Processing User Service Agreement using various Data Mining Techniques.

“**I have read and agree to the Terms**”

is the biggest lie on the web.

We aim to fix that!

WHY THIS TOPIC?

- **A new Deloitte survey found that over 90% of consumers accept legal terms and conditions without reading them.**
- **When faced with no choice, users are willing to accept potential consequences in exchange for access.**
- **For younger people, ages 18-34, the rate is even higher with 97% agreeing to conditions before reading.**

The High Level Steps:

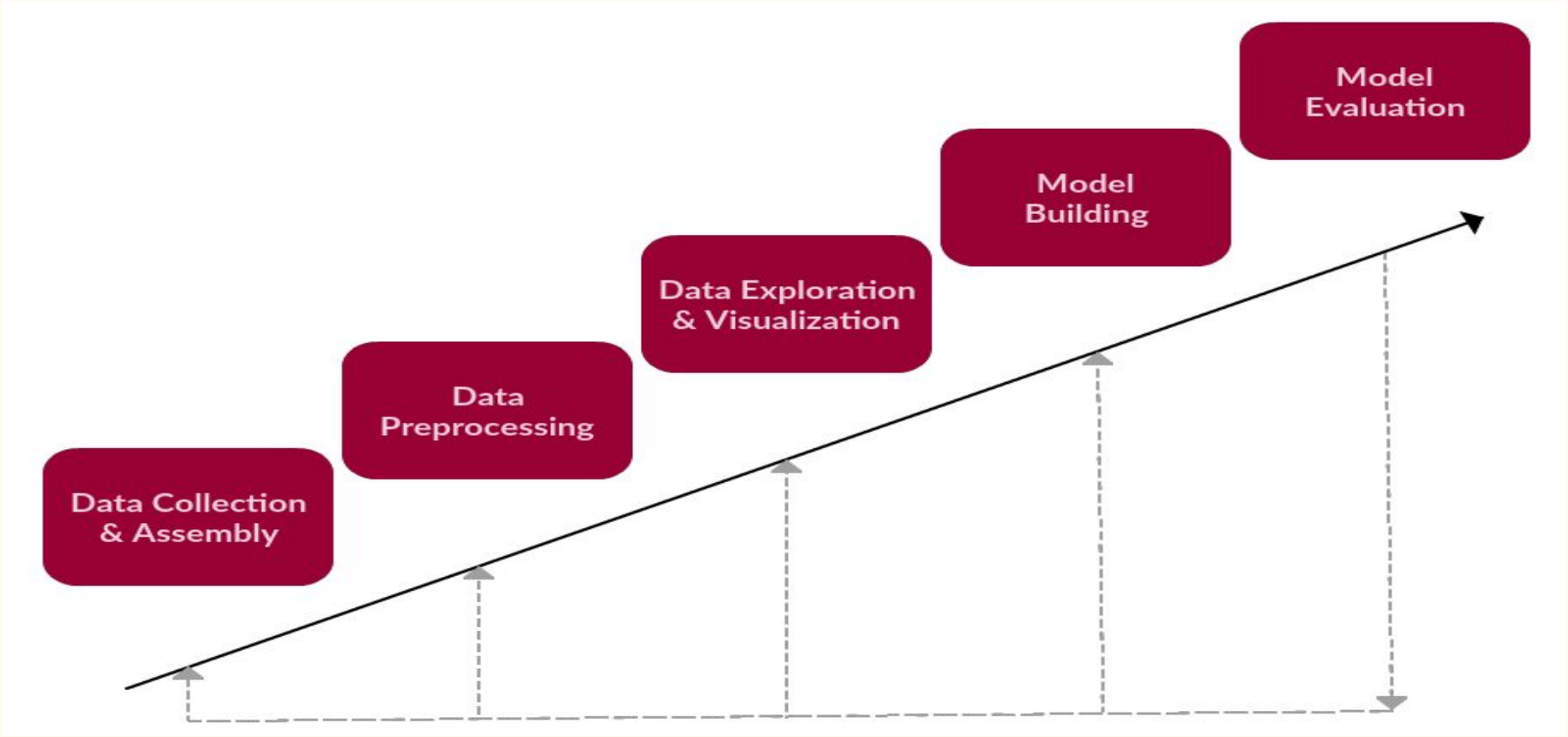


Image source:kdnuggets

We are doing an extractive summarization.

