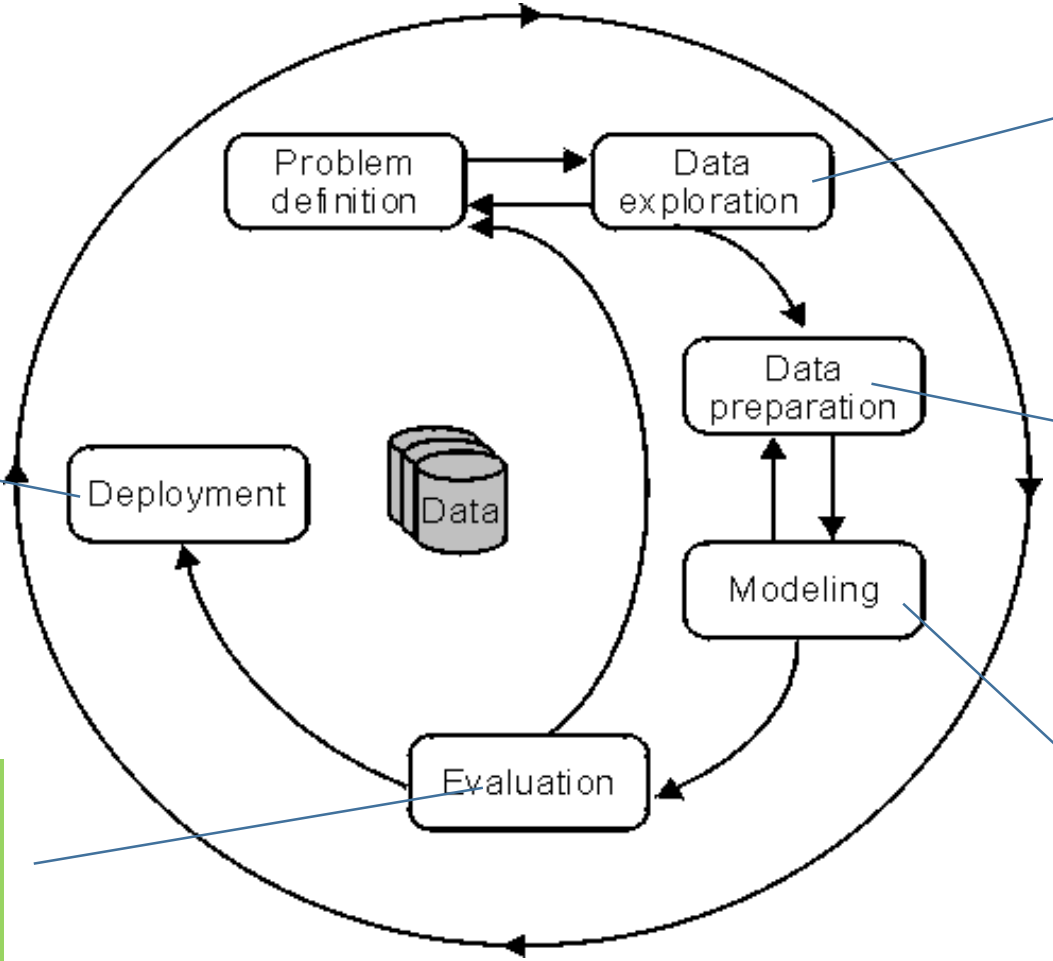
INTRODUCTION TO DATA MINING

By Kelly

Data Mining Process



CRISP DM process model

Descriptive statistics, Missing values, Feature Engineering

Data cleaning, appending, merging, remove inaccuracies, duplicates

Use Data Mining Models to solve your problem

Cross Validation, bias-variance trade-off, Testing the models

Monitoring and Maintenance plan

Data Mining Project - Airbnb

BACKGROUND

Airbnb is an online trusted community marketplace that enable people to discover, book and list unique accommodation around the world via internet.

Objective of the project: To develop a statistical models that could be used to facilitates with the prediction of the price of Airbnb property based on the information available in the Airbnb database.

