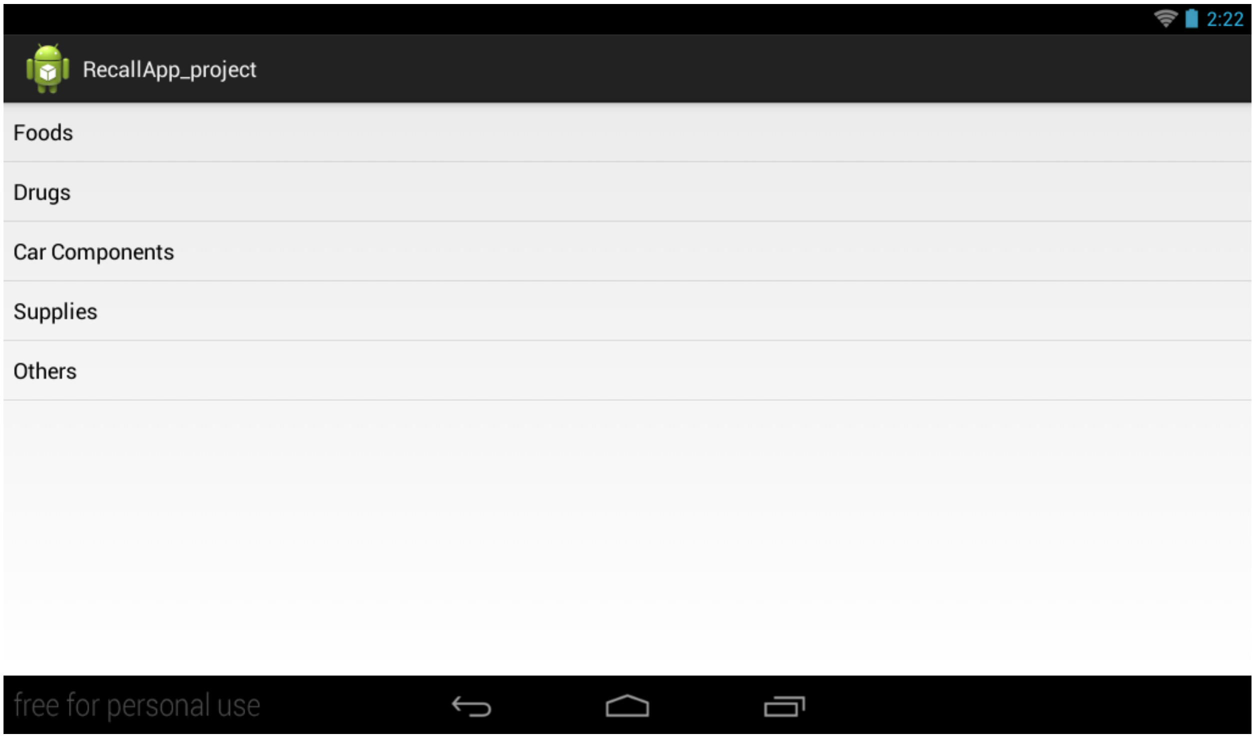
Application Interface

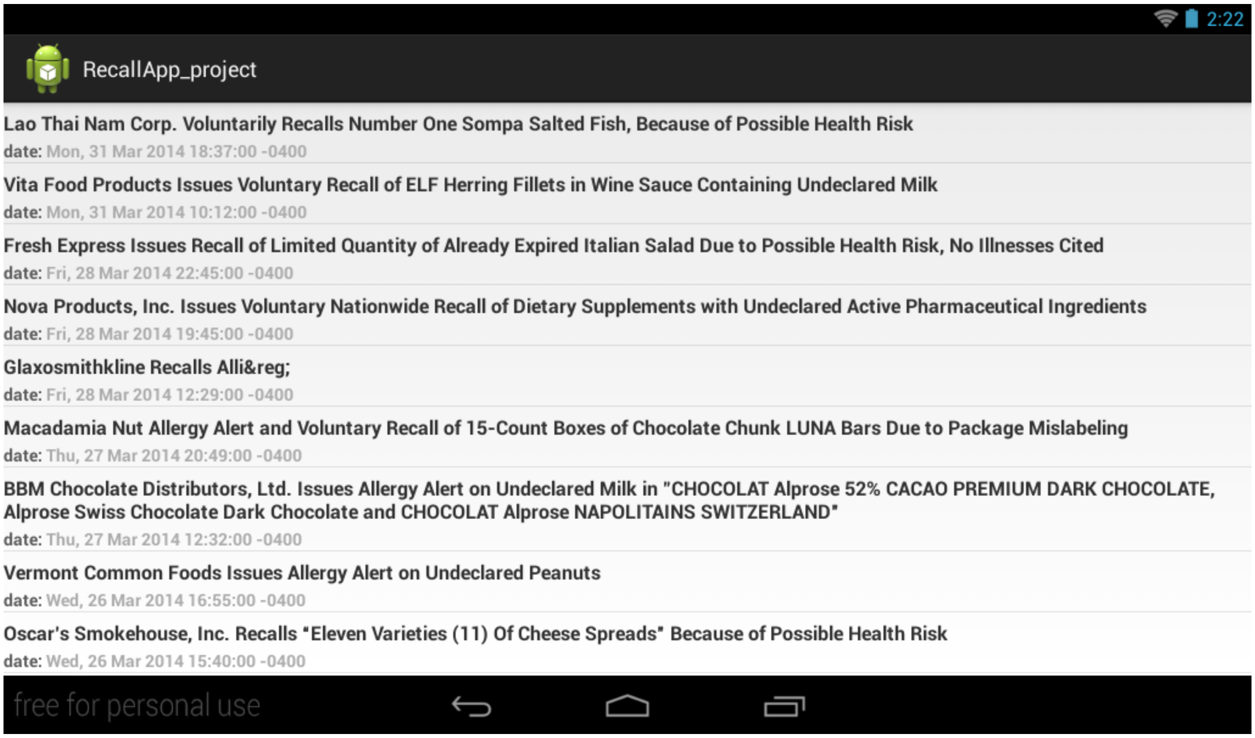
Most Likely finish the Application interface, furthermore we will change xml to json or HTML format. Here is some screenshots for our app.



Genymotion for personal use - Custom Phone 7 - 4.2.2 - API 17 - 1024x600 (1024x600, 160dpi) - 192.168.56.101

choose one of category for further action





Lao Thai Nam Corp. Voluntarily Recalls Number One Sompa Salted Fish, Because of Possible Health Risk
date: Mon, 31 Mar 2014 18:37:00 -0400

Vita Food Products Issues Voluntary Recall of ELF Herring Fillets in Wine Sauce Containing Undeclared Milk
date: Mon, 31 Mar 2014 10:12:00 -0400

Fresh Express Issues Recall of Limited Quantity of Already Expired Italian Salad Due to Possible Health Risk, No Illnesses Cited
date: Fri, 28 Mar 2014 22:45:00 -0400

Nova Products, Inc. Issues Voluntary Nationwide Recall of Dietary Supplements with Undeclared Active Pharmaceutical Ingredients
date: Fri, 28 Mar 2014 19:45:00 -0400

