# NEWS Recommendation

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A company is developing a news app named JhakaasNewsVala with a promises to deliver unique customer experiences. Company has identified target market of working professionals in age group 21-40. JhakaasNewsVala , is looking to optimize first-visit as lucrative as possible for users.

As part of their data science team we will try to optimize first-visit as lucrative as possible for users. Retention of user is also a critical path. We will approach this problem in a systematic way using CRISP-DM methodology. CRISP-DM breaks down into six major phases in data mining. Following are the steps which take place sequentially.

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modelling**
5. **Evaluation**
6. **Deployment**

**Business Understanding**

Jhakaas Newswala mobile application should be able to utilize the UI real estate appropriately to display diverse yet relatable news at the cold start. Statistics show that the number of users scrolling beyond the first set of stories drops off very quickly thus app should display only 8-10 news depending on the screen size.

Application would need to send variety of events to backend servers for better recommender system and user profiling. Events can be as follows:-

1. Initial news feed events: - This lets application know user have logged-in. Additional information would be news served as part of list and at which position.
2. Click-through events :- User clicks on the news link
3. Read percentage events: - Here events will be triggered after user has read more than 25 / 50 / 75% of article.
4. Engaged user events: - User has liked or commented on the news article.

Each event should consist of following data points. Few of the fields might not be populated for each event. For e.g. is\_engaged will be populated only when user likes or comments on the article.

User Id -- this can be the IP address of the system/mobile

Session Id – auto-generated id using timestamp & user id

Article Id – serial id which uniquely identifies the news article

Article Rank – the article sequence is shown to the user

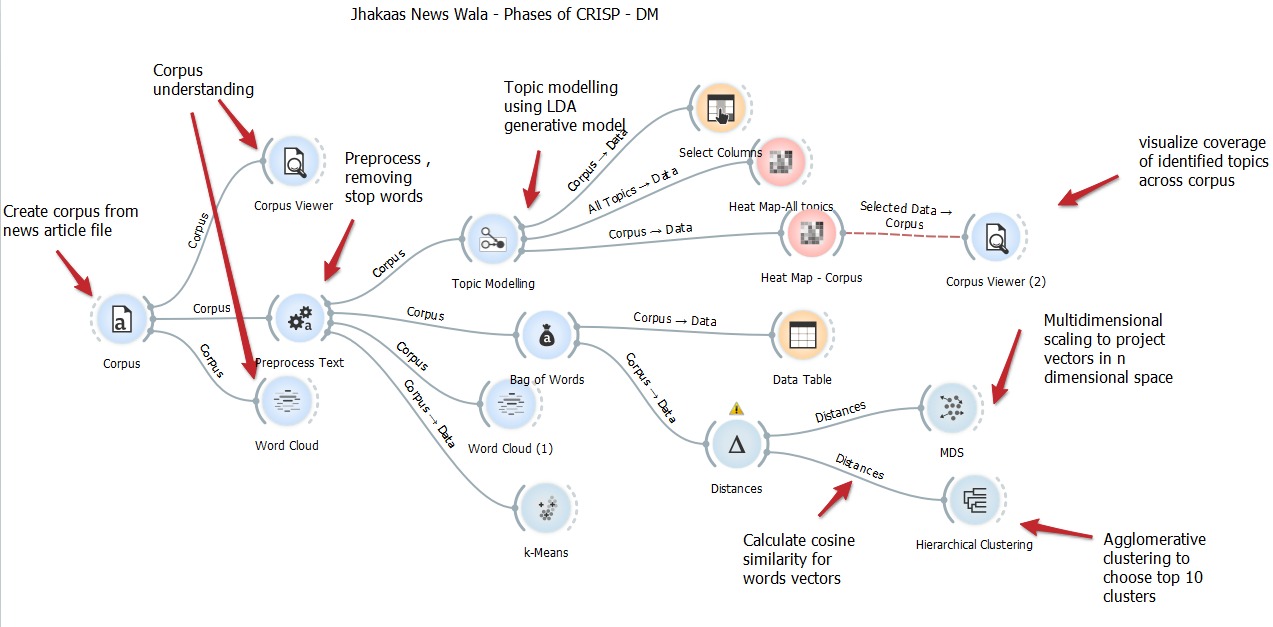
Click – a binary variable to indicate that user clicks on news link for reading the news