Data Collection

DATA COLLECTION & ASSEMBLY

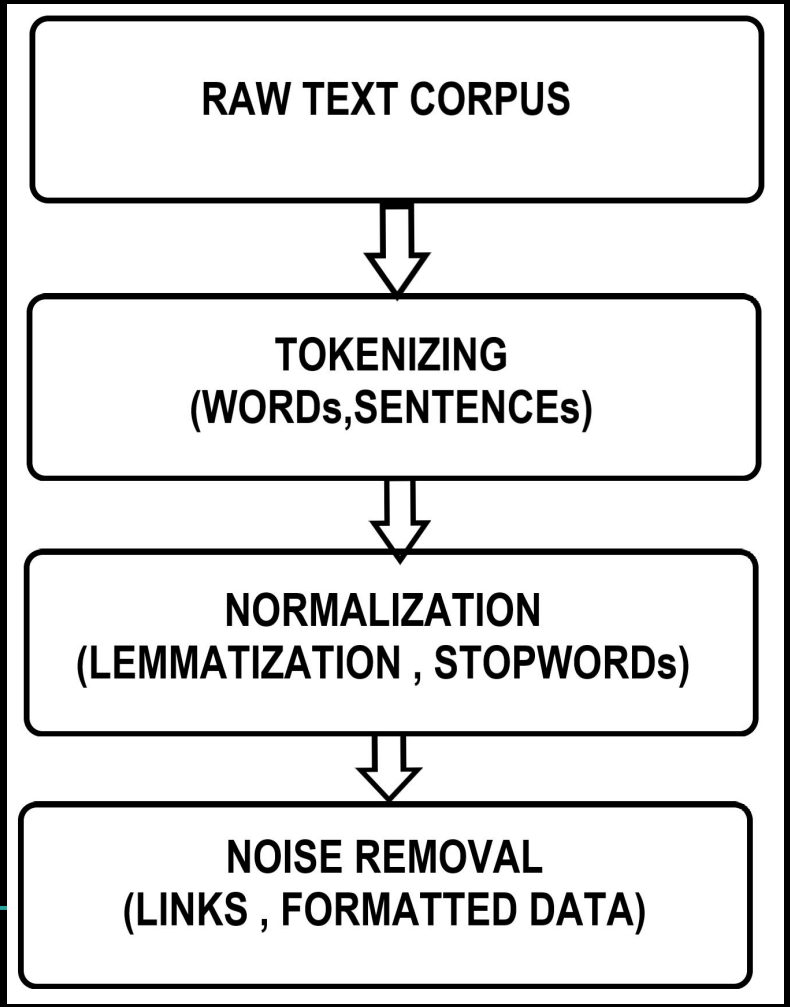
- ***There is no structured dataset present related to TERMS & CONDITIONS.***
- ***We collected raw data from various web applications and softwares.***

Data Collection

1. Visit the terms or privacy policy page of the software or the web apps.
2. Copied the policy and created the file for the same.
3. We have collected total 176 files in total with approx 300 lines per file.
4. And more than million words

PreProcessing of Data

Steps in our Preprocessing



Tokenizing

We are tokenizing raw data and slicing it by using the nltk library and some other self made techniques.

After tokenizing the output will be the tokens of words, sentences

Normalization

Lemmatizing:

We lemmatize our tokens and remove all the stopwords with the help of nltk library.

Noise Removal

We remove the noise from the tokens.

Noise : we define noise as the words like URL, Abbreviations and the 2 length words without any use.

Methodology

TWO STEP CLUSTERING

Procedure for Two Step Clustering

Step 1: After preprocessing the data , we calculate the feature value by computing the TF value,and with these values we have our feature vector.

Step 2 :With these feature vectors we have created, we performed K-Means Clustering to get the best relativity among the words from our corpus.To get the best ‘k’ for our clustering process we have used Elbow Method.

Step 3: After generating all the clusters,we analyzed the top two clusters matching our objective i.e. privacy . we then assigned each cluster a score,which is calculated as the inverse of the number of words in that particular cluster.This ends our step 1 of two step clustering.

Two Step Clustering(Cont.)

Step 4: After getting the desired clusters we have applied the below given algorithm on the clusters.

```
For Files in Corpus:
        For Sentences in Files:
                For Words in Sentences:
                        Check the cluster label for each word & Increment word count for that cluster
                Dictionary is formed in 'Z' format
                assign sentences a label on the basis of 'Label Formula'
        return Sentences with labels
```

With this algorithm we get the sentences with the labels for all the words in the corpus .with this Z format, dictionary we then apply the heuristics genrated in the Label format and then with this labelled sentences we extract the those labelled sentences from those important determined clusters and then show them in the final summary.

