FAKE REVIEW DETECTION

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Radhika Vaidya (131276)

Under the supervision of

Dr.Rajni Mohana

to

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat,
Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Fake Review Detection"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2016 to May 2017 under the supervision of **Dr.Rajni Mohana**(Assistant Professor Senior Grade (Computer Science Department).
The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Radhika Vaidya,131276

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)
Dr.Rajni Mohana
Assistant Professor (Senior Grade)
Computer Science Department
Dated:

Acknowledgement

I owe my profound gratitude to my project supervisor **Dr. Rajni Mohana**, who took keen interest and guided me all along in my project work titled **—Fake Review Detection** which is used to segregate the reviews whether it's true or fake , till the completion of my project by providing all the necessary information for developing the project. The project development helped me in research and I got to know a lot of new things this domain. I am really thankful to her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services. Reviews provide feedback to the companies about their product for any kind of improvement. The huge impact of reviews on customer's decision making motivates wrongdoers to create fake reviews to deliberately promote or demote a product. This is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for their profit. In order to provide right information to the customer detection of fake reviews is important. Manual detection of fake reviews is a time consuming task therefore we need an automated technique to detect the fake reviews. Natural Languages Processing(NLP) can be used to extract meaningful features from text content of reviews therefore it is possible to detect fake reviews using various machine learning techniques. In order to influence people fake review writers try to use words or topics that create an impact on readers mind. This difference in word choice pattern in fake and truthful review can be used as a method to identify fake reviews. Our work is based on this topic type differentiator to evaluate individual reviews. It is seen that fake review writers use words that are different from the truthful ones based on this an automated method is created using various machine learning techniques to segregate fake and truthful reviews. The method improves the efficiency and performance of fake review detection.

INTRODUCTION

1.1 INTRODUCTION

As the world is changing with the advancement in technology, so do the methods by which the customers are likely to buy things online. The growing trend of online shopping has made many people to read reviews before buying online products and provide their opinion for the same. Online shopping has a huge impact on the growing economy and so the companies are largely affected. Therefore, the chances of opinion spamming increases, as nowadays most of the customers rely on online reviews. The impact of such information on the organisation and the customer motivates people to promote or demote a product. Positive reviews have a huge impact on an organisations reputation and fame. Whereas, negative opinions can result into huge social and economic losses. Some organisation hires people to write positive reviews for their products or negative reviews for their competitors. Detection of these fake reviews has become an important issue in order to provide right and useful information.

Natural Language Processing is an area concerned with how computers help to understand human language, text and speech. This can be used to understand useful things. Detection of fake reviews can be done with natural language processing by using many Machine Learning Techniques. This helps to extract information and patterns from the text content of the reviews. The information is useful to compare various reviews and spot the fake ones. Users writing fake reviews choose different words or pattern in order to create impact on others. This can be used as a way to detect fake and truthful reviews. So far three categories of fake reviews have been detected,1)Untruthful opinions 2)reviews on different brands 3) non-reviews which includes various advertisements and links. The untruthful reviews are of most concern as they undermine the integrity of the online review system. It is difficult to distinguish between fake and real reviews by manually reading them therefore detection of these review spam is a challenging task.

Consumer reviews play an important in understanding the direct market scenario of the products and the company .It facilitates the idea of decision making for which we call for a Decision support Systems. Decision support systems decode and find solutions to various rapidly changing and advanced problems. The existing machine learning methods can be split into supervised and unsupervised approaches. Second, they can be split into three categories by their features: behavioural, linguistic or those using a combination of these two.

1.2 PROBLEM STATEMENT

As the use of online products and application has been increasing at a rapid rate, the competition in the market is increasing day by day. Business officials in order to fame their products and defame other competitor products may get people from other marketplace or hire them to post and give fake judgements on the products .So to safeguard the authenticity of the online products and opinions various steps and methods are needed to be applied.

Financially if we look at the situation, the opinions and reviews on various products may cost a lot to one company if people are actually relying on those reviews. If the reviews are original and authentic then the results are more obvious and fair, But if we ponder over the other side then the situation reverses. More and more fake reviews will cause a huge loss for the company which has got good and better products to sale which causes a huge financial loss to the growing company.

Fake Review Detection will help us to identify whether the opinion given to the product is a true or fake review. Various methods can be used to detect it but we our focusing on finding the results using Machine learning Techniques.

Our method focuses on the text content of the costumer reviews. People who tend to write fake reviews choose topics or words to influence online customers thus their word choice is different than others. This word choice pattern can be used for segregation of fake and truthful reviews. If these reviews are detected properly, then fake reviews may be automatically removed after detection, which may help to generate only the genuine information and serve as a boon to the companies and market especially the customers.

1.3 OBJECTIVE

To propose a method that has the potential to improve the performance and automate the task of fake review detection by identifying unique aspects of fake and truthful reviews.

The purposed method focuses on the text content of the reviews. We expect that the unnatural words used by fake review writers, to influence people, would be grouped into topics that are different from that of truthful reviews. The method focuses to model the process by means of the generative process of Latent Dirichlet Allocation(LDA) and use the resulting “topics” as differentiator between truthful and fake reviews.

1.4 METHODOLOGY

The method applied focus on the words of the review which different users use to write their reviews. Both the review are written for different purposes therefore their choice of words are also different. In order to create impact on others fake review writers writes review with a lot of information (hence their reviews are large in size).Also the chose words that will attract most of the customers to read them before buying a product. Words will include adjectives, adverbs and also words for comparison On the other hand a genuine review writer will write reviews based on its personal experience thus include more nouns, verbs and less of comparisons. This choice of words can be used to distinguish between these two kinds of reviews and help in review spam detection.

The dataset contains large number of reviews from seven different business domains. Hundred topics are extracted from theses reviews corresponding to each domain by using Latent Dirichlet Allocation. These hundred topics are then used as train models to classify individual review by calculating the score of each individual review.

We have two sub collection of reviews true type(D_T and fake type D_F)

Based on the formula:-

$$\mathbf{s} = \Sigma_{\, d\in \boldsymbol{D}T\cup \boldsymbol{D}F}\, \boldsymbol{\theta}_d$$
$$\boldsymbol{\Theta}_X = (\, \Sigma_{d\in \boldsymbol{D}X}\, \boldsymbol{\theta}_d\,)/\mathbf{s}$$

$\boldsymbol{\theta}$ is topic distribution of test review

$\boldsymbol{\Theta}_T$ is the global topic distribution of true review w.r.t LDA topics.

$\boldsymbol{\Theta}_F$ is the global topic distribution of true review w.r.t LDA topics

X={T,F} and / is element wise division.

The factors $w_T{}^i$ and $w_F{}^i$ are weights of the i-th topics in $\boldsymbol{D}_T$ and $\boldsymbol{D}_F$, respectively.

The weights reflect how widely a topic is spread across the reviews in either sub-collection(True and fake)

$$w_X{}^i = \log(|\{d\in \boldsymbol{D}_X : \theta_d{}^i > \tau\}| + 1 / |\{d\in \boldsymbol{D}_Y : \theta_d{}^i > \tau\}| + 1)$$

where $\{X,Y\}$ =$\{\, T,F\}$ or$\{\, F,T\}$ and $\tau = 1/\mathrm{K}$

K is the number of topics.

Finally, we combine the two factors for the final score:

$$\text{Score}_X(\ r) = \Sigma\theta_r^i \times (\Theta_X^i + \sigma \times w_X^i) \text{ where } X=\{T,F\}.$$

$\boldsymbol{\theta}r$ is topic distribution of test review r, which is the result of the inference in LDA. We consider that the global trend of topics would be more important and therefore wTi and wFi is multiplied by σ (0.2) in the current implementation.

The decision of whether or not r belongs to the fake review category is determined by comparing its scores for fake and truthful reviews: $scoreF\ r > scoreT\ r$

Apart from calculating the score the word choice pattern can also be used to distinguish between the reviews. All the reviews mails use five kinds of word type. These five categories are:-

- **Concrete Experience (CE):** Topics in this class mainly contain verbs.
- **Detailed Information (DI):** Topics usually composed of specific nouns.
- **General Comments (GC):** Topics contains describing abstract evaluations and adjectives.
- **Comparative Assessment (CA):** Topics in this class include comparative words like better, easier etc.
- **Reference and Recommendation (RR):** Includes words like refer, advice etc.

Many previous studies have found that true reviews mainly contain nouns and verbs therefore fall into first two categories. Last three categories belong to fake reviews. Whenever a fake review writer is writing a review he will try to compare the products to influence people to buy his product. Comparison will include words like better superior, lighter etc. Comparison can also be used to demote a particular kind of product. Fake reviews can also be used to false praise a product.

This is done so that most costumers buy the product therefore to praise the product writers will include a lot of adjectives in their reviews. To influence people to buy or not to by a product fake review writers will include words like “I advise you”, refer or recommend. Therefore fake reviews mostly contain last three categories.

For this hundred topics have been extracted from all domains and then classified into these categories

The classified topics are then train based on the probability of each word type in the topic using Support vector machines (SVM).This is then used to classy individual review for spam detection.

LITERATURE SURVEY

Nitin Jindal et al[1] identified 3 types of spam Type1- Untruthful reviews(fake reviews) Type2-Review on brands (targeting on different brands and are not product specific) Type3- Non-reviews(contains advertisements and irrelevant links). Amazon reviews were detected as duplicate and non-duplicate reviews Spam reviews of type2 and type3 are detected based on traditional classification learning as these two can be detected manually. Their research found that there are a large number of duplicate reviews on same product or same review on different products. These are written by group of users to create impact thus must contain Type1 reviews. These were detected by identifying (i) Duplicates from different userids on the same product.(ii) Duplicates from the same userid on different products (iii) Duplicates from different userids on different products by calculating similarity score of two reviews. They used logistic regression to find the probability of each review to be a spam. Three features that are content of the review, reviewer who wrote the review and the product being reviewed are used for training data for regression model.

Spammers groups influence a majority of costumers because of their size thus detection of these groups is equally important. Focusing on the issue Arjun et al[6] worked on detection of Spammers Groups instead of spotting fake reviews alone. A group of reviewers who work collaboratively to write fake reviews take control of the sentiment of a particular product influencing costumers because of their size. The purposed method focuses on detection of these groups along with fake reviews. Candidate groups are identifies using frequent itemset mining. These candidate groups are ranked based on their likelihoods for being spam called as GS Rank.GS Rank is based on Group content similarity, Group member content similarity, Group early time frame and labelling given to each group. Spamming behaviour indicator is based on both group and individual behaviour indicators. For each feature belonging to group or individual Statistical validation is done. Behavioural distribution of each group is used to identify the spammers groups. The proposed technique does not use from the traditional supervised learning approach for spam detection because of the inherent nature of problem which makes the classic supervised learning approach less effective. Experimental results show that the proposed method outperforms various detection techniques like supervised classification, regression, and learning to rank algorithms but it is a very time consuming task.

Yuging Lu et al.[7] proposed a method for the simultaneous detection of fake reviews and fakers .Factor Graph method incorporates features of reviews and reviewers which were divided into categories (i) review related (ii) reviewer related (iii) features between reviewers and reviews (iv) review group features based on the classification the local factor(probability of fakeness).Based on the these features three factors are defined namely local feature factor(find the probability of a reviewer or review to be fake) , group domain factor(using review group rating to measure the reliability of all the rating given to a particular product) and cross domain factor(relationship between reviewer and his product). All the review information is transformed into Review graph method. This Review Graph model containing all he features is used to create a Model. Model learning and inference of the model help to detect the fake reviews and reviewers in united framework. Experimental results show that method outperforms all of the other baseline methods significantly with respect to both efficiency and accuracy.

In the same year Kuldeep et al.[4] proposed a method for automatic review spam detection of fake reviews. In addition to some previous criteria like link in the reviews, all capital reviews, product and companies comparison one new criteria was introduced by taking in account the rating given to the product by the user. The sentiment of text is compared against the rating provided by the customers. The numerical value of sentiment of text is produced by using existing sentiment analysis tools like PhpInsight and Alchemy and is compared with the rating provided by the customer for consistency. PhpInsight is a Bayesian classifier that classifies the words of dictionary as positive, negative and neutral. Alchemy provides natural processing tools like tagging, topic categorisation, and language detection. In short it uses machine learning techniques to analyse the content of the review. The numerical value produced by sentiment analysis tool was compared with ratings of the review provided by the user and inconsistency between the two indicates a possibility of review spam.

Dewang et al. [3] addressed the issue by focussing on the lexical and the syntactic features. Lexical features are used to analyse each word in the review. POS is used as a lexical feature. Type-Token ratio is calculated which is the ratio of number of words of type adjective, noun, verb and adverb to the total number of words in the review. This is used to calculate lexical diversity. The purpose of this is to calculate the review complexity. Syntactic development for language learners is provided by syntactic features. Syntactic complexity is calculated by the number of clause and one main clause with some subservient clause joined together known as T-unit. Researchers used 16 new lexical and 25 new syntactic features. Features value is calculated one by one using koRpus (R package) and used as a training data for classification of further reviews.

Apart from all the reviews posted Negative reviews have a huge impact on the costumers thus it is important to detect authentic and manipulative negative reviews. Analyses of authentic and manipulative reviews by Sneshasish Banerjee et al.[3] showed review readability review genre and review writing style can be used to distinguish between the two kinds of reviews. These types of reviews are written with different purposes which can be reflected in their content. Manipulative review writers use sophisticated language for the review so that most of the costumers can follow them. Too simplistic reviews attract less people. Readability can be used to differentiate between the fake and true reviews. Authentic reviews are based on real events while the manipulative therefore are considered to be imaginative and authentic are considered to be informative out of therefore genre of text (conversational, text-oriented, informative and imaginative). Four genre of text are conversational, text-oriented Manipulative reviews contain different distribution of adjectives, nouns, verbs and adverbs than authentic reviews. Third, the writing style of both types of review writers is different. Manipulative reviews contain more affective cues (mainly negative affective cues), perceptual words and they also use more future tense to recommend to other customer.

Table1: Different Techniques for Fake Review detection

S No	Authors	Topic	Year	Website covered	Tools & Technology	Type detected
1.	Nitin Jindal & Bing liu	Opinion Spamming & Analysis	2008	Amazon	i)Natural Language Processing (Logistic Regression) ii)Data mining Techniques	i)Untruthful Opinions ii)Review on brands iii)Non reviews
2.	Arjun Mukherjee, Bing Liu, Natalic Glance	Identification fake reviews groups in customer reviews	2012	Amazon	Behavioural and relation models	Spammer Groups.
3.	Andrej ,Gregoe ,Karosak	Enhancing detection of opinion spam groups	2013	Facebook & Twitter	Quantitative Psycho-linguistic text analytics tool	Spam Groups
4.	Yuqing Lu, Zhang, Yudong Xiao, Yangguang	Detection of fake reviews and fake reviewers group simultaneously	2013	Amazon	Factor Graph Model	Fake reviews & spammers

S.No.	Authors	Topics	Year	Website Covered	Tools & Technologies	Type Detected
5.	Kuldeep Sharma, King-Ip Lin	Using Rating Consistency check for detection of fake reviews	2013	Amazon	i)Php Insight ii)Alchemy	Fake reviews
6..	Snehashish Banerjee & Alton Y..K. Chua	Analysis of authentic & manipulative reviews	2014	Amazon	Binomial Logistic regression for data analysis	Authentic & manipulative.
7.	Aakas Zhiyuli, Xun Liang, Yige Wang	Recognizing Deceptive Reviews	2015	Amazon	Sentiment analysis on attributes of products	Fake Reviews
8.	Rupesh Kumar Dewang A.K Singh	Using new set of lexical and syntactic features to identify fake reviews	2015	Amazon	Naive Bayes, Decision Tree. WEKA Toolkit for classification	Fake Reviews
9.	Kyungyup Daniel Lee, Kyungah Han, Sung-Hyon Myaeng	Capturing Word choice patterns of fake and truthful reviews	2016	Yelp	i)MALLET for topic modelling ii)LIBSVM for classification	Fake Reviews

SYSTEM DEVELOPMENT

3.1 DATASET

Dataset contains reviews from Yelp, a commercial review site. It contains reviews from seven business domains. The Reviews are already labelled as fake and true by yelp's own filtering method. Large-scale data is classified using behavioural-meta features including profile of each reviewer. There is no guarantee that this method is completely error free but accuracy is obtained by using several sets of tests.