Time spent – Amount of time spent on the news article

Scroll percentage – scroll % depicting user actually read article till which point

Is\_engaged – a binary variable to indicate the user is interested in the news.

For further steps we have created a flow diagram using Orange for better understanding. As this tool quickly analyses the data We have performed multiple clustering techniques to compare and come up with each other. Lets go through this step by step :-



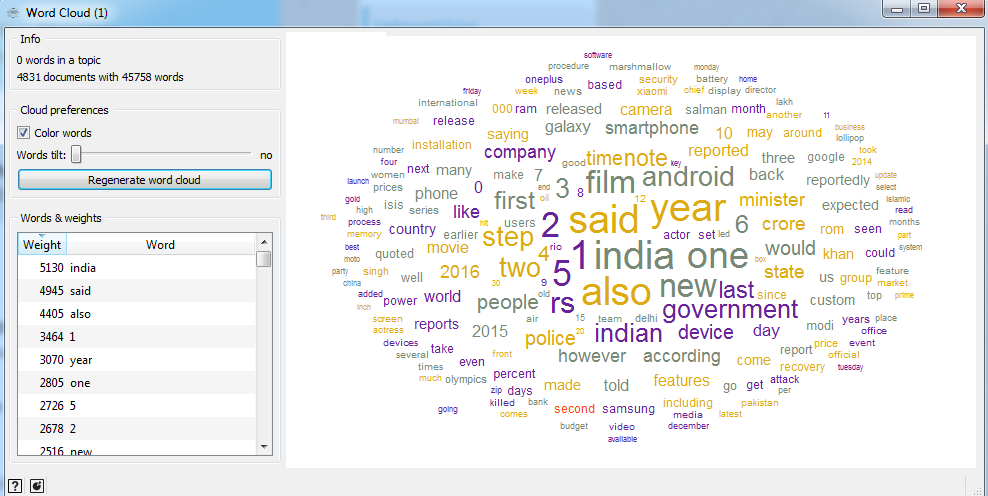
**Data Understanding**

Let’s input the news corpus NewsArticle.csv to the system and get some sights into it.

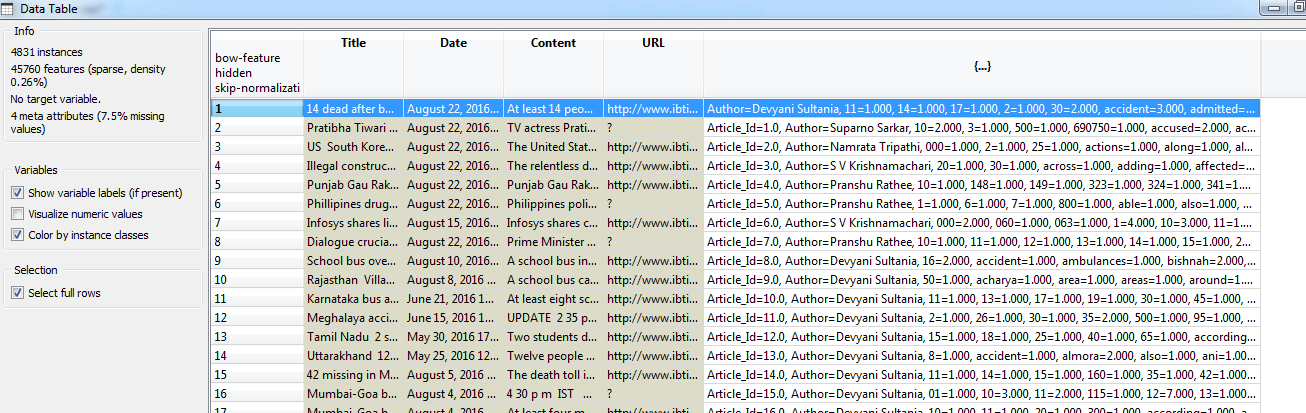
There are total 4831 news articles, average words in the article content is about 220 words.