'Label Formula'

```
value = 0
if Cluster Label belongs to Important Cluster:
        value+= Word Count *Constant Number * Cluster Score
else:
        value+= Word Count * 1* Cluster Score
Sentence label = int(value)% Total Cluster
```

'Z' Format

Sentence : {'Cluster1' : word count, 'Cluster2': word count ..... 'ClusterN': word count }

RULE BASED MINING

STEPS FOR RULE BASED MINING

Step1: Filtered the unique words from the corpus.

Step2: Weighted each word in corpus with tf-idf score.

Step3: With the help of heuristic dictionary(which basically a dictionary of words which is created by us to enhance our result to fulfil our project objective). The word in our heuristic dictionary are used as **antecedent** and the unique words in our corpus is used as **consequent.**On this basis we have created the Confidence Matrix.

STEPS FOR RULE BASED MINING(contd.)

step4: Finding association rules : we had find the min-support and min-confidence by hit and trial initially using mean then later working with different parameters.

Step5: In this we find the relevant sentences by mapping the various association rules to them , if any of the rule exist we are creating a map with the association rule and the associated sentences along the weight of each sentence where the weight of a sentence is calculated as.

Weight of sentence W_s

$W_s = \sum$ tfidf(word) for all words belongs to Sentence **S**

STEPS FOR RULE BASED MINING(contd.)

Step6: Later for each rule found and associated sentence , weight we have taken mean of those values then selecting only those sentence greater than the threshold(mean).Only these sentence will be in present in our summary.

Ground Truth

Ground Truth for Summary

We manually generated the ground truth by reading some of the test files and then compare from them.

So What about fairness?

Since this is a human generated Privacy summary then we can say it also not perfect and hence we can say it is somewhat biased towards our project objective .

But instead we are showing these Accuracy results on this ground truth only.

Accuracy

Accuracy based on similarity

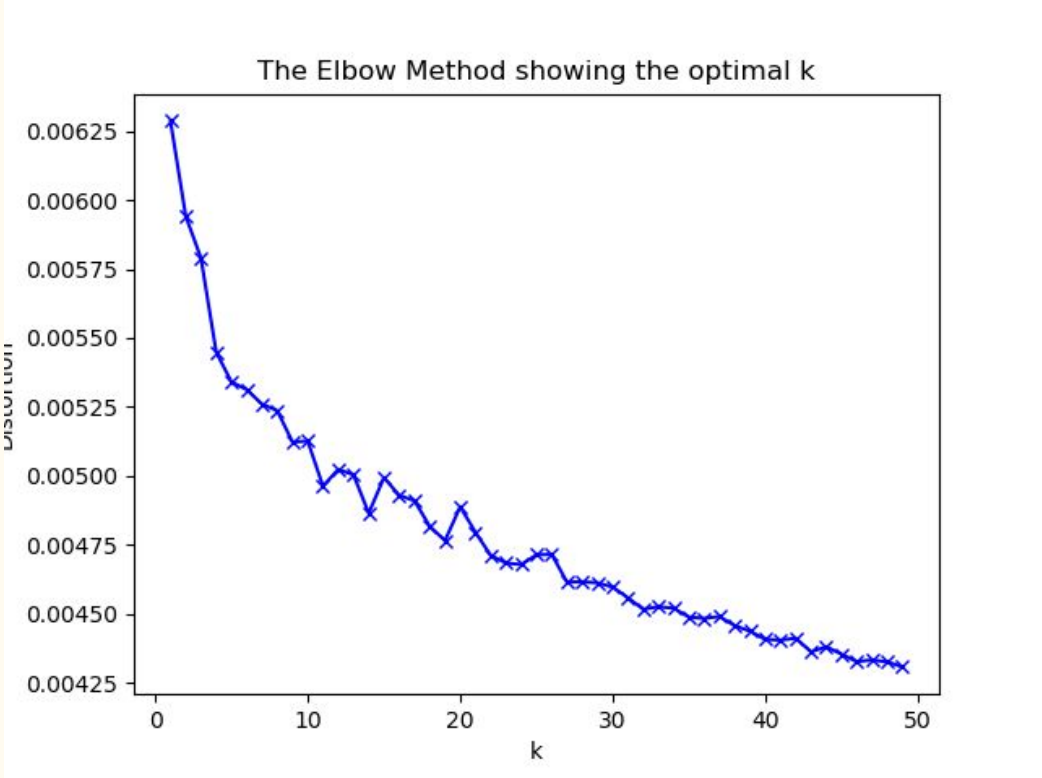
Our Accuracy is based on the similarity of our Ground Truth privacy summary and our system generated Privacy Summary.