Dataset contain positive looking reviews as their stars score is high. These include fake and truthful reviews. For the detection purpose reviews having less than 150 characters were eliminated to filter out insignificant reviews. Reviews without 5 stars were also filtered out as fake review writers rarely give low score for a positive fake review. Domain containing less than 15 reviews is not included as it is not sufficient for statistical analysis.

Table2: List of reviews in the dataset.

Category	#of reviews	#of truthful reviews	# of fake Reviews
Electronics	440	220	220
Fashion	2690	1345	1345
Hospital	200	100	100
Hotel	1100	550	550
Insurance	26	13	13
Restaurant	37980	18990	37980

Input : A text document containing reviews

Total no of text documents: 35(5 for each 7 domain

Output : Individual review identified as fake or true

3.2 ALGORITHM

1. Input the five documents of each category.
2. Extraction of 100 topics (cluster of words) using LDA implement in MALLET
3. Topic models are used as training set to calculate global topic distribution of two sub categories.
4. Probability of each topic is calculated in the review to be tested.
5. Weighted topic distribution is use as a differentiator by calculating the score of the review
6. Review identified as fake or true.

   Based on type of word used:-

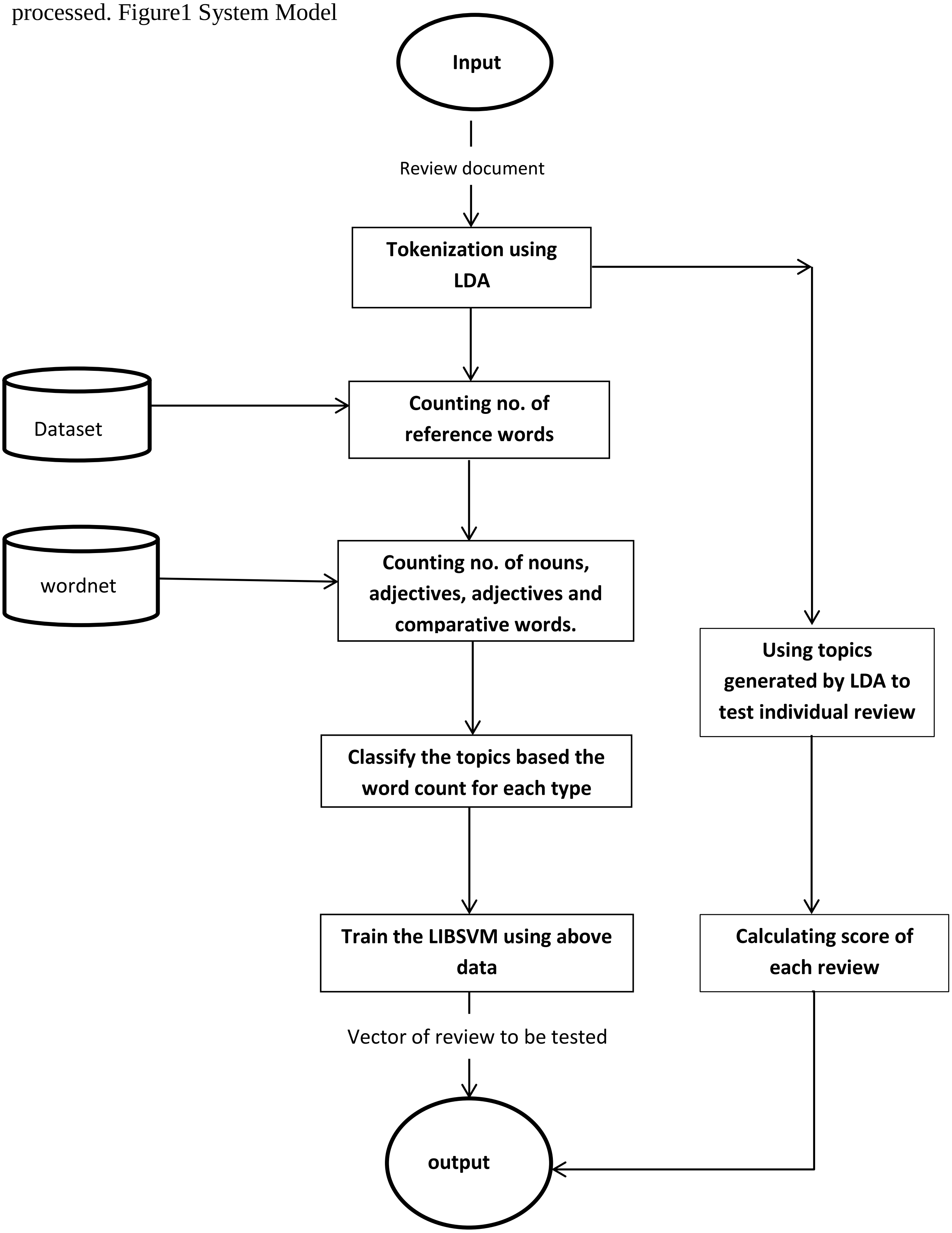
   1. Input the five documents of each category
   2. Extract hundred topics from the input documents.
   3. Counting number of nouns, adjective, adverbs and comparative words in each topic using POS tagging.
   4. Identify the category of topic based on word count of each type (categories mentioned above)
   5. Divide the topics into two classes truthful and fake.

```
  For each topic
            if CE+DI words >  GC+CA+RR words then class 1
else
        class2
```

   6. Train the SVM classifier using this data.
   7. Convert each review to be tested into vector
   8. Test the review to be fake or true.

3.3 SYSTEM MODEL

The following system model shows the flow of how the data will be transmitted and processed. Figure1 System Model


Input
Review document
Tokenization using LDA
Dataset
Counting no. of reference words
wordnet
Counting no. of nouns, adjectives, adjectives and comparative words.
Using topics generated by LDA to test individual review
Classify the topics based the word count for each type
Train the LIBSVM using above data
Calculating score of each review
Vector of review to be tested
output

3.4 LATENT DIRICHLET ALLOCATION

Topic modeling is an important part of natural language processing. It is used to analyse large scale data in an unsupervised manner. It defines the topics from which the document is created by defining the patterns among the words in the document. Latent Dieichlet Allocation (LDA) is the most popular model for topic modeling and also the simplest one. There is wide range of applications of LDA like document classification, sentiment analysis, and bioinformatics. The only observable feature the model sees in a document are the words and the hidden random variables are the topic distribution per document .LDA is a probabilistic generative model which defines the various topics in the document. In our method a topic is a collection of words which usually over together A topic can be defined as a probability distribution over a cluster of words.

WORKING OF LDA ALGORITHM

Various parameters are:-

i) N-number of words in the documents.
ii) M-number of documents.

Parameters to be defined

i) K-Number of topics
ii) Alpha- concentration parameter defining per document topic distribution.
iii) Beta-concentration parameter defining per topic word distribution

Analysing the documents hit and trial method can be used to define the number of topics.

For each topic t LDA algorithm computes two things

i) p(topic|document)- proportion of words in the document d that are currently assigned to topic t(let say value is a)

p(word|topic)-proportion of assignment to topic t over all documents that come from the word w.

LDA Model

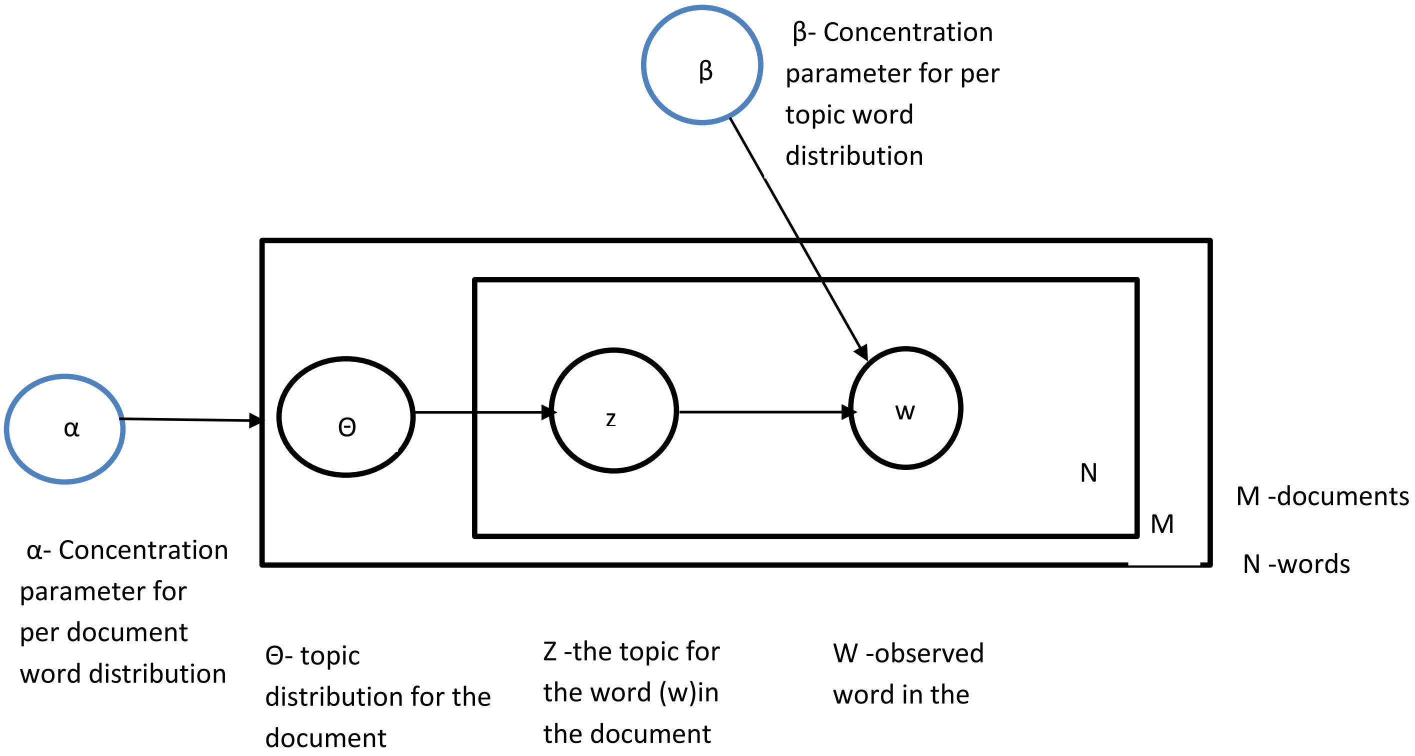

β
β- Concentration parameter for per topic word distribution
α
Θ
z
w
N
M
M -documents
N -words
α- Concentration parameter for per document word distribution
Θ- topic distribution for the document
Z -the topic for the word (w)in the document
W -observed word in the

Figure 2 : LDA Model

Steps:

1. Randomly assign a topic out of K topics to every word in the document.
   This will give topic distribution for the document and word distribution for each topic.
2. For each word w in the document d go through each word and compute p(topic|document) and p(word|topic)
3. Reassign a new topic to the word w based on the probability
   p(topic t|document d) and p(word w|topic t)
   it is done based on the assumption that every assignment of the words to the topic is correct except for the current word w.
4. Repeat step 3 several times to get accurate results.

3.5 POS TAGGING

Part-Of-Speech tagging s the process of assigning tags to each word in a sentence. Tagger is software that reads the text and automatically assign to tags to each word based on the definition of word and its relationship with other words. Various type of information is used by the taggers to assign tags which include dictionaries, rules and lexicons. There are various application of POS tagging like indexing of text, speech processing. POS tagging is mainly use to identify verbs, nouns, adjectives and adverbs in the text.

A POS tagger uses the following steps for POS tagging:

1. Tokenization of text-The text is first divided into tokens for assigning tags by analysis. These tokens are mainly words but also contain punctuations.
2. Ambiguity detection-Use lexicon for unknown words as it provides list of word forms. Also there may be ambiguity as some words can be used as nouns as well as verbs.
3. Ambiguity resolution- It is important to resolve ambiguity to have appropriate tags. It is based on the information of the word and the tag sequence in the particular sentence.

Tagset

Table 3: List of Tags

The most common tags used are:-

Tags used	Description
NN	Singular noun
NNS	Plural noun
NNP	Singular Proper noun
NNPS	Plural Proper noun
VB	Verb Base Form
VBD	Verb past tense
VBG	Verb present participle
VBN	Verb past participle
VBP	Verb non-3rd person singular present
VBZ	Verb 3rd person singular present
RB	Adverb
RBR	Adverb comparative
RBS	Adverb superlative
JJ	Adjective
JJS	Adjective comparative
JJS	Adjective superlative
MD	Modal
FW	Foreign word
RP	Particle

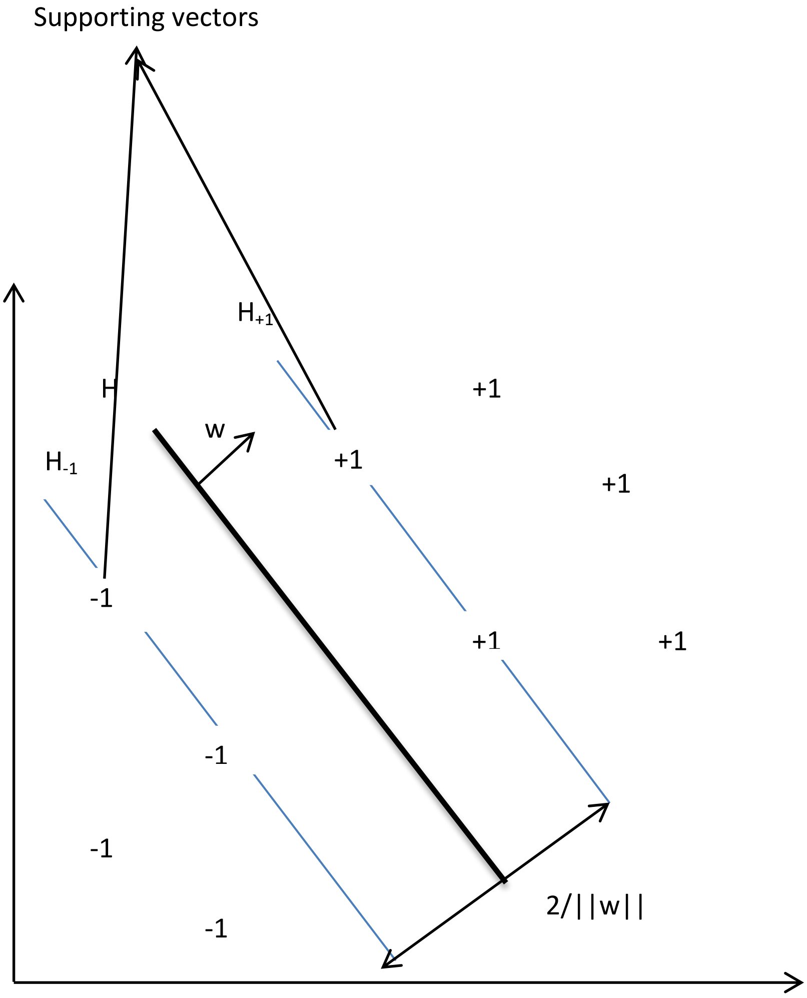
3.6 Support Vector Machine

Support vector machine is a supervised algorithm, uses train data to test the incoming data for classification and regression to test the incoming data. Supervised Vector classification is based on deciding the hyper plane that will classify the training vectors. SVM can also be used to classify data into multiple classes.

Let say we have to train the data by classifying the vectors into two classes (+1 & -1) .There are number of hyperplanes that can be defined to classify the two sets. We have to identify the optimum hyperplane that will separate the two kinds of vectors. The optimum hyperplane will be that define the maximum margin between the two sets of vectors (i.e. the supporting vectors).

Supporting vectors are the one are nearest to the hyperplane separating the two classes. Between this margin there are no training vectors. The testing data is classified based on these training vectors hence this is called supervised method.