**Data Preparation**

As part of preprocessing text we will be removing stop words and punctuations. After the processing we get bag of words. Let’s examine the Word Cloud of these bag of words.

****

As per the number of word occurances they are assigned some weightage as well when viewed using data table.

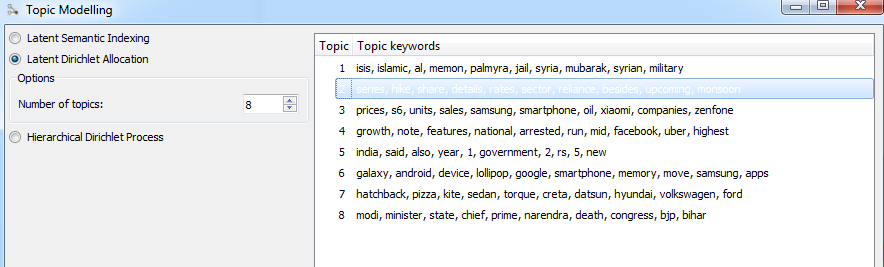


**Modelling**

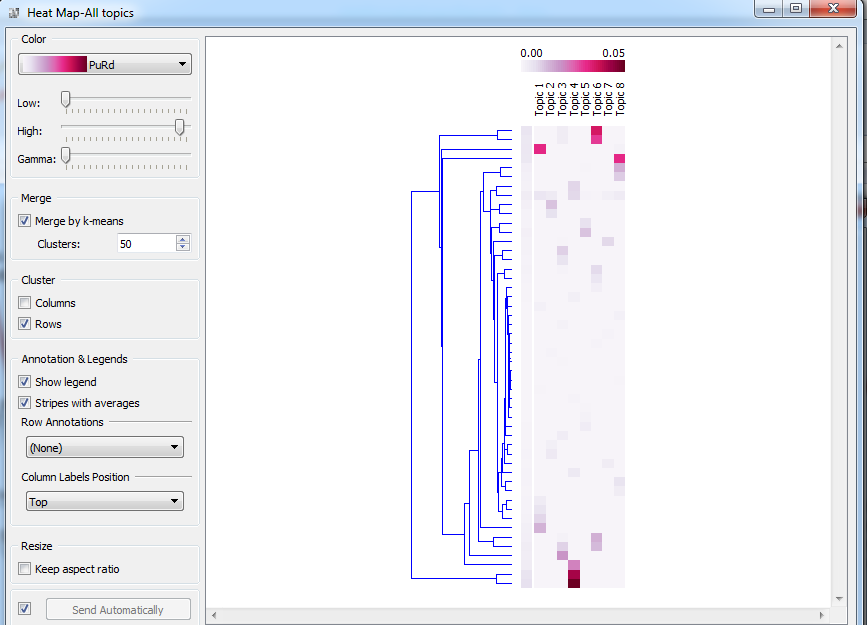
We are using different modelling technique for topic modelling and identifying the similarity between topics. And understanding the relationship between the documents.

**LDA topic modelling**

LDA represents documents as **mixtures of topics** that spit out words with certain probabilities.