How we get Similarity:

- We use the Cosine similarity to do the job for us.
- We have implemented this similarity for 1-GRAM analysis

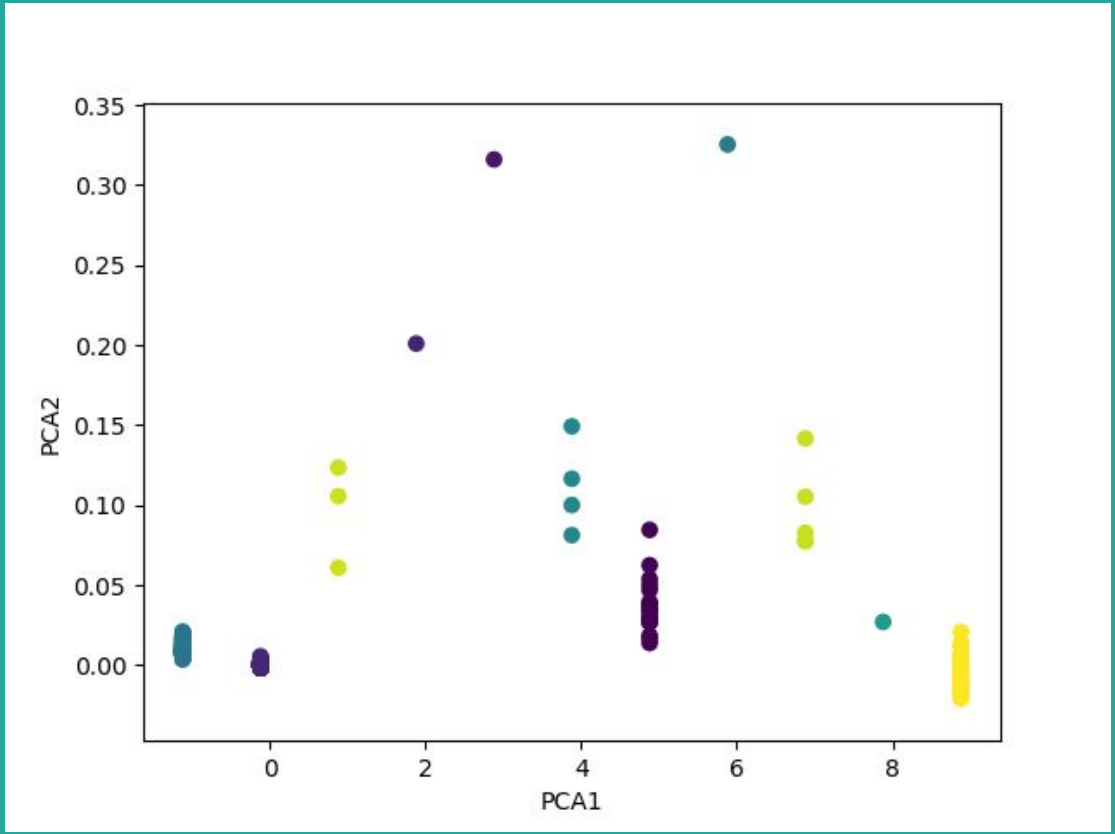
Method1: RESULTS

Method to select Cluster



Cluster

-



Accuracy Results method1

Filename	Accuracy with respect to ground truth
YouTube	75.5%
DuckDuckGo	64%
Facebook	83.45%
Paytm	60.89%
Loco	85.59%
Jabong	55.1%
GitHub	71.56%