Figure 3: Optimum Hyperplane For SVM



Hyperplane H (wx_i+c) is used to classify the test vectors with maximum width and satisfy the following inequality of each training vector in the sets.

$$wx_i+c > +1 \text{ if } y_i = +1$$

$$wx_i+c < -1 \text{ if } y_i = -1$$

can be written as

$$y_i(wx_i+c) >= +1$$

class of each training data is

If $(wx_i+c>0)$ then +1

If $(wx_i+c<0)$then -1

w is the vector from origin perpendicular to the hyperplane H.

The distance of hyperplane H from the origin is $|c|/||w||$.Hyperplane of the class +1(H_{+1}) satisfy the equation $wx_i+c=+1$ therefore its distance from origin is $|+1-c|/||w||$.

Similarly hyperplane of class -1(H_{-1}) satisfy the equation $wx_i+c= -1$.its distance from the origin is $|-1-c|/||w||$. The distance of the two hyperplanes H_{+1} and H_{-1} is $2/||w||$.

In order to find the optimum hyperplane we need to maximize $2/||w||$ or minimising $||w||^2/2$.

SVM determines this hyperplane for accurate classification of the data.

3.7 Implementation

Extraction of topics using LDA implemented in MALLET.

There are various implementation of Latent Dirichlet Allocation(LDA) like in mat lab, MALLET(Java based). For our project we have used MALLET. It is a Java-based package for natural language processing. Mallet can be used in various applications related to text analysis.

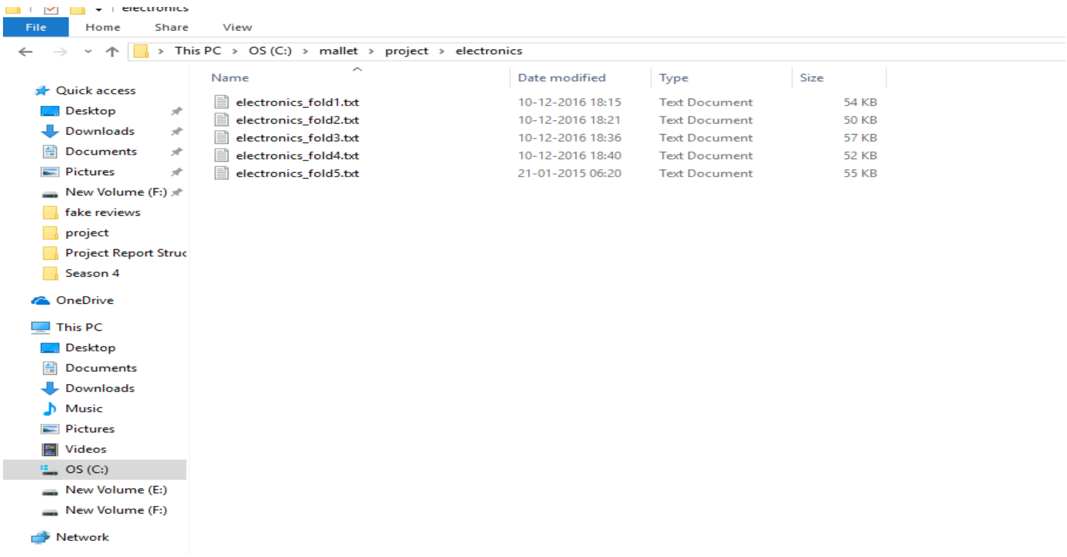
The Topic models are used to analyze large amount of unlabelled text. LDA provide topic models. MALLET provides efficient implementation of LDA which is based on sampling.

The following process include reviews of electronics category tokenize into 100 topics. Each topic includes 10-15 words.

Step1:-

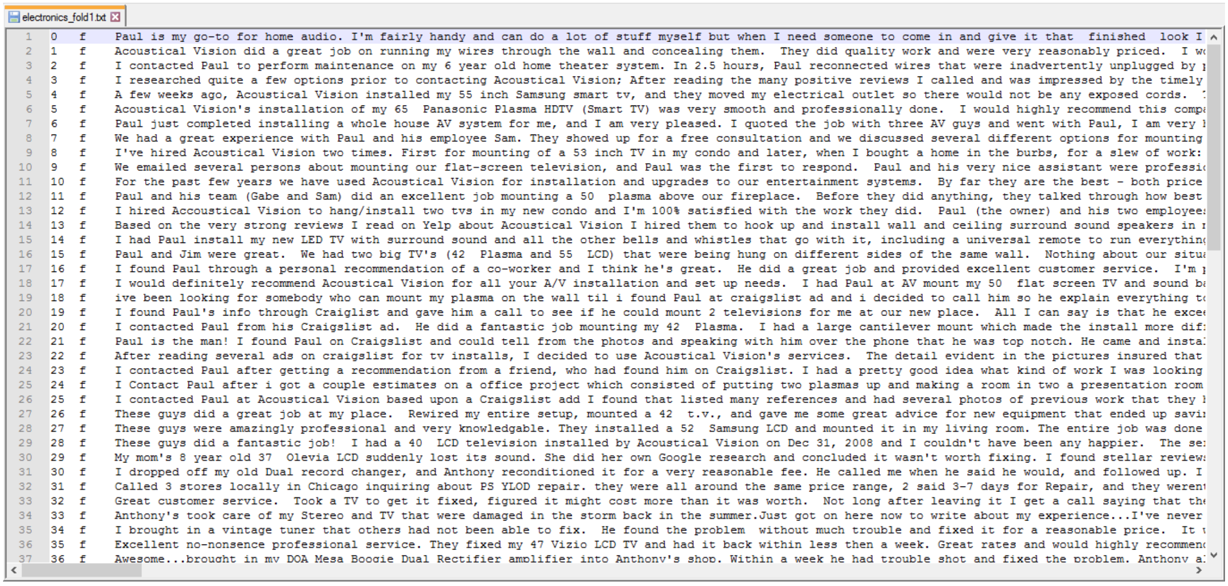
Data collection

Each domain contains 5 text files containing reviews.



Name	Date modified	Type	Size
electronics_fold1.txt	10-12-2016 18:15	Text Document	54 KB
electronics_fold2.txt	10-12-2016 18:21	Text Document	50 KB
electronics_fold3.txt	10-12-2016 18:36	Text Document	57 KB
electronics_fold4.txt	10-12-2016 18:40	Text Document	52 KB
electronics_fold5.txt	21-01-2015 06:20	Text Document	55 KB

Data is in the form of text file containing reviews.

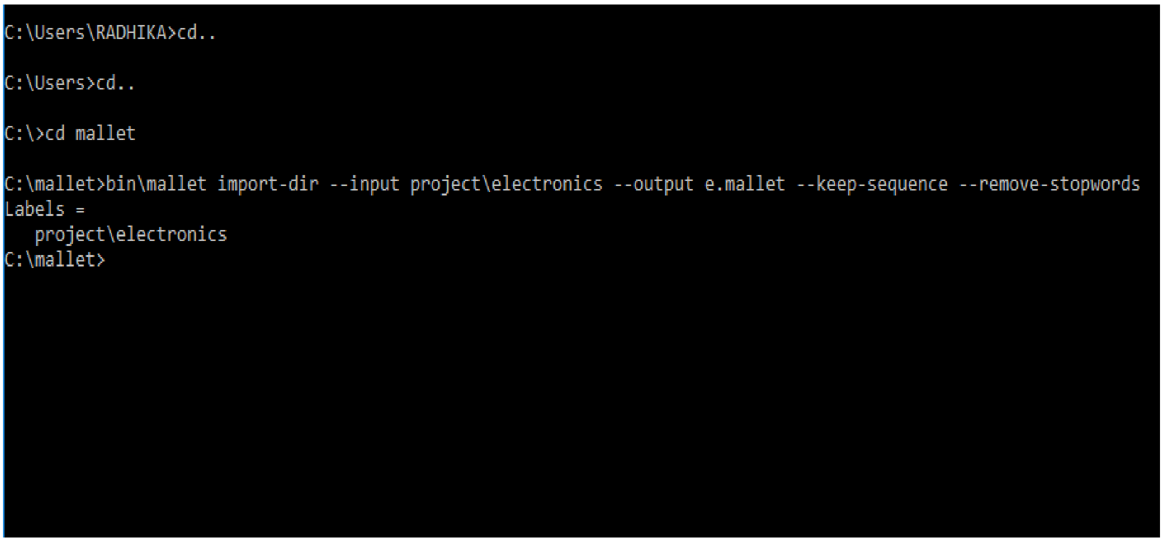


electronics_fold1.txt

```
0	f	Paul is my go-to for home audio. I'm fairly handy and can do a lot of stuff myself but when I need someone to come in and give it that  finished  look I
1	f	Acoustical Vision did a great job on running my wires through the wall and concealing them.  They did quality work and were very reasonably priced.  I w
2	f	I contacted Paul to perform maintenance on my 6 year old home theater system. In 2.5 hours, Paul reconnected wires that were inadvertently unplugged by
3	f	I researched quite a few options prior to contacting Acoustical Vision; After reading the many positive reviews I called and was impressed by the timely
4	f	A few weeks ago, Acoustical Vision installed my 55 inch Samsung smart tv, and they moved my electrical outlet so there would not be any exposed cords.
5	f	Acoustical Vision's installation of my 65  Panasonic Plasma HDTV (Smart TV) was very smooth and professionally done.  I would highly recommend this comp
6	f	Paul just completed installing a whole house AV system for me, and I am very pleased. I quoted the job with three AV guys and went with Paul, I am very
7	f	We had a great experience with Paul and his employee Sam. They showed up for a free consultation and we discussed several different options for mounting
8	f	I've hired Acoustical Vision two times. First for mounting of a 53 inch TV in my condo and later, when I bought a home in the burbs, for a slew of work:
9	f	We emailed several persons about mounting our flat-screen television, and Paul was the first to respond.  Paul and his very nice assistant were professi
10	f	For the past few years we have used Acoustical Vision for installation and upgrades to our entertainment systems.  By far they are the best - both price
11	f	Paul and his team (Gabe and Sam) did an excellent job mounting a 50  plasma above our fireplace.  Before they did anything, they talked through how best
12	f	I hired Accoustical Vision to hang/install two tvs in my new condo and I'm 100% satisfied with the work they did.  Paul (the owner) and his two employee
13	f	Based on the very strong reviews I read on Yelp about Acoustical Vision I hired them to hook up and install wall and ceiling surround sound speakers in
14	f	I had Paul install my new LED TV with surround sound and all the other bells and whistles that go with it, including a universal remote to run everythin
15	f	Paul and Jim were great.  We had two big TV's (42  Plasma and 55  LCD) that were being hung on different sides of the same wall.  Nothing about our situ
16	f	I found Paul through a personal recommendation of a co-worker and I think he's great.  He did a great job and provided excellent customer service.  I'm
17	f	I would definitely recommend Acoustical Vision for all your A/V installation and set up needs.  I had Paul at AV mount my 50  flat screen TV and sound b
18	f	ive been looking for somebody who can mount my plasma on the wall til i found Paul at craigslist ad and i decided to call him so he explain everything t
19	f	I found Paul's info through Craiglist and gave him a call to see if he could mount 2 televisions for me at our new place.  All I can say is that he exce
20	f	I contacted Paul from his Craigslist ad.  He did a fantastic job mounting my 42  Plasma.  I had a large cantilever mount which made the install more dif
21	f	Paul is the man! I found Paul on Craigslist and could tell from the photos and speaking with him over the phone that he was top notch. He came and insta
22	f	After reading several ads on craigslist for tv installs, I decided to use Acoustical Vision's services.  The detail evident in the pictures insured that
23	f	I contacted Paul after getting a recommendation from a friend, who had found him on Craigslist. I had a pretty good idea what kind of work I was looking
24	f	I Contact Paul after i got a couple estimates on a office project which consisted of putting two plasmas up and making a room in two a presentation room
25	f	I contacted Paul at Acoustical Vision based upon a Craigslist add I found that listed many references and had several photos of previous work that they
26	f	These guys did a great job at my place.  Rewired my entire setup, mounted a 42  t.v., and gave me some great advice for new equipment that ended up savi
27	f	These guys were amazingly professional and very knowledgable. They installed a 52  Samsung LCD and mounted it in my living room. The entire job was done
28	f	These guys did a fantastic job!  I had a 40  LCD television installed by Acoustical Vision on Dec 31, 2008 and I couldn't have been any happier.  The se
29	f	My mom's 8 year old 37  Olevia LCD suddenly lost its sound. She did her own Google research and concluded it wasn't worth fixing. I found stellar review
30	f	I dropped off my old Dual record changer, and Anthony reconditioned it for a very reasonable fee. He called me when he said he would, and followed up. I
31	f	Called 3 stores locally in Chicago inquiring about PS YLOD repair. they were all around the same price range, 2 said 3-7 days for Repair, and they werent
32	f	Great customer service.  Took a TV to get it fixed, figured it might cost more than it was worth.  Not long after leaving it I get a call saying that the
33	f	Anthony's took care of my Stereo and TV that were damaged in the storm back in the summer.Just got on here now to write about my experience...I've never
34	f	I brought in a vintage tuner that others had not been able to fix.  He found the problem  without much trouble and fixed it for a reasonable price.  It w
35	f	Excellent no-nonsence professional service. They fixed my 47 Vizio LCD TV and had it back within less then a week. Great rates and would highly recommen
36	f	Awesome...brought in my DOA Mesa Boogie Dual Rectifier amplifier into Anthony's shop. Within a week he had trouble shot and fixed the problem. Anthony a
```

Step 2:-

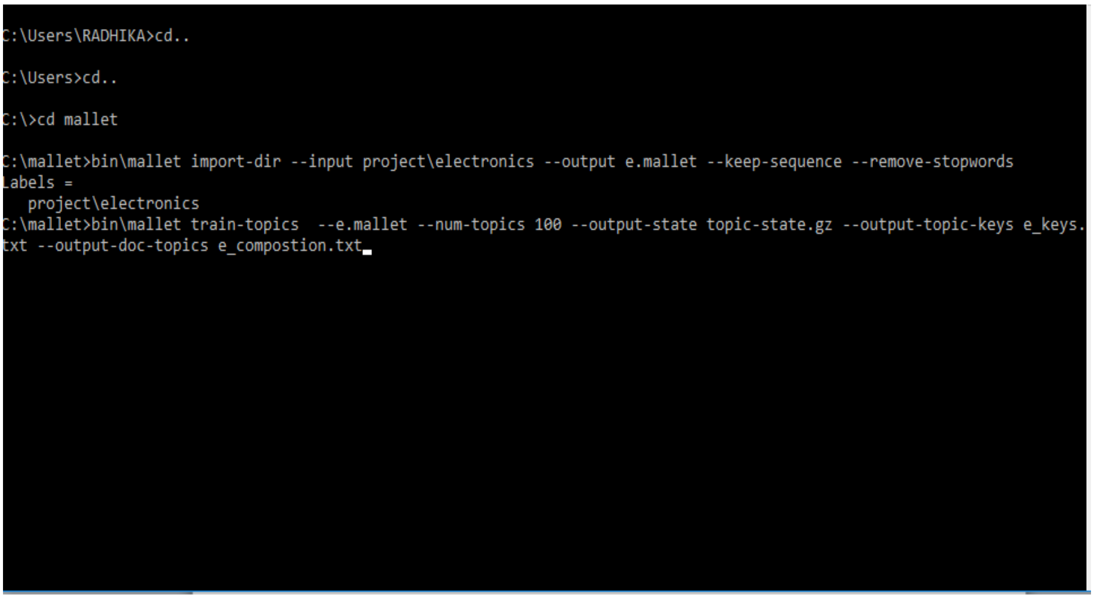
Importing the folder and converting the input file into Mallet format. Stopwords (including if,the,and) are also eliminated because they occur in large number and obstruct analysis of the text.



```
C:\Users\RADHIKA>cd..

C:\Users>cd..

C:\>cd mallet

C:\mallet>bin\mallet import-dir --input project\electronics --output e.mallet --keep-sequence --remove-stopwords
Labels =
   project\electronics
C:\mallet>
```