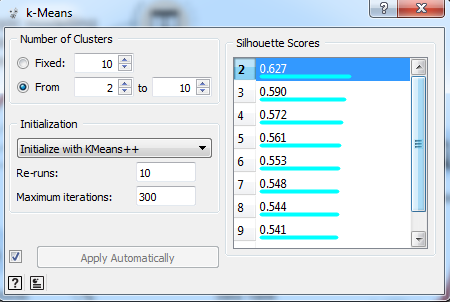


**Various topics and their coverage in corpus**



**K-Means**

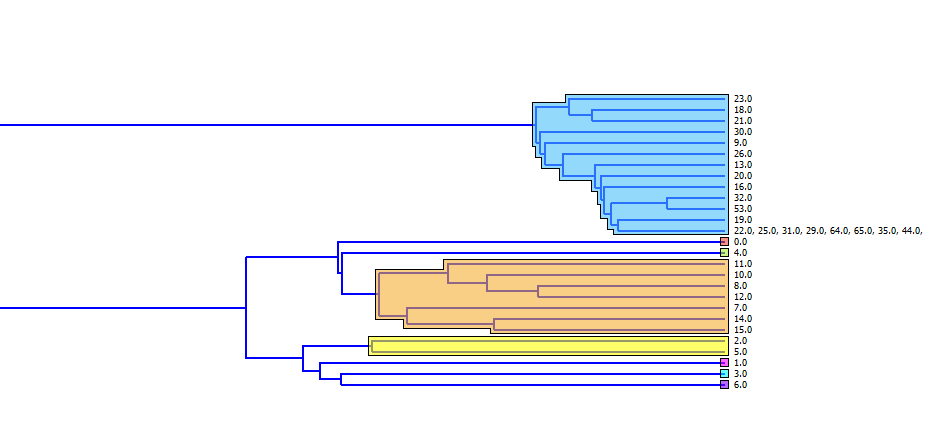
Looking at the silhouette scores we can infer that there are 6 distinctive clusters.

****

**Hierarchical/Agglomerative clustering**

It can be used to find out possible clusters of these news articles and we can use dendrogram to draw benchmark to split corpus of stories in few clusters.

Agglomerative creates one cluster for each document, Join the most similar pair into one cluster and repeat the process until some condition is fulfilled. Similarity in agglomerative can be obtained using ward’s method – combination.



**News as vector in multidimensional space**

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. An MDS algorithm aims to place each object in N-dimensional space such that the between-object distances are preserved as well as possible. Each object is then assigned coordinates in each of the N dimensions. The number of dimensions of an MDS plot N can exceed 2 and is specified a priori.



**Evaluating Clusters**

Looking at all the above clustering methods we can derive that there are 6 major clusters in the given corpus. These are few methods to validate clustering data –

Precision: Given all predicted labels (for given class X) , how many instances were correctly predicted?

Recall: For all instances that should have a label X, how many of these were correctly captured.

F1-score:  This is a weighted average of the true positive rate (recall) and precision.

We can build confusion matrix to showcase true positive, true negative, false positive, false negative & calculate above evaluations.

**Generation of ClickStream Data**

We would be assuming attributes of clicks be gaussian distribution, will be defining range of these values, mean and standard deviation to generate the data using gensample.

**User Profiling and Recommendation system:**

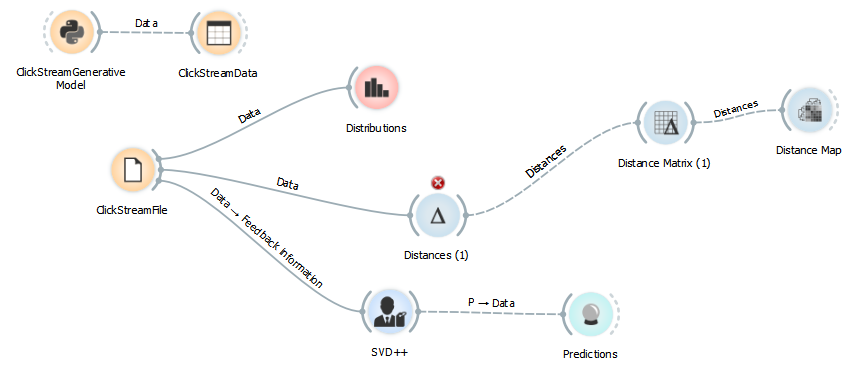
We will be designing recommender system which is based on content-based filtering. Content-based filtering method will be based on a description of the news item and a profile of the user’s preferences. In a content-based recommender system, keywords will be used to describe the news items and a user profile is built to indicate the type of news item this user likes. In other words, the algorithm tries to recommend news items that are similar to those that a user liked in the past or is examining in the present. In particular, various news items will be compared with items previously rated by the user and the best-matching items are recommended.