Summary Results

another way to prevent search leakage is by using something called a post request, which has the effect of not showing your search in your browser, and, as a consequence, does not send it to other sites. you can turn on post requests on our settings page, but it has its own issues. post requests usually break browser back buttons, and they make it impossible for you to easily share your search by copying and pasting it out of your web browser's address bar.
other search engines save your search history. usually your searches are saved along with the date and time of the search, some information about your computer (e.g. your ip address, user agent and often a unique identifier stored in a browser cookie), and if you are logged in, your account information (e.g. name and email address).
in addition, when you visit any site, your computer automatically sends information about it to that site (including your user agent and ip address). this information can often be used to identify you directly.
similarly, we may add an affiliate code to some ecommerce sites (e.g. amazon & ebay) that results in small commissions being paid back to duckduckgo when you make purchases at those sites. we do not use any third parties to do the code insertion, and we do not work with any sites that share personally identifiable information (e.g. name, address, etc.) via their affiliate programs. this means that no information is shared from duckduckgo to the sites, and the only information that is collected from this process is product information, which is not tied to any particular user and which we do not save or store on our end. it is completely analogous to the search result case from the previous paragraph--we can see anonymous product info such that we cannot tie them to any particular person (or even tie multiple purchases together). this whole affiliate process is an attempt to keep advertising to a minimal level on duckduckgo.
duckduckgo does not collect or share personal information. that is our privacy policy in a nutshell. the rest of this page tries to explain why you should care. with only the timestamp and computer information, your searches can often be traced directly to you. with the additional account information, they are associated directly with you.

Summary of Duck Duck Go

Method2: Results

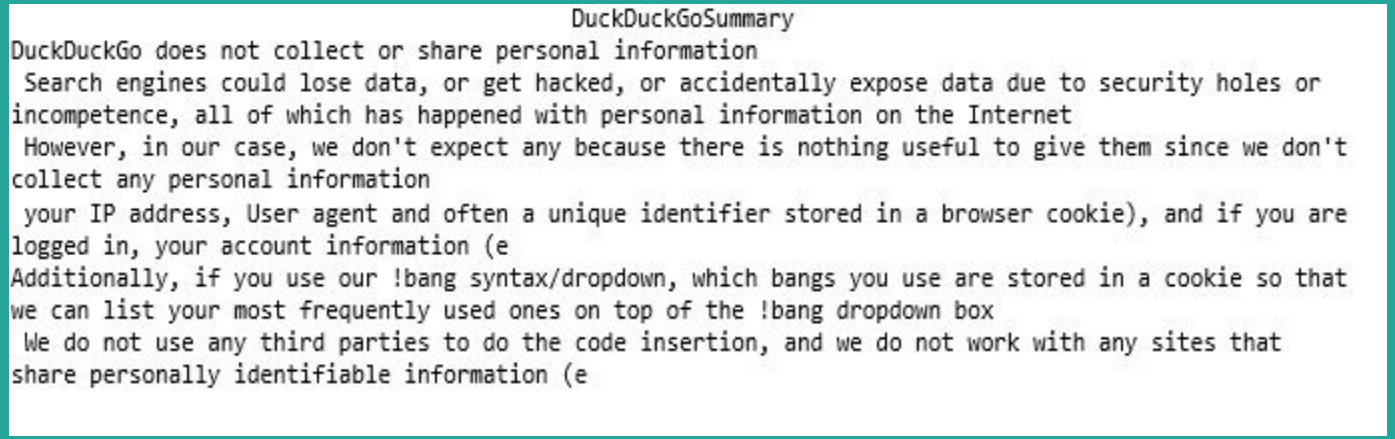
Rules generated

```
protection --> information
protection --> use
money --> law
money --> may
money --> account
money --> information
money --> make
money --> use
money --> service
privacy --> may
privacy --> use
nonexclusive --> policy
nonexclusive --> product
nonexclusive --> arise
nonexclusive --> connection
nonexclusive --> agree
nonexclusive --> claim
nonexclusive --> must
nonexclusive --> remain
```

Accuracy Results method2

Filename	Accuracy with respect to ground truth
YouTube	87.5%
DuckDuckGo	60%
Facebook	88.23%
Paytm	68.42%
Loco	82.75%
Jabong	77.5%
GitHub	78.12%

Summary Results


DuckDuckGoSummary
DuckDuckGo does not collect or share personal information
Search engines could lose data, or get hacked, or accidentally expose data due to security holes or incompetence, all of which has happened with personal information on the Internet
However, in our case, we don't expect any because there is nothing useful to give them since we don't collect any personal information
your IP address, User agent and often a unique identifier stored in a browser cookie), and if you are logged in, your account information (e
Additionally, if you use our !bang syntax/dropdown, which bangs you use are stored in a cookie so that we can list your most frequently used ones on top of the !bang dropdown box
We do not use any third parties to do the code insertion, and we do not work with any sites that share personally identifiable information (e

Summary of Duck Duck Go

Literature survey:

http://faculty.iitmandi.ac.in/~arti/mtp/shubham.pdf

https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-revisited/

https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html

https://waset.org/publications/3514/mining-association-rules-from-unstructured-documents