Step 3:-

Extracting 100 topics from the folder

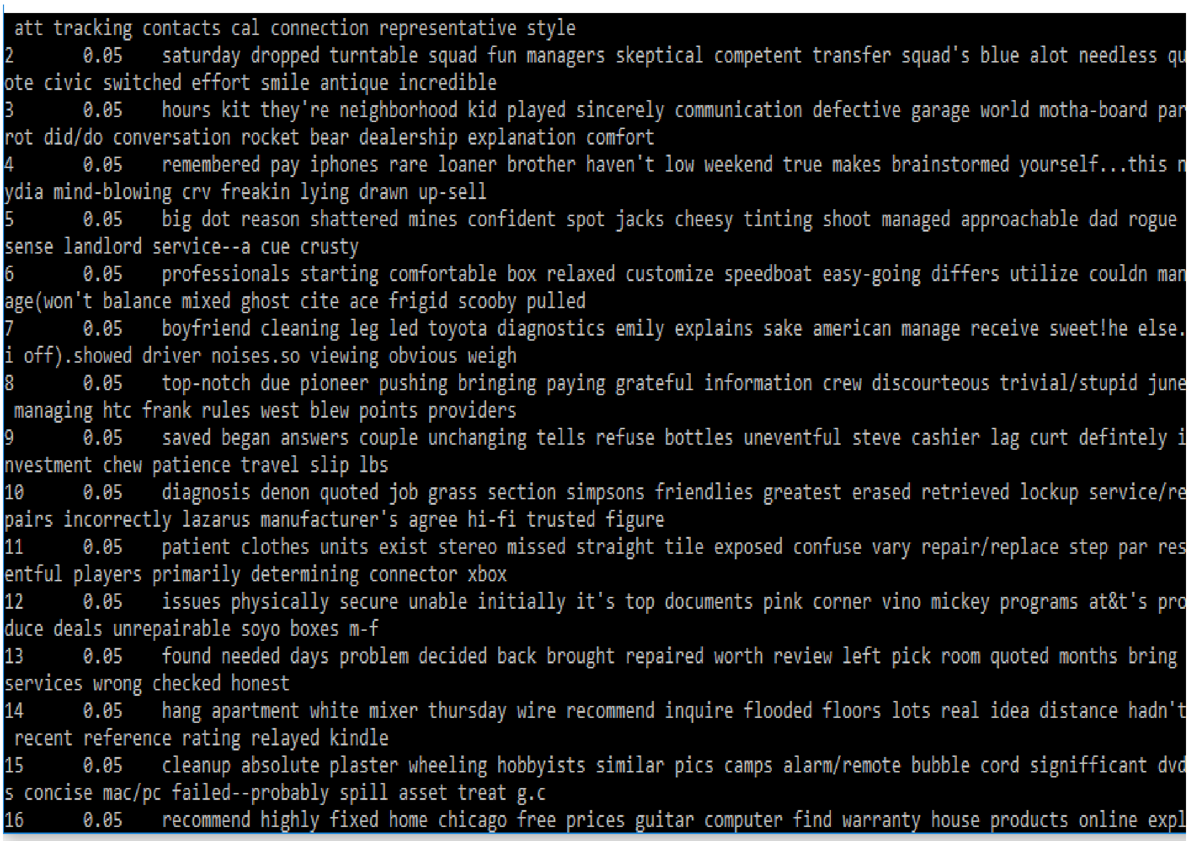


```
C:\Users\RADHIKA>cd..

C:\Users>cd..

C:\>cd mallet

C:\mallet>bin\mallet import-dir --input project\electronics --output e.mallet --keep-sequence --remove-stopwords
Labels =
   project\electronics
C:\mallet>bin\mallet train-topics  --e.mallet --num-topics 100 --output-state topic-state.gz --output-topic-keys e_keys.
txt --output-doc-topics e_compostion.txt
```

It will give the following output:-

- Output every word in every document and the topic it belongs to into a compressed file.
- outputs a text document containing topics which in turn contains cluster of words(e_keys.txt)
- outputs a text file indicating the percentage of each topic within each text file imported (e_composition.txt).

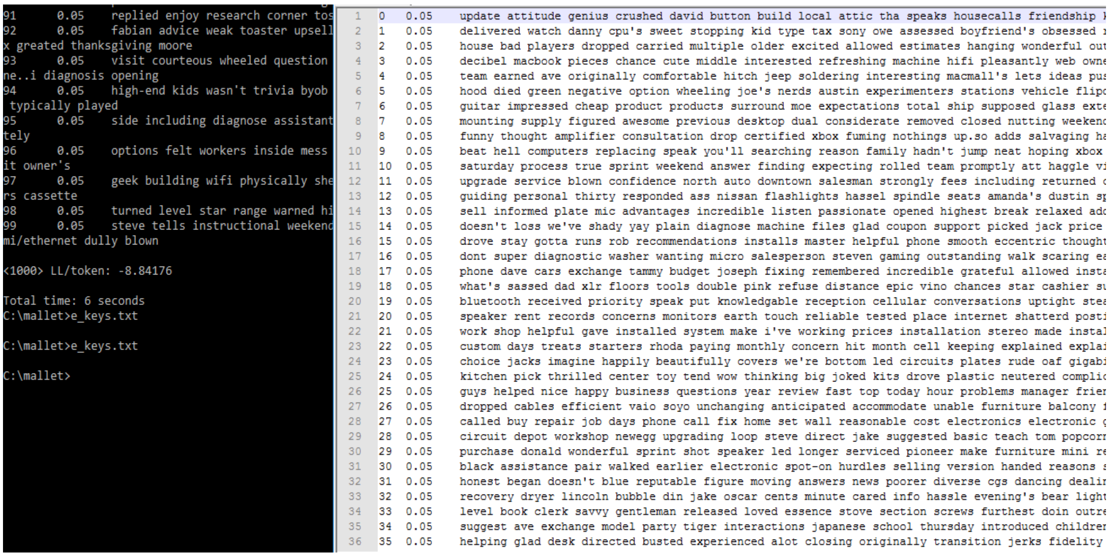
Topics are extracted in the following manner



```
 att tracking contacts cal connection representative style
2       0.05    saturday dropped turntable squad fun managers skeptical competent transfer squad's blue alot needless qu
ote civic switched effort smile antique incredible
3       0.05    hours kit they're neighborhood kid played sincerely communication defective garage world motha-board par
rot did/do conversation rocket bear dealership explanation comfort
4       0.05    remembered pay iphones rare loaner brother haven't low weekend true makes brainstormed yourself...this n
ydia mind-blowing crv freakin lying drawn up-sell
5       0.05    big dot reason shattered mines confident spot jacks cheesy tinting shoot managed approachable dad rogue
sense landlord service--a cue crusty
6       0.05    professionals starting comfortable box relaxed customize speedboat easy-going differs utilize couldn man
age(won't balance mixed ghost cite ace frigid scooby pulled
7       0.05    boyfriend cleaning leg led toyota diagnostics emily explains sake american manage receive sweet!he else.
i off).showed driver noises.so viewing obvious weigh
8       0.05    top-notch due pioneer pushing bringing paying grateful information crew discourteous trivial/stupid june
 managing htc frank rules west blew points providers
9       0.05    saved began answers couple unchanging tells refuse bottles uneventful steve cashier lag curt defintely i
nvestment chew patience travel slip lbs
10      0.05    diagnosis denon quoted job grass section simpsons friendlies greatest erased retrieved lockup service/re
pairs incorrectly lazarus manufacturer's agree hi-fi trusted figure
11      0.05    patient clothes units exist stereo missed straight tile exposed confuse vary repair/replace step par res
entful players primarily determining connector xbox
12      0.05    issues physically secure unable initially it's top documents pink corner vino mickey programs at&t's pro
duce deals unrepairable soyo boxes m-f
13      0.05    found needed days problem decided back brought repaired worth review left pick room quoted months bring
services wrong checked honest
14      0.05    hang apartment white mixer thursday wire recommend inquire flooded floors lots real idea distance hadn't
 recent reference rating relayed kindle
15      0.05    cleanup absolute plaster wheeling hobbyists similar pics camps alarm/remote bubble cord significant dvd
s concise mac/pc failed--probably spill asset treat g.c
16      0.05    recommend highly fixed home chicago free prices guitar computer find warranty house products online expl
```

Step4:-

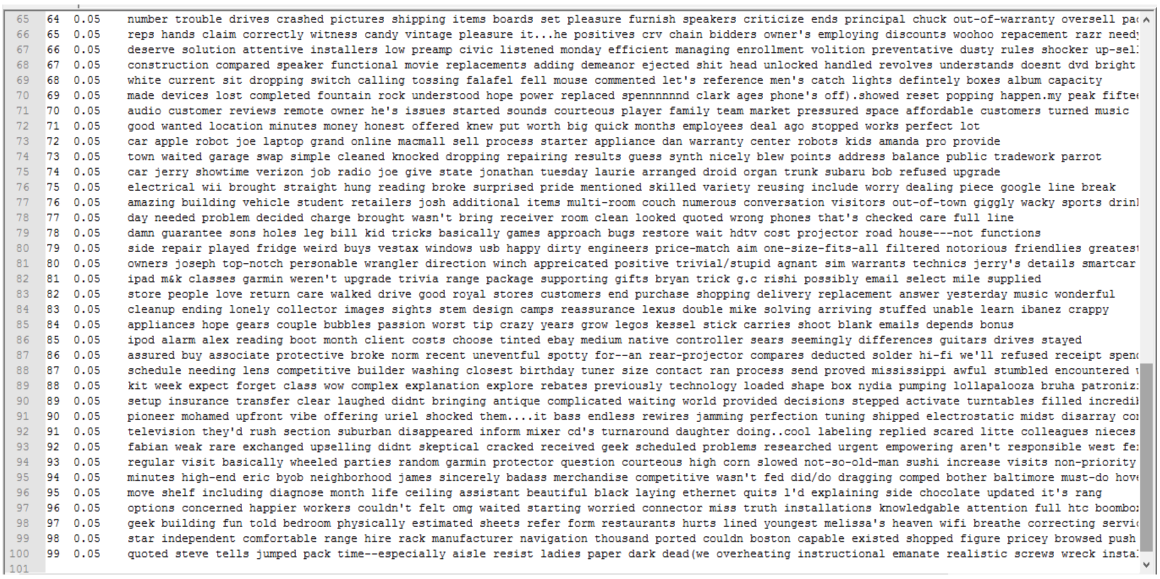
Open e_keys it contains series of paragraph .Each paragraph is a topic. There will be 100 (0 to 99) paragraphs in the text document. Paragraph 0 is topic 1,paragraph 1 is topic 2 and so on.



```
91      0.05    replied enjoy research corner tos
92      0.05    fabian advice weak toaster upsell
x greated thanksgiving moore
93      0.05    visit courteous wheeled question
ne..i diagnosis opening
94      0.05    high-end kids wasn't trivia byob
 typically played
95      0.05    side including diagnose assistant
tely
96      0.05    options felt workers inside mess
it owner's
97      0.05    geek building wifi physically she
rs cassette
98      0.05    turned level star range warned hi
99      0.05    steve tells instructional weekend
mi/ethernet dully blown

<1000> LL/token: -8.84176

Total time: 6 seconds
C:\mallet>e_keys.txt

C:\mallet>e_keys.txt

C:\mallet>
```

```
0   0.05    update attitude genius crushed david button build local attic tha speaks housecalls friendship
1   0.05    delivered watch danny cpu's sweet stopping kid type tax sony owe assessed boyfriend's obsessed
2   0.05    house bad players dropped carried multiple older excited allowed estimates hanging wonderful out
3   0.05    decibel macbook pieces chance cute middle interested refreshing machine hifi pleasantly web owne
4   0.05    team earned ave originally comfortable hitch jeep soldering interesting macmall's lets ideas pus
5   0.05    hood died green negative option wheeling joe's nerds austin experimenters stations vehicle flip
6   0.05    guitar impressed cheap product products surround moe expectations total ship supposed glass exte
7   0.05    mounting supply figured awesome previous desktop dual considerate removed closed nutting weekenc
8   0.05    funny thought amplifier consultation drop certified xbox fuming nothings up.so adds salvaging ha
9   0.05    beat hell computers replacing speak you'll searching reason family hadn't jump neat hoping xbox
10  0.05    saturday process true sprint weekend answer finding expecting rolled team promptly att haggle vi
11  0.05    upgrade service blown confidence north auto downtown salesman strongly fees including returned
12  0.05    guiding personal thirty responded ass nissan flashlights hassel spindle seats amanda's dustin sp
13  0.05    sell informed plate mic advantages incredible listen passionate opened highest break relaxed adc
14  0.05    doesn't loss we've shady yay plain diagnose machine files glad coupon support picked jack price
15  0.05    drove stay gotta runs rob recommendations installs master helpful phone smooth eccentric thought
16  0.05    dont super diagnostic washer wanting micro salesperson steven gaming outstanding walk scaring ea
17  0.05    phone dave cars exchange tammy budget joseph fixing remembered incredible grateful allowed insta
18  0.05    what's sassed dad xlr floors tools double pink refuse distance epic vino chances star cashier su
19  0.05    bluetooth received priority speak put knowledgable reception cellular conversations uptight stea
20  0.05    speaker rent records concerns monitors earth touch reliable tested place internet shatterd posti
21  0.05    work shop helpful gave installed system make i've working prices installation stereo made instal
22  0.05    custom days treats starters rhoda paying monthly concern hit month cell keeping explained explai
23  0.05    choice jacks imagine happily beautifully covers we're bottom led circuits plates rude oaf gigabi
24  0.05    kitchen pick thrilled center toy tend wow thinking big joked kits drove plastic neutered complic
25  0.05    guys helped nice happy business questions year review fast top today hour problems manager frien
26  0.05    dropped cables efficient vaio soyo unchanging anticipated accommodate unable furniture balcony
27  0.05    called buy repair job days phone call fix home set wall reasonable cost electronics electronic
28  0.05    circuit depot workshop newegg upgrading loop steve direct jake suggested basic teach tom popcorn
29  0.05    purchase donald wonderful sprint shot speaker led longer serviced pioneer make furniture mini re
30  0.05    black assistance pair walked earlier electronic spot-on hurdles selling version handed reasons
31  0.05    honest began doesn't blue reputable figure moving answers news poorer diverse cgs dancing dealir
32  0.05    recovery dryer lincoln bubble din jake oscar cents minute cared info hassle evening's bear light
33  0.05    level book clerk savvy gentleman released loved essence stove section screws furthest doin outre
34  0.05    suggest ave exchange model party tiger interactions japanese school thursday introduced childrer
35  0.05    helping glad desk directed busted experienced alot closing originally transition jerks fidelity
```