To abstract the features of the news items in the system, an item presentation algorithm called the tf–idf representation will be used for vector space representation.

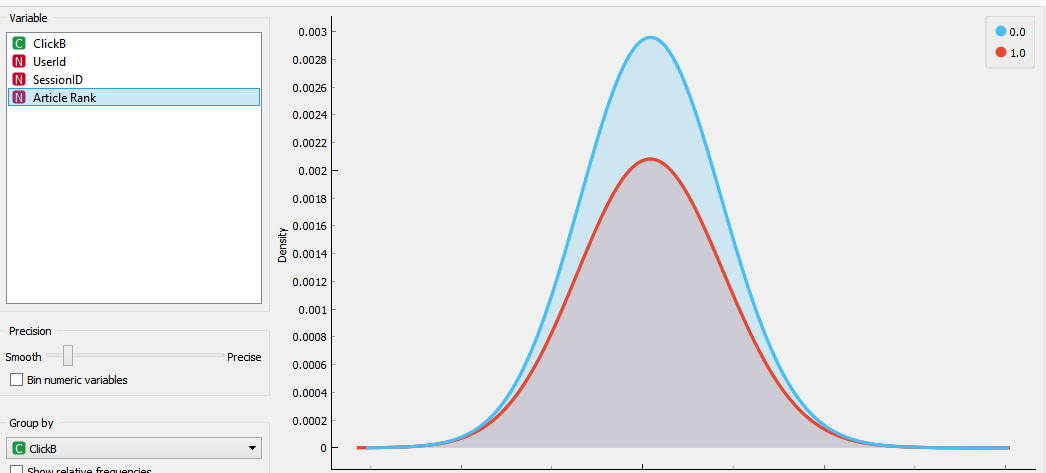
We need to create the content-based profile of users based on a weighted vector of news item features. The weights denote the importance of each news feature to the user and can be computed from individually rated content vectors. If user1 read 5 sports articles and user2 also reads 5 sports articles from given 10 articles, then we can say that the average is 5/10 and so on. Based on this we can also find the near user profile to recommend similar articles.

Beta is (nearest neighbour user profile calculated based on cosine similarity between existing users and beta should keep on decreasing) and Lambda value will keep increasing for each user click.

So if a new user comes, we recommend 10 items out of which 8 item are based on user profile, one on Avg UP and one on Nearest UP



**Distributions of variables in the data sets**



Pearson correlation can be projected among users to identify with similar interests so if new user comes in we can recommend what other users with similar interests clicked.

Strategy:

We plan to build two models one for content based and another is personalized with various technique as showed above in work flow.

**References**

<http://www.casmacat.eu/uploads/Main/SentenceClustering-Chinea.pdf>

<http://stp.lingfil.uu.se/~santinim/ml/UnsupervisedLearningMagnusRosell_Slides.pdf>

<http://www.techniajournal.com/attachments/article/109/Sentence%20Clustering-based%20Summarization.pdf>

<http://www.zora.uzh.ch/id/eprint/56367/1/56367.pdf>

<https://dzone.com/articles/clickstream-analysis-and-data-mining-techniques-10>

<https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35599.pdf>

<https://blog.biolab.si/tag/recommendation/>

Also some inspirations are from the presentations made in classroom