DATASET

The dataset was obtained from the Airbnb's publicly available data website, Inside Airbnb. There are 16254 instances and 27 attributes. The dataset contains information about the Airbnb listings in Sydney from 2016 to 2017.
Variables captured in the dataset include data such as price, number of amenities, property type, number of beds etc. With these variables, we hope to analyze and predict the price of other Airbnb listings Sydney.

MODEL SELECTION

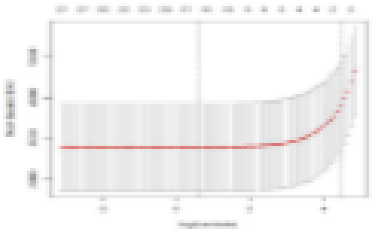
LASSO Regression
Regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces for the prediction on the price of an Airbnb listing.

Random Forest
Random Forest generates plenty of training data sets and tweaks it in such a way that it de-correlates the trees. This reduced variance is by training on different samples of the data.

Neural Network
Neural networks undergoes a learning process that studies the interactions between neurons and the weights applied to each input at each neuron. The weights are constantly adjusted by an activation function to optimise predictions.

FINDING

LASSO Regression



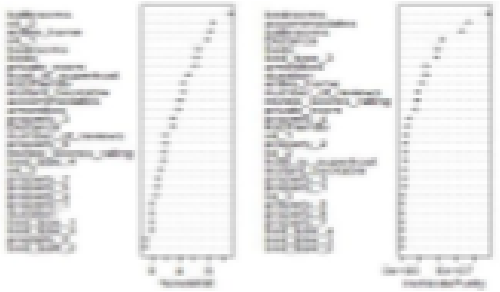
Based on the cross validation of Lasso Regression with different penalizing parameter, λ, Lasso regression model with λ=1.86383 generates the least MSE. The improved model will contain 16 variables .The table above would show the predictor variables that are highly significant to the model.

Test error: 15609.23 (MSE)

Regression Tree

From the diagram, we are able to observed that number of bedrooms of a listing is the most important factor in determining price. The next important factor to consider would be listings with more bathrooms as such listing would tend to fetch a higher price
Test error: 13469 (MSE)

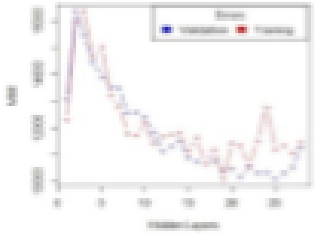
Random Forest



From the diagram, it shows that the number of bedrooms is one of the most important factors to the Gini index, while the number of bathrooms is an important variable in reducing MSE

Out of bag error: 24694.7
Test error: 12568.11 (MSE)

Neural Network



The optimal number of neurons in the hidden layer is 19 after cross validating of neural networks with different number of hidden layers. The diagram above is the plot of MSE against number of hidden layers. The neural network with 19 neurons is found to have a MSE of

DISCUSSION

The different models are best suited for different situations . Based on our findings, random forest has the lowest test error, and hence the most accurate model for our problem. However, the other models have their advantages . Lasso regression is preferred model to use if were to make future numerical predictions . Regression tree, on the other hand, has the best visualization of the data, and is best used to explain to non-experts . Neural network, however, could be further trained and has the potential to make even more accurate predictions as compared to the other models .