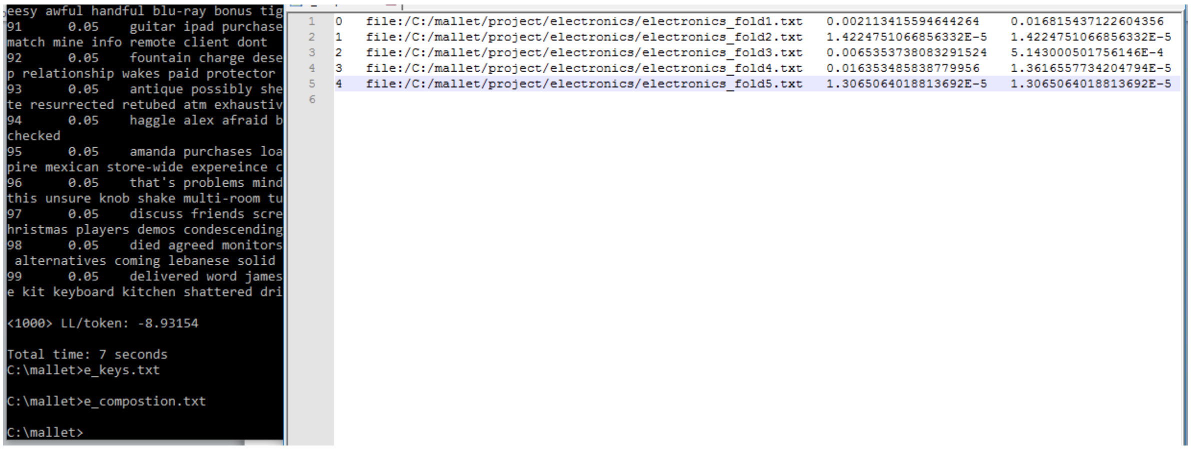
e_keys.txt



```
64  0.05    number trouble drives crashed pictures shipping items boards set pleasure furnish speakers criticize ends principal chuck out-of-warranty oversell pa
65  0.05    reps hands claim correctly witness candy vintage pleasure it...he positives crv chain bidders owner's employing discounts woohoo repacement razr need
66  0.05    deserve solution attentive installers low preamp civic listened monday efficient managing enrollment volition preventative dusty rules shocker up-sel
67  0.05    construction compared speaker functional movie replacements adding demeanor ejected shit head unlocked handled revolves understands doesnt dvd bright
68  0.05    white current sit dropping switch calling tossing falafel fell mouse commented let's reference men's catch lights defintely boxes album capacity
69  0.05    made devices lost completed fountain rock understood hope power replaced spennnnnnd clark ages phone's off).showed reset popping happen.my peak fifte
70  0.05    audio customer reviews remote owner he's issues started sounds courteous player family team market pressured space affordable customers turned music
71  0.05    good wanted location minutes money honest offered knew put worth big quick months employees deal ago stopped works perfect lot
72  0.05    car apple robot joe laptop grand online macmall sell process starter appliance dan warranty center robots kids amanda pro provide
73  0.05    town waited garage swap simple cleaned knocked dropping repairing results guess synth nicely blew points address balance public tradework parrot
74  0.05    car jerry showtime verizon job radio joe give state jonathan tuesday laurie arranged droid organ trunk subaru bob refused upgrade
75  0.05    electrical wii brought straight hung reading broke surprised pride mentioned skilled variety reusing include worry dealing piece google line break
76  0.05    amazing building vehicle student retailers josh additional items multi-room couch numerous conversation visitors out-of-town giggly wacky sports drin
77  0.05    day needed problem decided charge brought wasn't bring receiver room clean looked quoted wrong phones that's checked care full line
78  0.05    damn guarantee sons holes leg bill kid tricks basically games approach bugs restore wait hdtv cost projector road house---not functions
79  0.05    side repair played fridge weird buys vestax windows usb happy dirty engineers price-match aim one-size-fits-all filtered notorious friendlies greatest
80  0.05    owners joseph top-notch personable wrangler direction winch appreicated positive trivial/stupid agnant sim warrants technics jerry's details smartcar
81  0.05    ipad m&k classes garmin weren't upgrade trivia range package supporting gifts bryan trick g.c rishi possibly email select mile supplied
82  0.05    store people love return care walked drive good royal stores customers end purchase shopping delivery replacement answer yesterday music wonderful
83  0.05    cleanup ending lonely collector images sights stem design camps reassurance lexus double mike solving arriving stuffed unable learn ibanez crappy
84  0.05    appliances hope gears couple bubbles passion worst tip crazy years grow legos kessel stick carries shoot blank emails depends bonus
85  0.05    ipod alarm alex reading boot month client costs choose tinted ebay medium native controller sears seemingly differences guitars drives stayed
86  0.05    assured buy associate protective broke norm recent uneventful spotty for--an rear-projector compares deducted solder hi-fi we'll refused receipt spen
87  0.05    schedule needing lens competitive builder washing closest birthday tuner size contact ran process send proved mississippi awful stumbled encountered
88  0.05    kit week expect forget class wow complex explanation explore rebates previously technology loaded shape box nydia pumping lollapalooza bruha patroniz
89  0.05    setup insurance transfer clear laughed didnt bringing antique complicated waiting world provided decisions stepped activate turntables filled incredi
90  0.05    pioneer mohamed upfront vibe offering uriel shocked them....it bass endless rewires jamming perfection tuning shipped electrostatic midst disarray co
91  0.05    television they'd rush section suburban disappeared inform mixer cd's turnaround daughter doing..cool labeling replied scared litte colleagues nieces
92  0.05    fabian weak rare exchanged upselling didnt skeptical cracked received geek scheduled problems researched urgent empowering aren't responsible west fe
93  0.05    regular visit basically wheeled parties random garmin protector question courteous high corn slowed not-so-old-man sushi increase visits non-priority
94  0.05    minutes high-end eric byob neighborhood james sincerely badass merchandise competitive wasn't fed did/do dragging comped bother baltimore must-do hov
95  0.05    move shelf including diagnose month life ceiling assistant beautiful black laying ethernet quits l'd explaining side chocolate updated it's rang
96  0.05    options concerned happier workers couldn't felt omg waited starting worried connector miss truth installations knowledgable attention full htc boombo
97  0.05    geek building fun told bedroom physically estimated sheets refer form restaurants hurts lined youngest melissa's heaven wifi breathe correcting servi
98  0.05    star independent comfortable range hire rack manufacturer navigation thousand ported couldn boston capable existed shopped figure pricey browsed push
99  0.05    quoted steve tells jumped pack time--especially aisle resist ladies paper dark dead(we overheating instructional emanate realistic screws wreck insta
```

Step4:- open e_compostion.txt

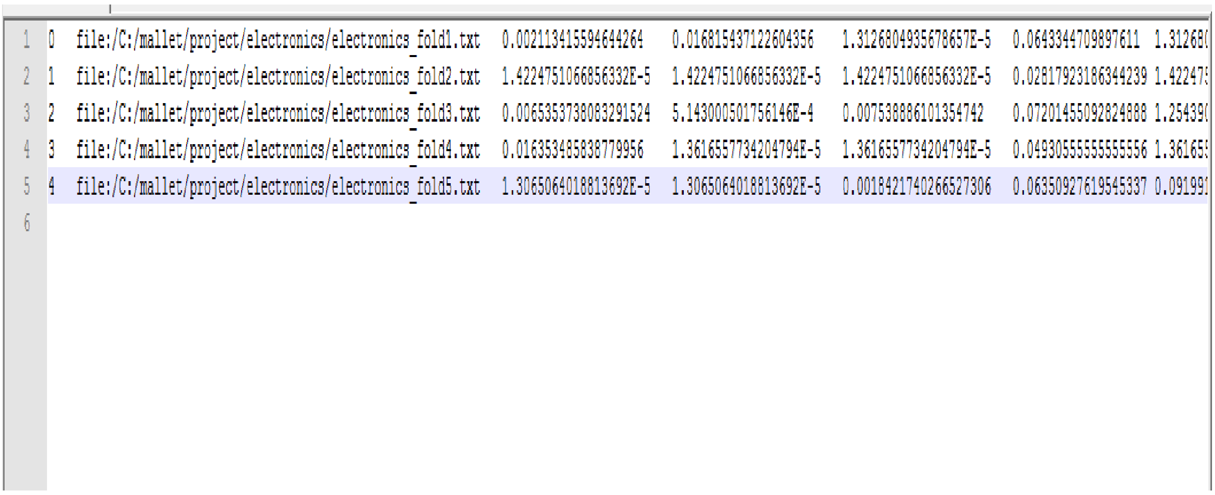
This provides the percentage of each topic in different documents.



```
eesy awful handful blu-ray bonus tig
91      0.05    guitar ipad purchase
match mine info remote client dont
92      0.05    fountain charge dese
p relationship wakes paid protector
93      0.05    antique possibly she
te resurrected retubed atm exhaustiv
94      0.05    haggle alex afraid b
checked
95      0.05    amanda purchases loa
pire mexican store-wide expereince c
96      0.05    that's problems mind
this unsure knob shake multi-room tu
97      0.05    discuss friends scre
hristmas players demos condescending
98      0.05    died agreed monitors
 alternatives coming lebanese solid
99      0.05    delivered word james
e kit keyboard kitchen shattered dri

<1000> LL/token: -8.93154

Total time: 7 seconds
C:\mallet>e_keys.txt

C:\mallet>e_compostion.txt

C:\mallet>
```

```
0   file:/C:/mallet/project/electronics/electronics_fold1.txt   0.002113415594644264    0.016815437122604356
1   file:/C:/mallet/project/electronics/electronics_fold2.txt   1.4224751066856332E-5   1.4224751066856332E-5
2   file:/C:/mallet/project/electronics/electronics_fold3.txt   0.006535373808329152 4   5.143000501756146E-4
3   file:/C:/mallet/project/electronics/electronics_fold4.txt   0.016353485838779956    1.3616557734204794E-5
4   file:/C:/mallet/project/electronics/electronics_fold5.txt   1.3065064018813692E-5   1.3065064018813692E-5
```

e_compostion.txt

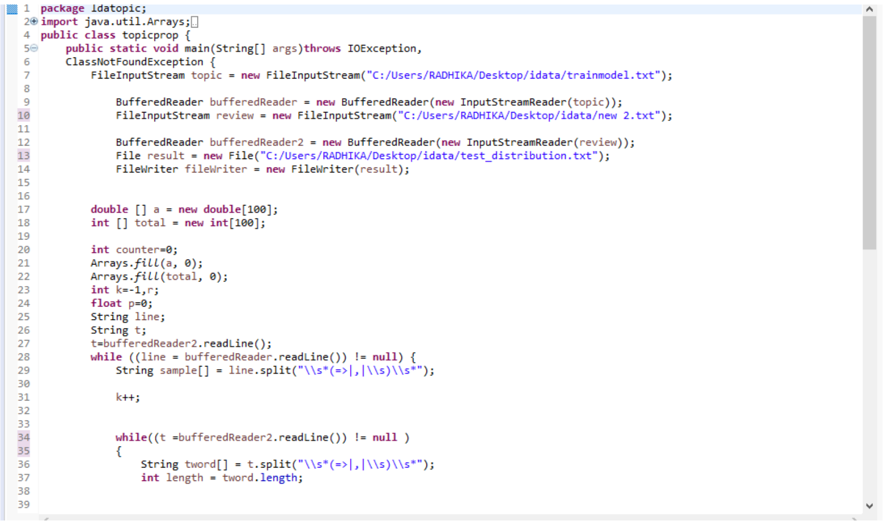


```
0   file:/C:/mallet/project/electronics/electronics_fold1.txt   0.002113415594644264    0.016815437122604356   1.3126804935678657E-5   0.0643344709897611  1.31268
1   file:/C:/mallet/project/electronics/electronics_fold2.txt   1.4224751066856332E-5   1.4224751066856332E-5  1.4224751066856332E-5   0.02817923186344239 1.42247
2   file:/C:/mallet/project/electronics/electronics_fold3.txt   0.0065353738083291524   5.143000501756146E-4   0.007538886101354742    0.07201455092824888 1.25439
3   file:/C:/mallet/project/electronics/electronics_fold4.txt   0.016353485838779956    1.3616557734204794E-5  1.3616557734204794E-5   0.04930555555555556 1.36165
4   file:/C:/mallet/project/electronics/electronics_fold5.txt   1.3065064018813692E-5   1.3065064018813692E-5  0.0018421740266527306   0.06350927619545337 0.09199
```

Calculating global to topic distribution for sub collections(True and fake reviews)

Input is given as the LDA topic models(100) and file containing true reviews.

For all the reviews the weight of each topic is calculated and stored in an array.



```java
package ldatopic;
import java.util.Arrays;
public class topicprop {
    public static void main(String[] args)throws IOException,
    ClassNotFoundException {
        FileInputStream topic = new FileInputStream("C:/Users/RADHIKA/Desktop/idata/trainmodel.txt");

            BufferedReader bufferedReader = new BufferedReader(new InputStreamReader(topic));
            FileInputStream review = new FileInputStream("C:/Users/RADHIKA/Desktop/idata/new 2.txt");

            BufferedReader bufferedReader2 = new BufferedReader(new InputStreamReader(review));
            File result = new File("C:/Users/RADHIKA/Desktop/idata/test_distribution.txt");
            FileWriter fileWriter = new FileWriter(result);


        double [] a = new double[100];
        int [] total = new int[100];

        int counter=0;
        Arrays.fill(a, 0);
        Arrays.fill(total, 0);
        int k=-1,r;
        float p=0;
        String line;
        String t;
        t=bufferedReader2.readLine();
        while ((line = bufferedReader.readLine()) != null) {
            String sample[] = line.split("\\s*(=>|,|\\s)\\s*");

            k++;


            while((t =bufferedReader2.readLine()) != null )
            {
                String tword[] = t.split("\\s*(=>|,|\\s)\\s*");
                int length = tword.length;


```

Similarly we get an array for fake reviews.

Global topic distribution for each review is calculated as

$$\mathbf{s} = \Sigma_{d \in \boldsymbol{D}T \cup \boldsymbol{D}F} \, \boldsymbol{\theta}_d$$

$$\mathbf{\Theta}_X = (\Sigma_{d \in \boldsymbol{D}X} \, \boldsymbol{\theta}_d)/\mathbf{s}$$

$\boldsymbol{\theta}$ is topic distribution of test review

$\mathbf{\Theta}_T$ is the global topic distribution of true review w.r.t LDA topics.

$\mathbf{\Theta}_F$ is the global topic distribution of true review w.r.t LDA topics

X={T,F} and / is element wise division.

Calculating weight for each topic (1 to 100)

This is calculated using the formula

$$w_{X}^{i}=\log(|\{d\in\boldsymbol{D}_{X}:\theta_{d}^{i}>\tau\}|+1/|\{d\in\boldsymbol{D}_{Y}:\theta_{d}^{i}>\tau\}|+1)$$

where $\{X,Y\}=\{T,F\}$ or $\{F,T\}$ and $\tau=1/\mathrm{K}$

K is the number of topics.

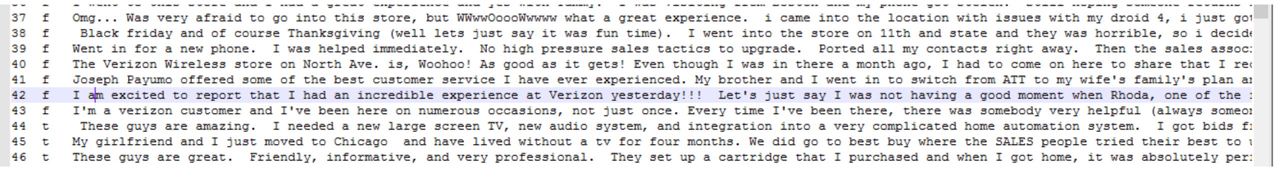
$\mathrm{K}=100, \tau=0.01$

Testing individual review

In the same program input is given a review to be tested and the its score is calculated

Test review:- I am excited to report that I had an incredible experience at Verizon yesterday!!! Let's just say I was not having a good moment when Rhoda, one of the floor managers noticed my dismay and demonstrated impeccable customer service!! She listened to my concern and went above & beyond my expectations to resolve my issue. Rhoda Yuen is truly an asset to the Verizon team and bravo to Verizon for employing someone with outstanding customer service skills!! I am a happy Verizon customer!!!

Review taken from the dataset(fake review)



```
f   Omg... Was very afraid to go into this store, but WWwwOoooWwwww what a great experience.  i came into the location with issues with my droid 4, i just got
f    Black friday and of course Thanksgiving (well lets just say it was fun time).  I went into the store on 11th and state and they was horrible, so i decide
f   Went in for a new phone.  I was helped immediately.  No high pressure sales tactics to upgrade.  Ported all my contacts right away.  Then the sales assoc
f   The Verizon Wireless store on North Ave. is, Woohoo! As good as it gets! Even though I was in there a month ago, I had to come on here to share that I re
f   Joseph Payumo offered some of the best customer service I have ever experienced. My brother and I went in to switch from ATT to my wife's family's plan an
f   I am excited to report that I had an incredible experience at Verizon yesterday!!!  Let's just say I was not having a good moment when Rhoda, one of the
f   I'm a verizon customer and I've been here on numerous occasions, not just once. Every time I've been there, there was somebody very helpful (always someon
t    These guys are amazing.  I needed a new large screen TV, new audio system, and integration into a very complicated home automation system.  I got bids fr
t   My girlfriend and I just moved to Chicago  and have lived without a tv for four months. We did go to best buy where the SALES people tried their best to
t   These guys are great.  Friendly, informative, and very professional.  They set up a cartridge that I purchased and when I got home, it was absolutely per
```

$$\mathrm{Score}_{X}(\mathrm{r})=\Sigma\theta_{r}^{i}\times(\Theta_{X}^{i}+\sigma\times w_{X}^{i}) \text{ where } X=\{T,F\}.$$

$\boldsymbol{\theta}r$ is topic distribution of test review r, which is the result of the inference in LDA

$\mathrm{Score}_{T}(\mathrm{r})=0.867440338$

$\mathrm{Score}_{F}(\mathrm{r})=0.8944$

Fake score is more than true score hence fake review.