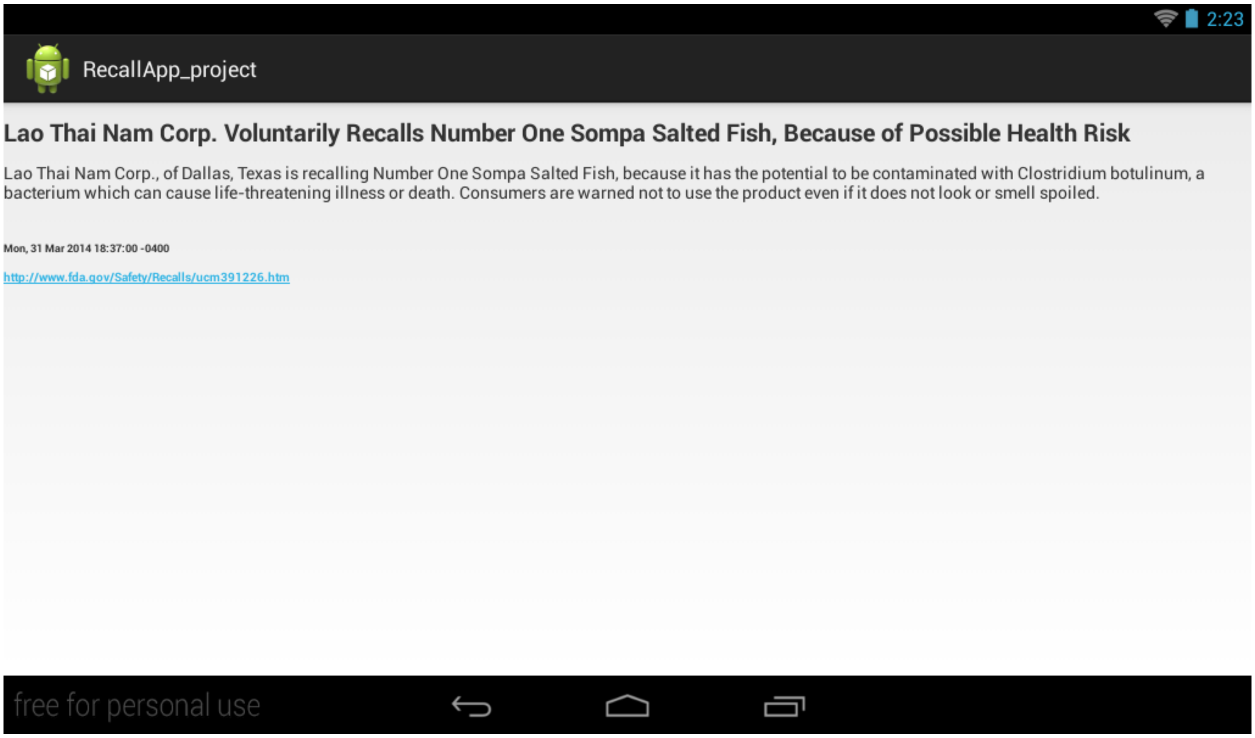
Glaxosmithkline Recalls Alli&reg;
date: Fri, 28 Mar 2014 12:29:00 -0400

Macadamia Nut Allergy Alert and Voluntary Recall of 15-Count Boxes of Chocolate Chunk LUNA Bars Due to Package Mislabeling
date: Thu, 27 Mar 2014 20:49:00 -0400

BBM Chocolate Distributors, Ltd. Issues Allergy Alert on Undeclared Milk in "CHOCOLAT Alprose 52% CACAO PREMIUM DARK CHOCOLATE, Alprose Swiss Chocolate Dark Chocolate and CHOCOLAT Alprose NAPOLITAINS SWITZERLAND"
date: Thu, 27 Mar 2014 12:32:00 -0400

Vermont Common Foods Issues Allergy Alert on Undeclared Peanuts
date: Wed, 26 Mar 2014 16:55:00 -0400

Oscar's Smokehouse, Inc. Recalls "Eleven Varieties (11) Of Cheese Spreads" Because of Possible Health Risk
date: Wed, 26 Mar 2014 15:40:00 -0400



Lao Thai Nam Corp. Voluntarily Recalls Number One Sompa Salted Fish, Because of Possible Health Risk

Lao Thai Nam Corp., of Dallas, Texas is recalling Number One Sompa Salted Fish, because it has the potential to be contaminated with Clostridium botulinum, a bacterium which can cause life-threatening illness or death. Consumers are warned not to use the product even if it does not look or smell spoiled.

Mon, 31 Mar 2014 18:37:00 -0400

http://www.fda.gov/Safety/Recalls/ucm391226.htm