Types of Data Mining Models

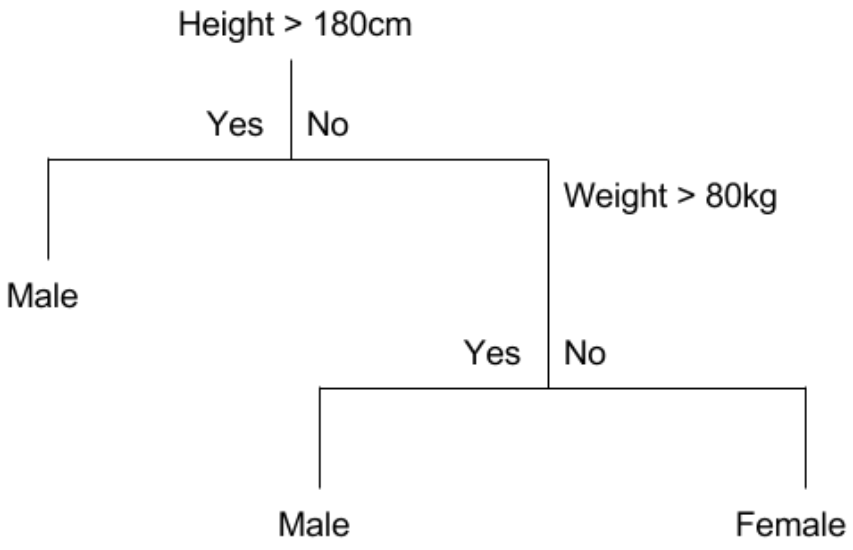
Supervised	Unsupervised
Regression: Linear, Multiple, Regularised, LASSO, Ridge, Weighted, Logistic	K means Clustering
Neural Network: Forward propagation, Backward propagation	Market basket analysis
Classification: Regression Tree, Random Forest	Principal Component analysis
Xgboost	
Support vector machines	

LASSO Regression

Lasso regression is a type of **linear regression** that uses shrinkage. The lasso procedure encourages simple, sparse models. This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

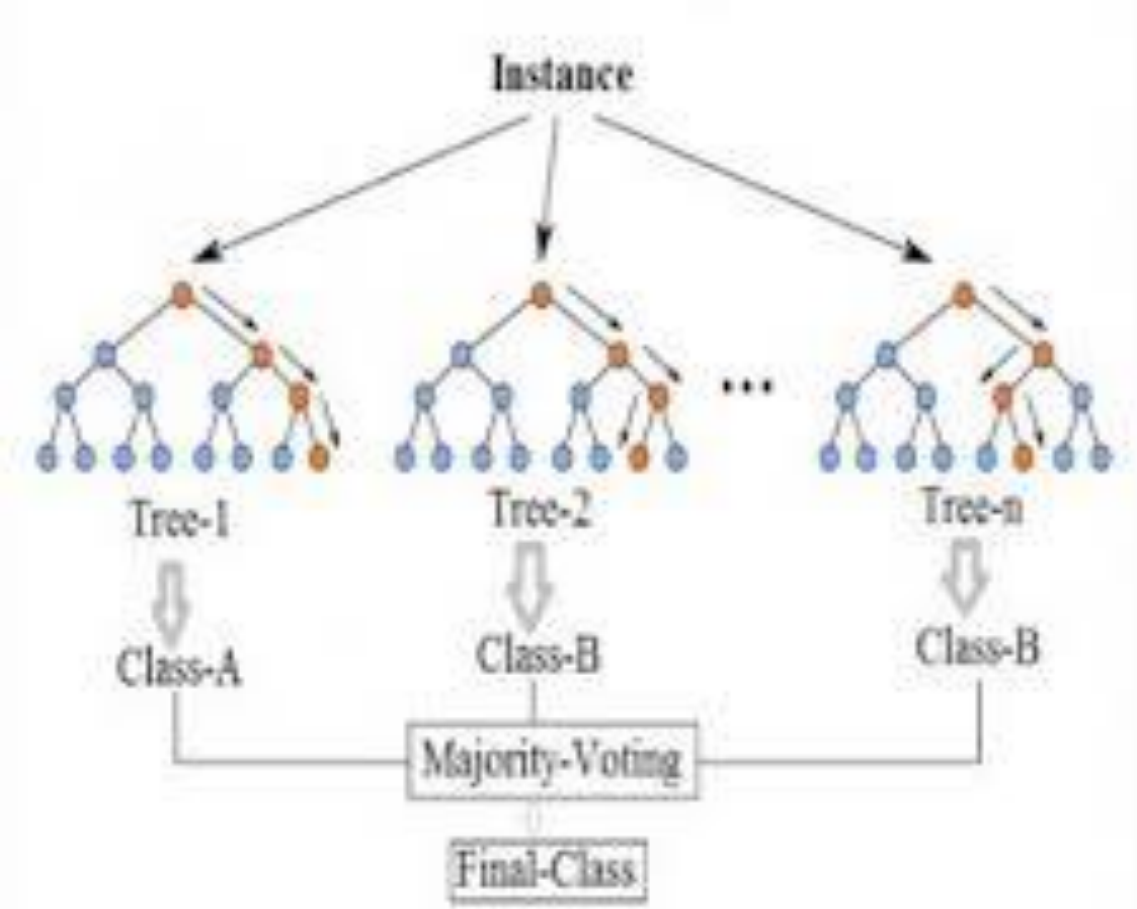
Cost Function - $\min\left(\left|\left|Y - X\theta\right|\right|_2^2 + \lambda||\theta||_1\right)$