The method provides the correct result.

Type of words used in the review

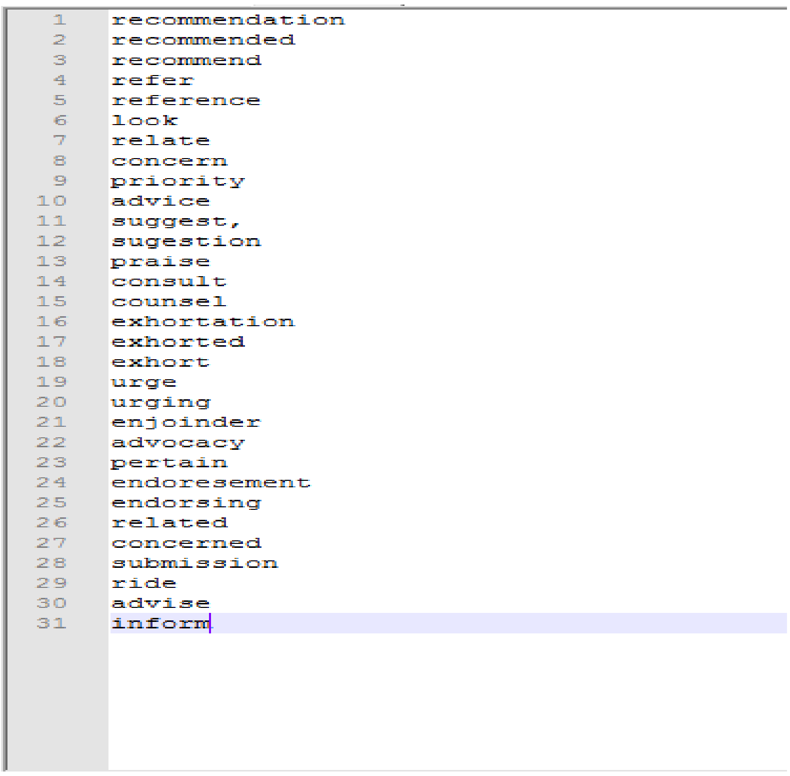
Categorizing words of each topic into five categories.

Each topic contains approximately 20 words. These words are of different kinds. We are mainly interested in five categories of words:

i) Nouns
ii) Verbs
iii) Adjectives
iv) Comparative words
v) Reference words

For first four categories we have used Stanford POS tagger. There is no standard dataset for the fifth category which includes words like reference, recommend. So we have created our own dataset using their synonyms from wordnet. For each topic word count of each type is calculated.

Dataset for fifth category



```
recommendation
recommended
recommend
refer
reference
look
relate
concern
priority
advice
suggest,
sugestion
praise
consult
counsel
exhortation
exhorted
exhort
urge
urging
enjoinder
advocacy
pertain
endoresement
endorsing
related
concerned
submission
ride
advise
inform
```

Steps for counting word of different category in a topic

Step1:-

Counting word of category five (RR)

We have created a java program for this purpose. All the words in the data set mentioned above are stored in an array.

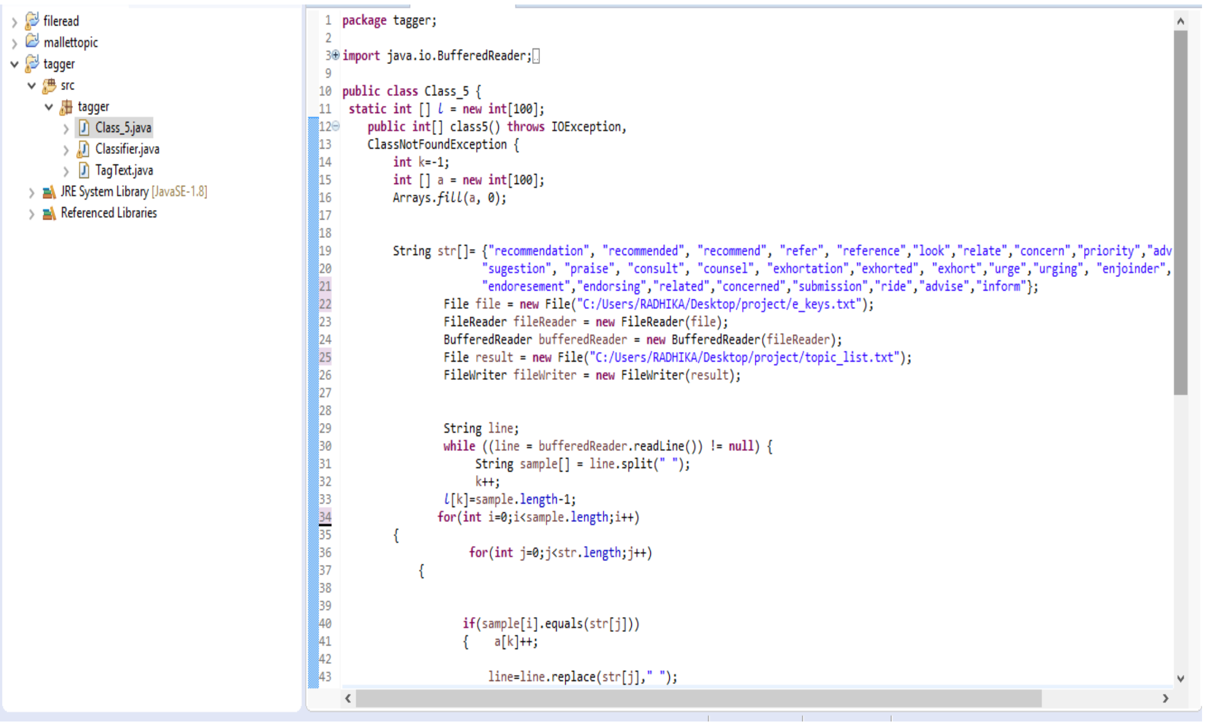
For class Class_5.java

Input: - Document containing topics (e_keys)

Output:-Count the number of words of this type and store the result in an array.

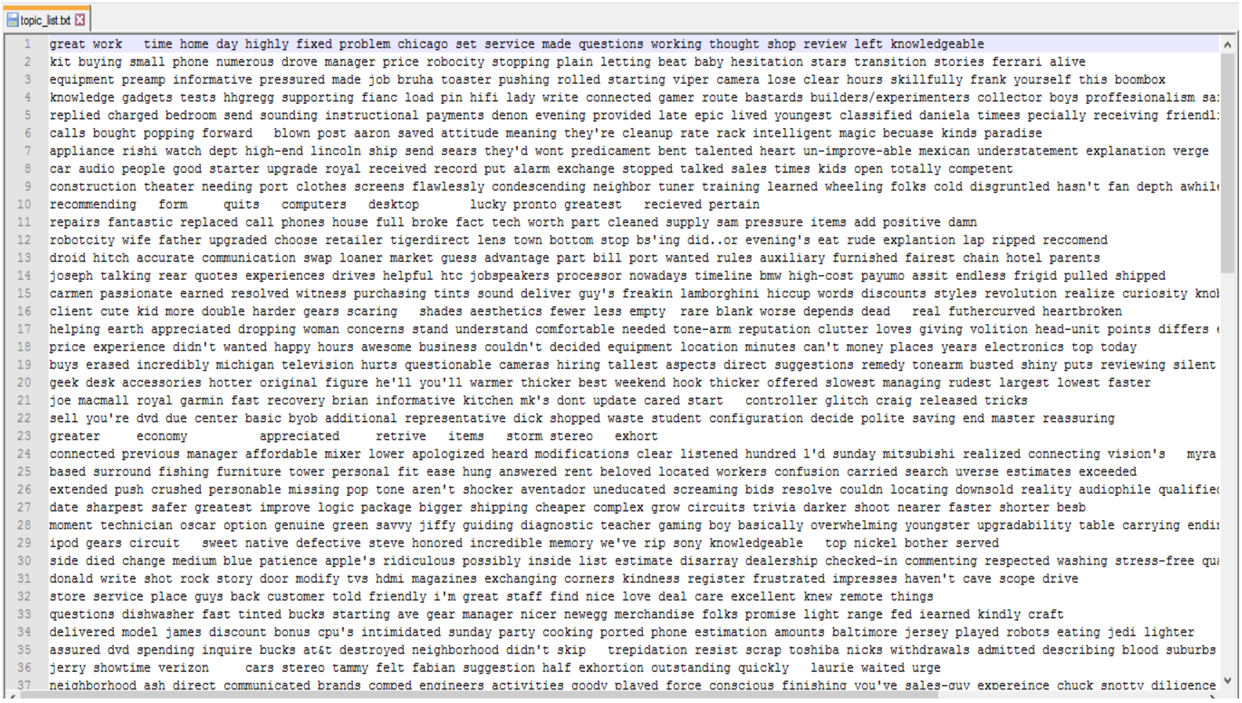
Eliminate these type of words from each topic and store the results in another text file(topic_list.txt)

Class_5.java



```
package tagger;

import java.io.BufferedReader;

public class Class_5 {
 static int [] l = new int[100];
    public int[] class5() throws IOException,
    ClassNotFoundException {
        int k=-1;
        int [] a = new int[100];
        Arrays.fill(a, 0);


        String str[]= {"recommendation", "recommended", "recommend", "refer", "reference","look","relate","concern","priority","adv
                      "sugestion", "praise", "consult", "counsel", "exhortation","exhorted", "exhort","urge","urging", "enjoinder",
                      "endoresement","endorsing","related","concerned","submission","ride","advise","inform"};
                File file = new File("C:/Users/RADHIKA/Desktop/project/e_keys.txt");
                FileReader fileReader = new FileReader(file);
                BufferedReader bufferedReader = new BufferedReader(fileReader);
                File result = new File("C:/Users/RADHIKA/Desktop/project/topic_list.txt");
                FileWriter fileWriter = new FileWriter(result);


                String line;
                while ((line = bufferedReader.readLine()) != null) {
                     String sample[] = line.split(" ");
                     k++;
                l[k]=sample.length-1;
               for(int i=0;i<sample.length;i++)
        {
                    for(int j=0;j<str.length;j++)
            {


                   if(sample[i].equals(str[j]))
                   {    a[k]++;

                       line=line.replace(str[j]," ");
```

Output file topic_list.txt



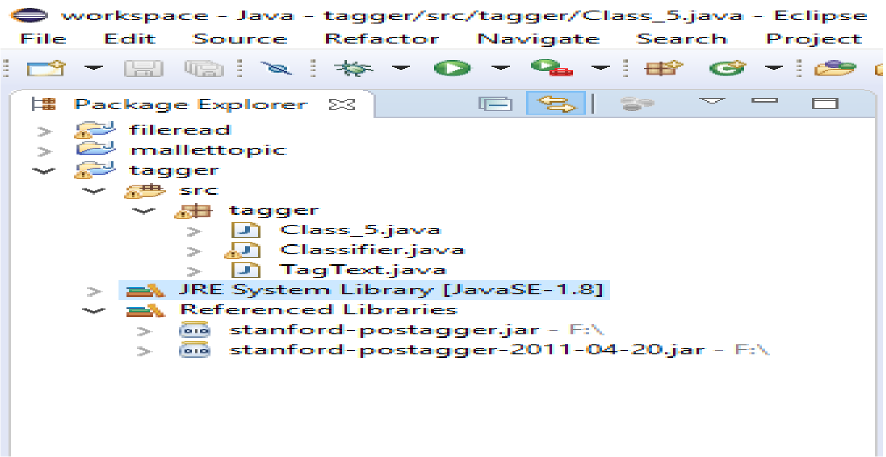
```
great work   time home day highly fixed problem chicago set service made questions working thought shop review left knowledgeable
kit buying small phone numerous drove manager price robocity stopping plain letting beat baby hesitation stars transition stories ferrari alive
equipment preamp informative pressured made job bruha toaster pushing rolled starting viper camera lose clear hours skillfully frank yourself this boombox
knowledge gadgets tests hhgregg supporting fianc load pin hifi lady write connected gamer route bastards builders/experimenters collector boys proffesionalism sa
replied charged bedroom send sounding instructional payments denon evening provided late epic lived youngest classified daniela timees pecially receiving friendl
calls bought popping forward   blown post aaron saved attitude meaning they're cleanup rate rack intelligent magic becuase kinds paradise
appliance rishi watch dept high-end lincoln ship send sears they'd wont predicament bent talented heart un-improve-able mexican understatement explanation verge
car audio people good starter upgrade royal received record put alarm exchange stopped talked sales times kids open totally competent
construction theater needing port clothes screens flawlessly condescending neighbor tuner training learned wheeling folks cold disgruntled hasn't fan depth awhil
recommending   form    quits   computers   desktop       lucky pronto greatest   recieved pertain
repairs fantastic replaced call phones house full broke fact tech worth part cleaned supply sam pressure items add positive damn
robotcity wife father upgraded choose retailer tigerdirect lens town bottom stop bs'ing did..or evening's eat rude explantion lap ripped reccomend
droid hitch accurate communication swap loaner market guess advantage part bill port wanted rules auxiliary furnished fairest chain hotel parents
joseph talking rear quotes experiences drives helpful htc jobspeakers processor nowadays timeline bmw high-cost payumo assit endless frigid pulled shipped
carmen passionate earned resolved witness purchasing tints sound deliver guy's freakin lamborghini hiccup words discounts styles revolution realize curiosity kno
client cute kid more double harder gears scaring   shades aesthetics fewer less empty  rare blank worse depends dead   real futhercurved heartbroken
helping earth appreciated dropping woman concerns stand understand comfortable needed tone-arm reputation clutter loves giving volition head-unit points differs
price experience didn't wanted happy hours awesome business couldn't decided equipment location minutes can't money places years electronics top today
buys erased incredibly michigan television hurts questionable cameras hiring tallest aspects direct suggestions remedy tonearm busted shiny puts reviewing silent
geek desk accessories hotter original figure he'll you'll warmer thicker best weekend hook thicker offered slowest managing rudest largest lowest faster
joe macmall royal garmin fast recovery brian informative kitchen mk's dont update cared start   controller glitch craig released tricks
sell you're dvd due center basic byob additional representative dick shopped waste student configuration decide polite saving end master reassuring
greater     economy          appreciated     retrive   items   storm stereo   exhort
connected previous manager affordable mixer lower apologized heard modifications clear listened hundred l'd sunday mitsubishi realized connecting vision's   myra
based surround fishing furniture tower personal fit ease hung answered rent beloved located workers confusion carried search uverse estimates exceeded
extended push crushed personable missing pop tone aren't shocker aventador uneducated screaming bids resolve couldn locating downsold reality audiophile qualifie
date sharpest safer greatest improve logic package bigger shipping cheaper complex grow circuits trivia darker shoot nearer faster shorter besb
moment technician oscar option genuine green savvy jiffy guiding diagnostic teacher gaming boy basically overwhelming youngster upgradability table carrying endi
ipod gears circuit   sweet native defective steve honored incredible memory we've rip sony knowledgeable   top nickel bother served
side died change medium blue patience apple's ridiculous possibly inside list estimate disarray dealership checked-in commenting respected washing stress-free qu
donald write shot rock story door modify tvs hdmi magazines exchanging corners kindness register frustrated impresses haven't cave scope drive
store service place guys back customer told friendly i'm great staff find nice love deal care excellent knew remote things
questions dishwasher fast tinted bucks starting ave gear manager nicer newegg merchandise folks promise light range fed iearned kindly craft
delivered model james discount bonus cpu's intimidated sunday party cooking ported phone estimation amounts baltimore jersey played robots eating jedi lighter
assured dvd spending inquire bucks at&t destroyed neighborhood didn't skip   trepidation resist scrap toshiba nicks withdrawals admitted describing blood suburbs
jerry showtime verizon     cars stereo tammy felt fabian suggestion half exhortion outstanding quickly   laurie waited urge
neighborhood ash direct communicated brands comped engineers activities goody played force conscious finishing you've sales-guy expereince chuck snotty diligence
```

The word count for each topic is stored in the array arr.

Step2:-

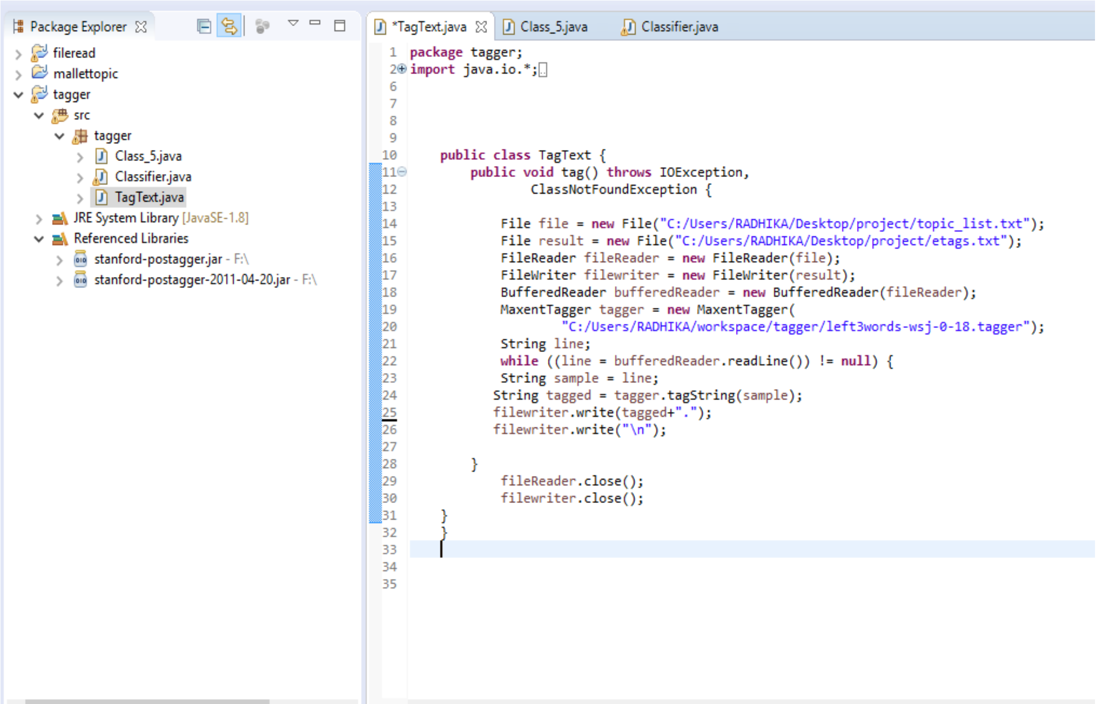
Counting words of types 1 to 4 using POS tagger

Stanford Pos tagger jar file is included in the java program


workspace - Java - tagger/src/tagger/Class_5.java - Eclipse
File
Edit
Source
Refactor
Navigate
Search
Project
Package Explorer
fileread
mallettopic
tagger
src
tagger
Class_5.java
Classifier.java
TagText.java
JRE System Library [JavaSE-1.8]
Referenced Libraries
stanford-postagger.jar - F:\
stanford-postagger-2011-04-20.jar - F:\

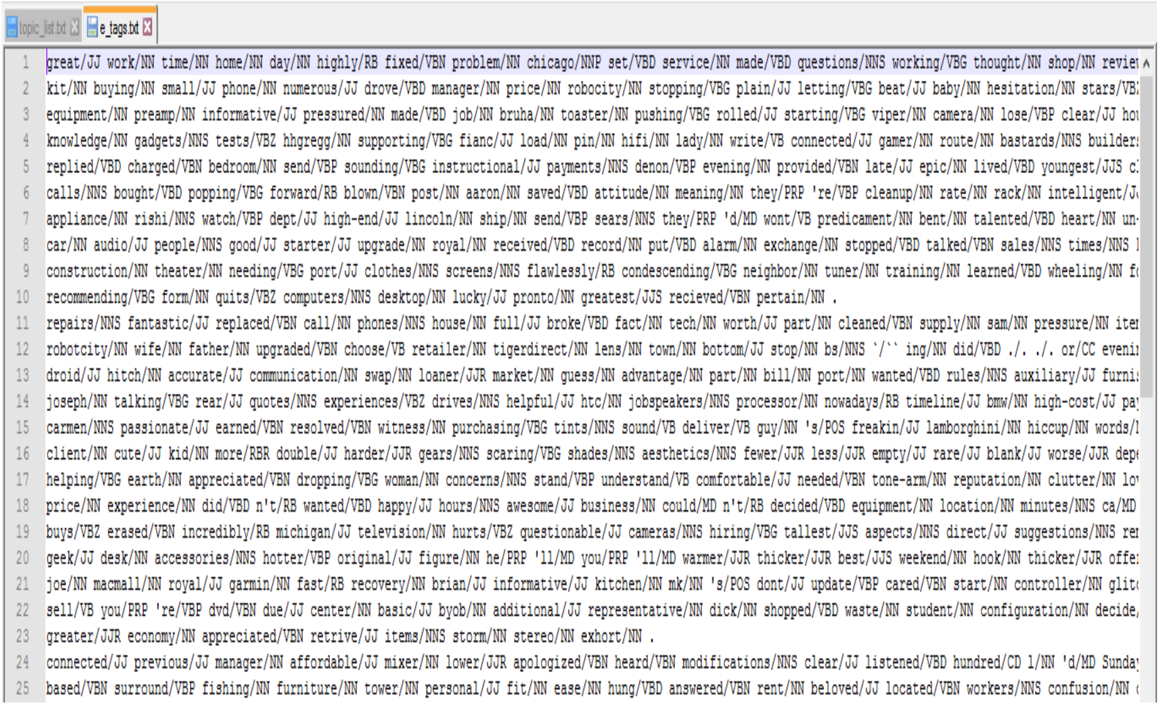
Input of this class TagText.java :- output from class Class_5 (topic_list.txt)

Output:-Text file containing tags for each word



```java
package tagger;
import java.io.*;




    public class TagText {
        public void tag() throws IOException,
                ClassNotFoundException {

            File file = new File("C:/Users/RADHIKA/Desktop/project/topic_list.txt");
            File result = new File("C:/Users/RADHIKA/Desktop/project/etags.txt");
            FileReader fileReader = new FileReader(file);
            FileWriter filewriter = new FileWriter(result);
            BufferedReader bufferedReader = new BufferedReader(fileReader);
            MaxentTagger tagger = new MaxentTagger(
                    "C:/Users/RADHIKA/workspace/tagger/left3words-wsj-0-18.tagger");
            String line;
            while ((line = bufferedReader.readLine()) != null) {
            String sample = line;
           String tagged = tagger.tagString(sample);
           filewriter.write(tagged+".");
           filewriter.write("\n");

        }
            fileReader.close();
            filewriter.close();
    }
    }
```

Output file of class TagText.java



```
great/JJ work/NN time/NN home/NN day/NN highly/RB fixed/VBN problem/NN chicago/NNP set/VBD service/NN made/VBD questions/NNS working/VBG thought/NN shop/NN revie
kit/NN buying/NN small/JJ phone/NN numerous/JJ drove/VBD manager/NN price/NN robocity/NN stopping/VBG plain/JJ letting/VBG beat/JJ baby/NN hesitation/NN stars/VB
equipment/NN preamp/NN informative/JJ pressured/NN made/VBD job/NN bruha/NN toaster/NN pushing/VBG rolled/JJ starting/VBG viper/NN camera/NN lose/VBP clear/JJ hou
knowledge/NN gadgets/NNS tests/VBZ hhgregg/NN supporting/VBG fianc/JJ load/NN pin/NN hifi/NN lady/NN write/VB connected/JJ gamer/NN route/NN bastards/NNS builders
replied/VBD charged/VBN bedroom/NN send/VBP sounding/VBG instructional/JJ payments/NNS denon/VBP evening/NN provided/VBN late/JJ epic/NN lived/VBD youngest/JJS c
calls/NNS bought/VBD popping/VBG forward/RB blown/VBN post/NN aaron/NN saved/VBD attitude/NN meaning/NN they/PRP 're/VBP cleanup/NN rate/NN rack/NN intelligent/J
appliance/NN rishi/NNS watch/VBP dept/JJ high-end/JJ lincoln/NN ship/NN send/VBP sears/NNS they/PRP 'd/MD wont/VB predicament/NN bent/NN talented/VBD heart/NN un-
car/NN audio/JJ people/NNS good/JJ starter/JJ upgrade/NN royal/NN received/VBD record/NN put/VBD alarm/NN exchange/NN stopped/VBD talked/VBN sales/NNS times/NNS
construction/NN theater/NN needing/VBG port/JJ clothes/NNS screens/NNS flawlessly/RB condescending/VBG neighbor/NN tuner/NN training/NN learned/VBD wheeling/NN f
recommending/VBG form/NN quits/VBZ computers/NNS desktop/NN lucky/JJ pronto/NN greatest/JJS recieved/VBN pertain/NN .
repairs/NNS fantastic/JJ replaced/VBN call/NN phones/NNS house/NN full/JJ broke/VBD fact/NN tech/NN worth/JJ part/NN cleaned/VBN supply/NN sam/NN pressure/NN ite
robotcity/NN wife/NN father/NN upgraded/VBN choose/VB retailer/NN tigerdirect/NN lens/NN town/NN bottom/JJ stop/NN bs/NNS `/`` ing/NN did/VBD ./. ./. or/CC evenin
droid/JJ hitch/NN accurate/JJ communication/NN swap/NN loaner/JJR market/NN guess/NN advantage/NN part/NN bill/NN port/NN wanted/VBD rules/NNS auxiliary/JJ furni
joseph/NN talking/VBG rear/JJ quotes/NNS experiences/VBZ drives/NNS helpful/JJ htc/NN jobspeakers/NNS processor/NN nowadays/RB timeline/JJ bmw/NN high-cost/JJ pa
carmen/NNS passionate/JJ earned/VBN resolved/VBN witness/NN purchasing/VBG tints/NNS sound/VB deliver/VB guy/NN 's/POS freakin/JJ lamborghini/NN hiccup/NN words/
client/NN cute/JJ kid/NN more/RBR double/JJ harder/JJR gears/NNS scaring/VBG shades/NNS aesthetics/NNS fewer/JJR less/JJR empty/JJ rare/JJ blank/JJ worse/JJR dep
helping/VBG earth/NN appreciated/VBN dropping/VBG woman/NN concerns/NNS stand/VBP understand/VB comfortable/JJ needed/VBN tone-arm/NN reputation/NN clutter/NN lo
price/NN experience/NN did/VBD n't/RB wanted/VBD happy/JJ hours/NNS awesome/JJ business/NN could/MD n't/RB decided/VBD equipment/NN location/NN minutes/NNS ca/MD
buys/VBZ erased/VBN incredibly/RB michigan/JJ television/NN hurts/VBZ questionable/JJ cameras/NNS hiring/VBG tallest/JJS aspects/NNS direct/JJ suggestions/NNS re
geek/JJ desk/NN accessories/NNS hotter/VBP original/JJ figure/NN he/PRP 'll/MD you/PRP 'll/MD warmer/JJR thicker/JJR best/JJS weekend/NN hook/NN thicker/JJR offe
joe/NN macmall/NN royal/JJ garmin/NN fast/RB recovery/NN brian/JJ informative/JJ kitchen/NN mk/NN 's/POS dont/JJ update/VBP cared/VBN start/NN controller/NN glit
sell/VB you/PRP 're/VBP dvd/VBN due/JJ center/NN basic/JJ byob/NN additional/JJ representative/NN dick/NN shopped/VBD waste/NN student/NN configuration/NN decide
greater/JJR economy/NN appreciated/VBN retrive/JJ items/NNS storm/NN stereo/NN exhort/NN .
connected/JJ previous/JJ manager/NN affordable/JJ mixer/NN lower/JJR apologized/VBN heard/VBN modifications/NNS clear/JJ listened/VBD hundred/CD 1/NN 'd/MD Sunda
based/VBN surround/VBP fishing/NN furniture/NN tower/NN personal/JJ fit/NN ease/NN hung/VBD answered/VBN rent/NN beloved/JJ located/VBN workers/NNS confusion/NN
```

Step3:-

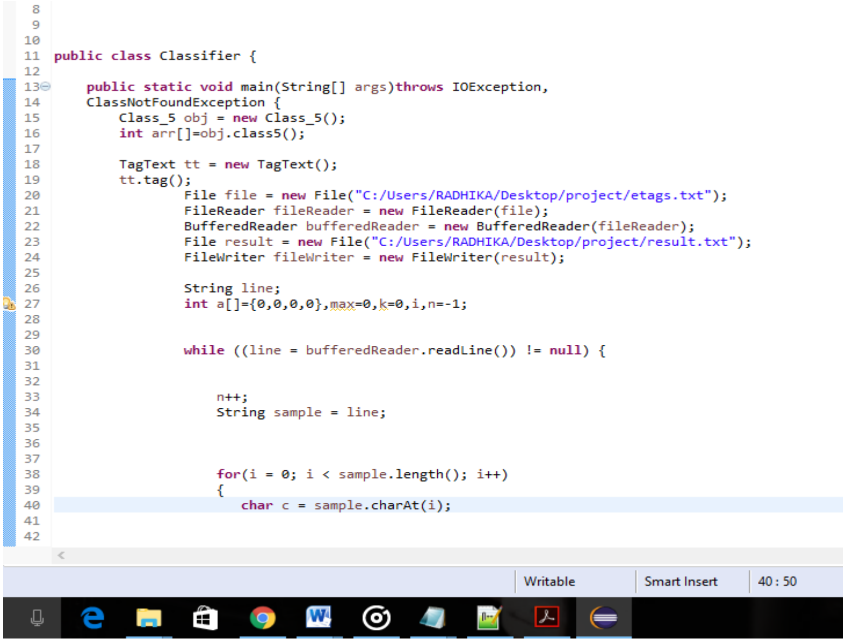
The output file from class TagText.java is send as an input for class Classifier which will count the number of words of different category for every topic.

For class Classifier.java

Input:-Output from class TagTxt.java (e_tags)

Output:-Text file containing two type of classes for training (result.txt)

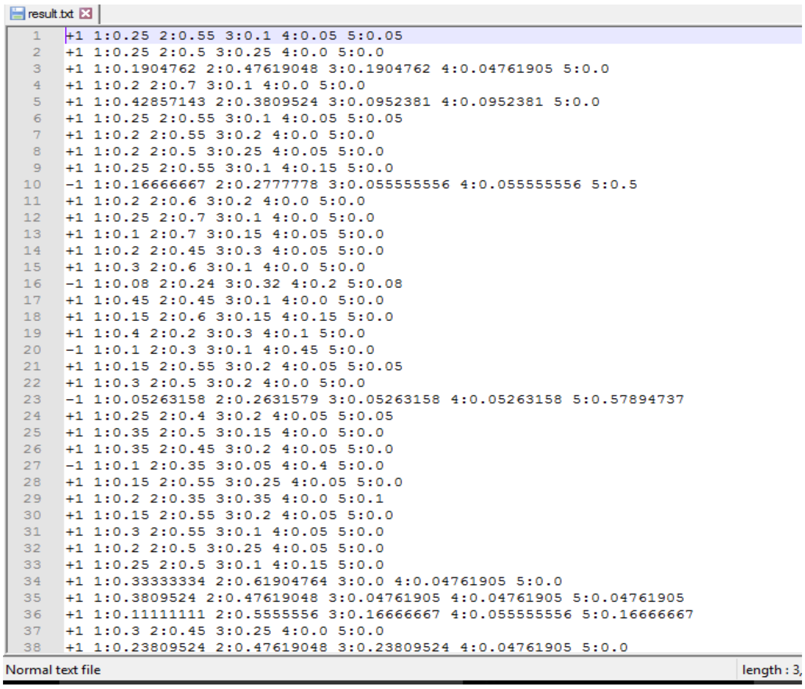
Class Classifier.java



```



public class Classifier {

    public static void main(String[] args)throws IOException,
    ClassNotFoundException {
        Class_5 obj = new Class_5();
        int arr[]=obj.class5();

        TagText tt = new TagText();
        tt.tag();
                File file = new File("C:/Users/RADHIKA/Desktop/project/etags.txt");
                FileReader fileReader = new FileReader(file);
                BufferedReader bufferedReader = new BufferedReader(fileReader);
                File result = new File("C:/Users/RADHIKA/Desktop/project/result.txt");
                FileWriter fileWriter = new FileWriter(result);

                String line;
                int a[]={0,0,0,0},max=0,k=0,i,n=-1;


                while ((line = bufferedReader.readLine()) != null) {


                    n++;
                    String sample = line;



                    for(i = 0; i < sample.length(); i++)
                    {
                       char c = sample.charAt(i);