Regression Tree



- Trees divide the predictor space into distinct and non-overlapping regions
- The decision of making strategic splits heavily affects a tree's accuracy.
    - Gini Index
    - Chi Square
- Stopping Criteria – e.g. stop once the number of observations per node becomes < 50.
- Pruning - remove irrelevant nodes

Random Forest



- Create a “forest” of decision trees
- Determine the prediction made by each decision tree
- Majority vote will be used to obtain the final decision

Cross Validation

- K-fold cross validation:

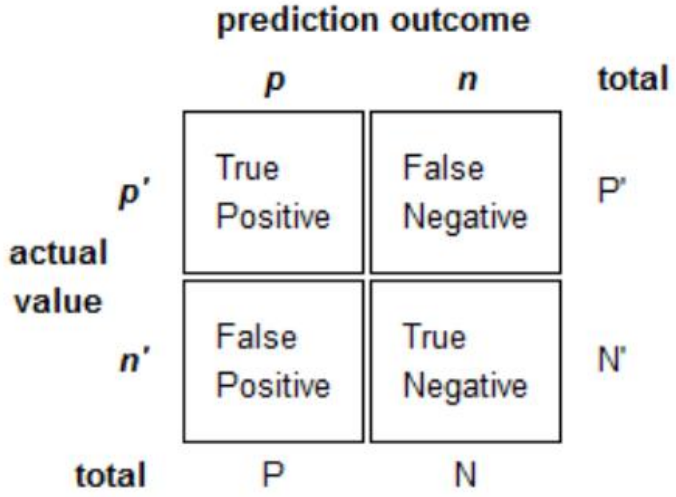
  Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

  - **This is a common practice**

Leave-out-one cross validation: when K=N (total number or data points)