```

Output of class Classifier.java- Topics for which wordcoount of first two categories

Is more are included in +1 class and rest are included in -1.



```
+1 1:0.25 2:0.55 3:0.1 4:0.05 5:0.05
+1 1:0.25 2:0.5 3:0.25 4:0.0 5:0.0
+1 1:0.1904762 2:0.47619048 3:0.1904762 4:0.04761905 5:0.0
+1 1:0.2 2:0.7 3:0.1 4:0.0 5:0.0
+1 1:0.42857143 2:0.3809524 3:0.0952381 4:0.0952381 5:0.0
+1 1:0.25 2:0.55 3:0.1 4:0.05 5:0.05
+1 1:0.2 2:0.55 3:0.2 4:0.0 5:0.0
+1 1:0.2 2:0.5 3:0.25 4:0.05 5:0.0
+1 1:0.25 2:0.55 3:0.1 4:0.15 5:0.0
-1 1:0.16666667 2:0.2777778 3:0.055555556 4:0.055555556 5:0.5
+1 1:0.2 2:0.6 3:0.2 4:0.0 5:0.0
+1 1:0.25 2:0.7 3:0.1 4:0.0 5:0.0
+1 1:0.1 2:0.7 3:0.15 4:0.05 5:0.0
+1 1:0.2 2:0.45 3:0.3 4:0.05 5:0.0
+1 1:0.3 2:0.6 3:0.1 4:0.0 5:0.0
-1 1:0.08 2:0.24 3:0.32 4:0.2 5:0.08
+1 1:0.45 2:0.45 3:0.1 4:0.0 5:0.0
+1 1:0.15 2:0.6 3:0.15 4:0.15 5:0.0
+1 1:0.4 2:0.2 3:0.3 4:0.1 5:0.0
-1 1:0.1 2:0.3 3:0.1 4:0.45 5:0.0
+1 1:0.15 2:0.55 3:0.2 4:0.05 5:0.05
+1 1:0.3 2:0.5 3:0.2 4:0.0 5:0.0
-1 1:0.05263158 2:0.2631579 3:0.05263158 4:0.05263158 5:0.57894737
+1 1:0.25 2:0.4 3:0.2 4:0.05 5:0.05
+1 1:0.35 2:0.5 3:0.15 4:0.0 5:0.0
+1 1:0.35 2:0.45 3:0.2 4:0.05 5:0.0
-1 1:0.1 2:0.35 3:0.05 4:0.4 5:0.0
+1 1:0.15 2:0.55 3:0.25 4:0.05 5:0.0
+1 1:0.2 2:0.35 3:0.35 4:0.0 5:0.1
+1 1:0.15 2:0.55 3:0.2 4:0.05 5:0.0
+1 1:0.3 2:0.55 3:0.1 4:0.05 5:0.0
+1 1:0.2 2:0.5 3:0.25 4:0.05 5:0.0
+1 1:0.25 2:0.5 3:0.1 4:0.15 5:0.0
+1 1:0.33333334 2:0.61904764 3:0.0 4:0.04761905 5:0.0
+1 1:0.3809524 2:0.47619048 3:0.04761905 4:0.04761905 5:0.04761905
+1 1:0.11111111 2:0.5555556 3:0.16666667 4:0.055555556 5:0.16666667
+1 1:0.3 2:0.45 3:0.25 4:0.0 5:0.0
+1 1:0.23809524 2:0.47619048 3:0.23809524 4:0.04761905 5:0.0
```

Creating Training data for LIBSVM

A library of Support Vector machine is used to train the data.

Training vector is

<label> <index1><value1> <index2><value2> . .

.

.

Label is a class label in our case + and -1

<index><value> forms the attribute where index is integer starting from1 and value is a real number.

For our training data set

Index ranges from 1to 5

Value is a real number indicating the probability of each category of words in each topic.

For example:-

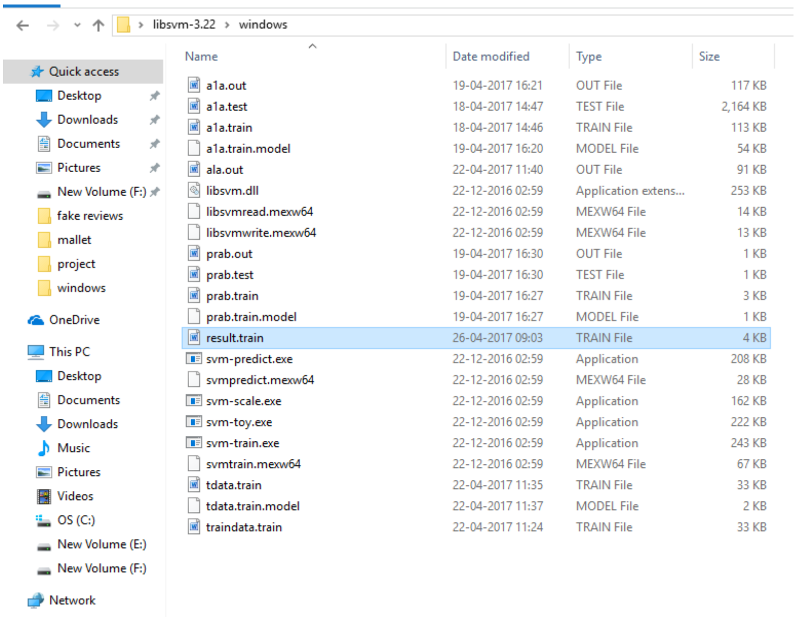
+1 1:0.4 2:0.2 3:0.3 4:0.1 5:0.0 (considered to be true)

-1 1:0.1 2:0.3 3:0.1 4:0.45 5:0.0 (considered to be fake)

To train the data in LIBSVM

Step1:-

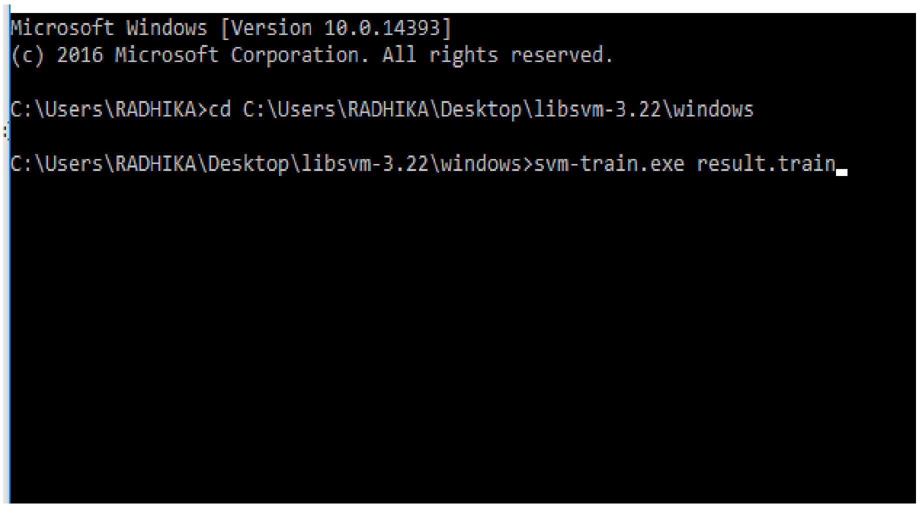
Include the result.train file in the windows folder of LIBSVM.



Name	Date modified	Type	Size
a1a.out	19-04-2017 16:21	OUT File	117 KB
a1a.test	18-04-2017 14:47	TEST File	2,164 KB
a1a.train	18-04-2017 14:46	TRAIN File	113 KB
a1a.train.model	19-04-2017 16:20	MODEL File	54 KB
ala.out	22-04-2017 11:40	OUT File	91 KB
libsvm.dll	22-12-2016 02:59	Application extens...	253 KB
libsvmread.mexw64	22-12-2016 02:59	MEXW64 File	14 KB
libsvmwrite.mexw64	22-12-2016 02:59	MEXW64 File	13 KB
prab.out	19-04-2017 16:30	OUT File	1 KB
prab.test	19-04-2017 16:30	TEST File	1 KB
prab.train	19-04-2017 16:27	TRAIN File	3 KB
prab.train.model	19-04-2017 16:27	MODEL File	1 KB
result.train	26-04-2017 09:03	TRAIN File	4 KB
svm-predict.exe	22-12-2016 02:59	Application	208 KB
svmpredict.mexw64	22-12-2016 02:59	MEXW64 File	28 KB
svm-scale.exe	22-12-2016 02:59	Application	162 KB
svm-toy.exe	22-12-2016 02:59	Application	222 KB
svm-train.exe	22-12-2016 02:59	Application	243 KB
svmtrain.mexw64	22-12-2016 02:59	MEXW64 File	67 KB
tdata.train	22-04-2017 11:35	TRAIN File	33 KB
tdata.train.model	22-04-2017 11:37	MODEL File	2 KB
traindata.train	22-04-2017 11:24	TRAIN File	33 KB

Step2:-

In command prompt give the path of windows directory and write the command for training.



```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\RADHIKA>cd C:\Users\RADHIKA\Desktop\libsvm-3.22\windows

C:\Users\RADHIKA\Desktop\libsvm-3.22\windows>svm-train.exe result.train
```

4.1 PERFORMANCE ANALYSIS

Method using weighted average score prove to be more efficient.

We have taken 10 true and 10 fake reviews for testing.

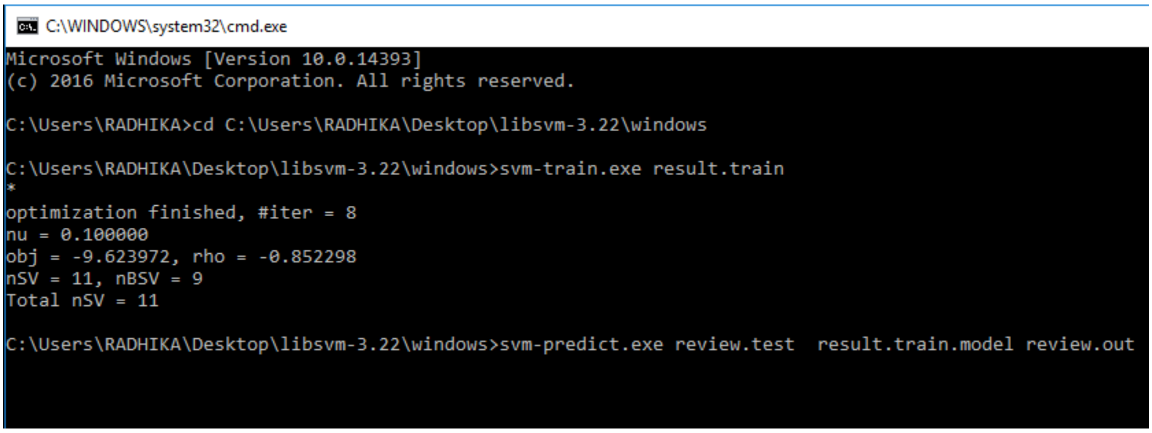
Score of each review is calculated and the method gives correct results for 17 reviews tested. For rest of the three score came out to be little different from expected.

For all the domains we have used the above steps to creating the training data. Training data is used for individual review testing for that each review is converted into factor form and then tested.

Table 4: Results of review testing

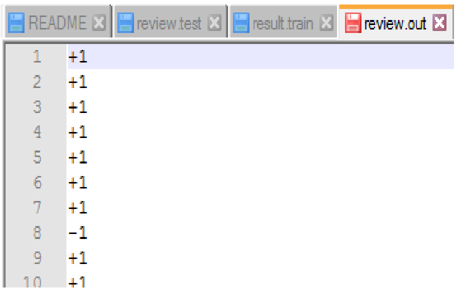
S.NO.	Review	Type	Score	Result
1.	RobotCity Workshop is such a cool small……….	True	t-0.543 f-0.467	correct
2.	Moved into an apartment with lots of equipment……	True	t-0.456 f-0.49	wrong
3.	I found these guys on Yelp and used………	True	t-0.752 f-0.678	correct
4.	Just called these guys on the phone……..	True	t-0.654 f-0.562	correct
5.	Joseph Payumo offered some of the best……..	False	t-0.273 f-0.432	correct

Testing reviews



```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\RADHIKA>cd C:\Users\RADHIKA\Desktop\libsvm-3.22\windows

C:\Users\RADHIKA\Desktop\libsvm-3.22\windows>svm-train.exe result.train
*
optimization finished, #iter = 8
nu = 0.100000
obj = -9.623972, rho = -0.852298
nSV = 11, nBSV = 9
Total nSV = 11

C:\Users\RADHIKA\Desktop\libsvm-3.22\windows>svm-predict.exe review.test  result.train.model review.out
```

Result is in this form



```
+1
+1
+1
+1
+1
+1
+1
-1
+1
+1
```

Result comes out to be correct for 6 reviews out of 10.

CONCLUSION

5.1) Challenges

Global topic distribution serves a good method for detecting fake reviews. It involves large number of calculation.

Even though the proposed method of word type provide automated and simple classification of reviews there is a lot of improvement to be taken in account.

The five categories included may not always be true for the entire fake and true reviews .A fake review writer can also use more nouns and verbs in the sentence. In such a case our method will not be able to produce correct results.

There is no proper criterion to choose number of topics from each domain. The choice of 100 topics can be correct for a domain but may not be for other.

All these problems are challenges which are required to be resolved for more accuracy of results.

5.2)Result

The topic choice and word choice pattern provides efficient method for detection of fake review. These reviews are written for different purposes the type of word and the topic can be used to different between two kind

The approach helps to identified various features of true and fake reviews which can be used as a differentiator between the two. The Model also automates the process of fake review detection. Even though there is a lot room for improvement the method proved to be simple, less time consuming and involving various machine learning techniques

5.3) Conclusion

Review spam detection is an area of research from the last few years because of its huge impact on the costumer’s decision making and companies’ name and structure. There are different methods for detection of fake reviews including manual methods, supervised learning or both. There is no proper standard technique for review spam detection as all the methods includes assumption that can be proved wrong at one point or another.

This motivates for further improvement in the area to provide useful and correct information to the customer.

6) REFERENCES:

1. Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In*Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 219-230). ACM.

2. Duh, A., Štiglic, G., & Korošak, D. (2013, October). Enhancing identification of opinion spammer groups. In *Proceedings of International Conference on Making Sense of Converging Media* (p. 326). ACM.

3. Dewang, R. K., & Singh, A. K. (2015, September). Identification of Fake Reviews Using New Set of Lexical and Syntactic Features. In *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015* (pp. 115-119). ACM.

4. Sharma, K., & Lin, K. I. (2013, April). Review spam detector with rating consistency check. In *Proceedings of the 51st ACM Southeast Conference*(p. 34). ACM.

5. Banerjee, S., & Chua, A. Y. (2014, January). A study of manipulative and authentic negative reviews. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication* (p. 76). ACM.

6. Mukherjee, A., Liu, B., & Glance, N. (2012, April). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web* (pp. 191-200). ACM.

7. Lu, Y., Zhang, L., Xiao, Y., & Li, Y. (2013, May). Simultaneously detecting fake reviews and review spammers using factor graph model. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 225-233). ACM

8. Zhiyuli, A., Liang, X., & Wang, Y. (2015, October). Discerning the Trend: Concealing Deceptive Reviews. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on* (pp. 1833-1838). IEEE.

9. Lee, K. D., Han, K. A., & Myaeng, S. H. (2016). Capturing Word Choice Patterns with LDA for Fake Review Detection in Sentiment Analysis. In*WIMS* (p. 9).

10. Web, As. "Suspicious Behavior Detection: Current Trends and Future Directions.

11. Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *2*(1).

12. MALLET learning : http://mallet.cs.umass.edu/

13. https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003

14. SVM learning: http://www.support –vector-machine.org/SVM_osh.html
15. LIBSVM: https: //www.csie.ntu.edu.tw/~cjlin/libsvm/
16. Machine Learning: https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/
17. POS Tagging: https://nlp.stanford.edu/software/tagger.shtml