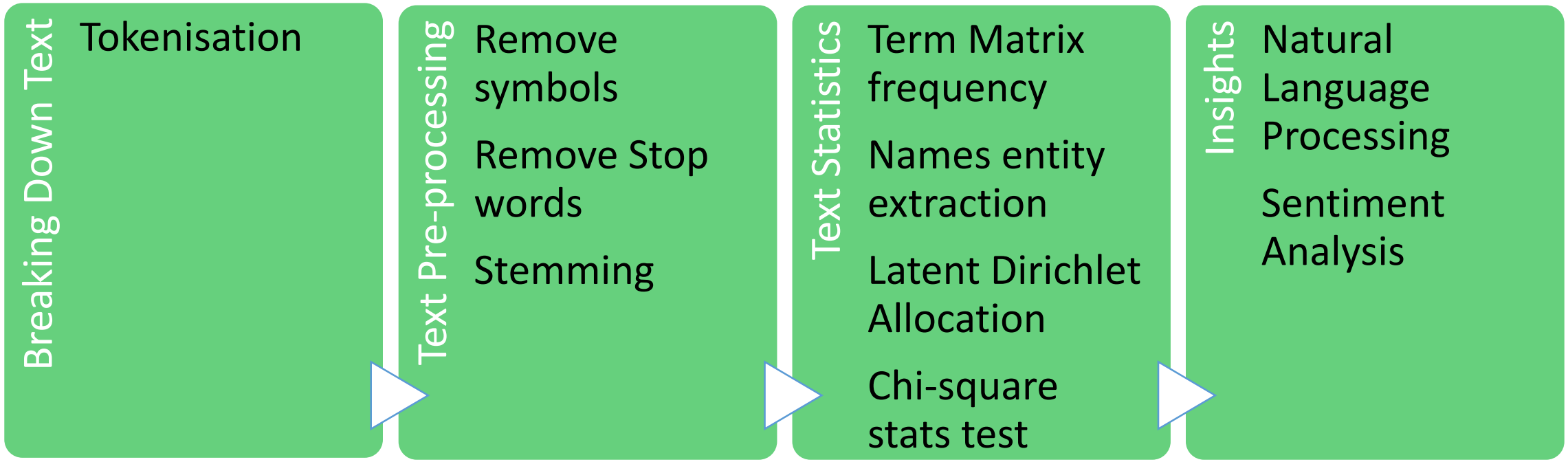
Data Model Evaluation

- Mean square error $MSE(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta} - \theta_0)^2 \right]$
- Residual sum of squares : $\sum(y_i - \bar{y})^2 = \sum(\bar{y}_i - \bar{y})^2 + \sum(y_i - \bar{y}_i)^2$
- Confusion Matrix – Categorical predictor

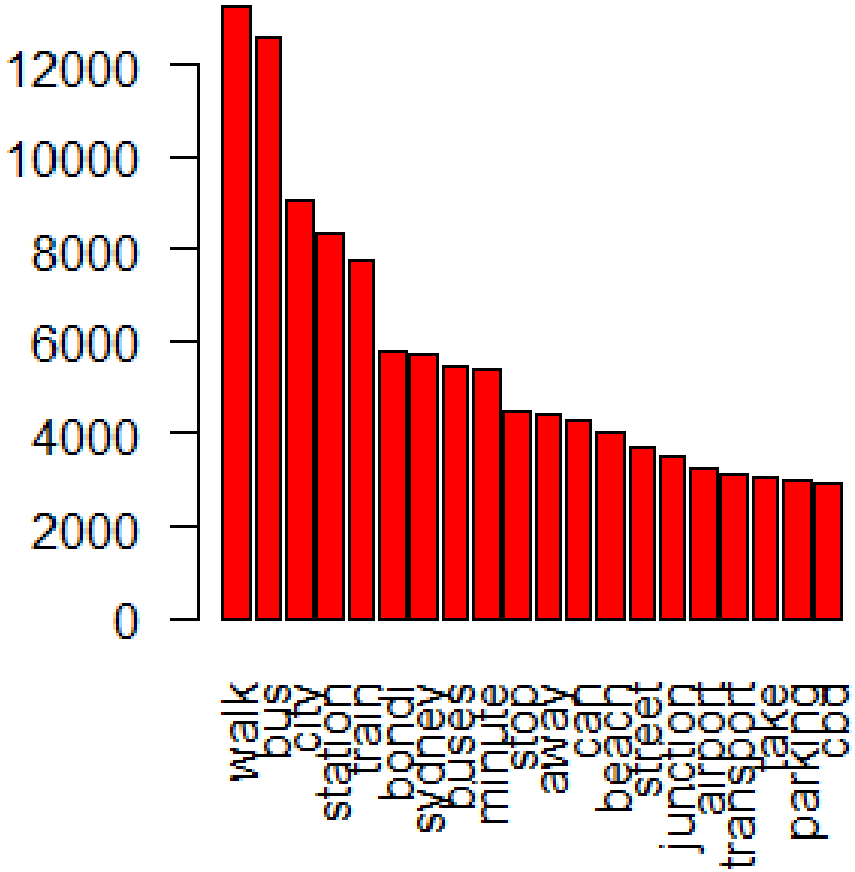


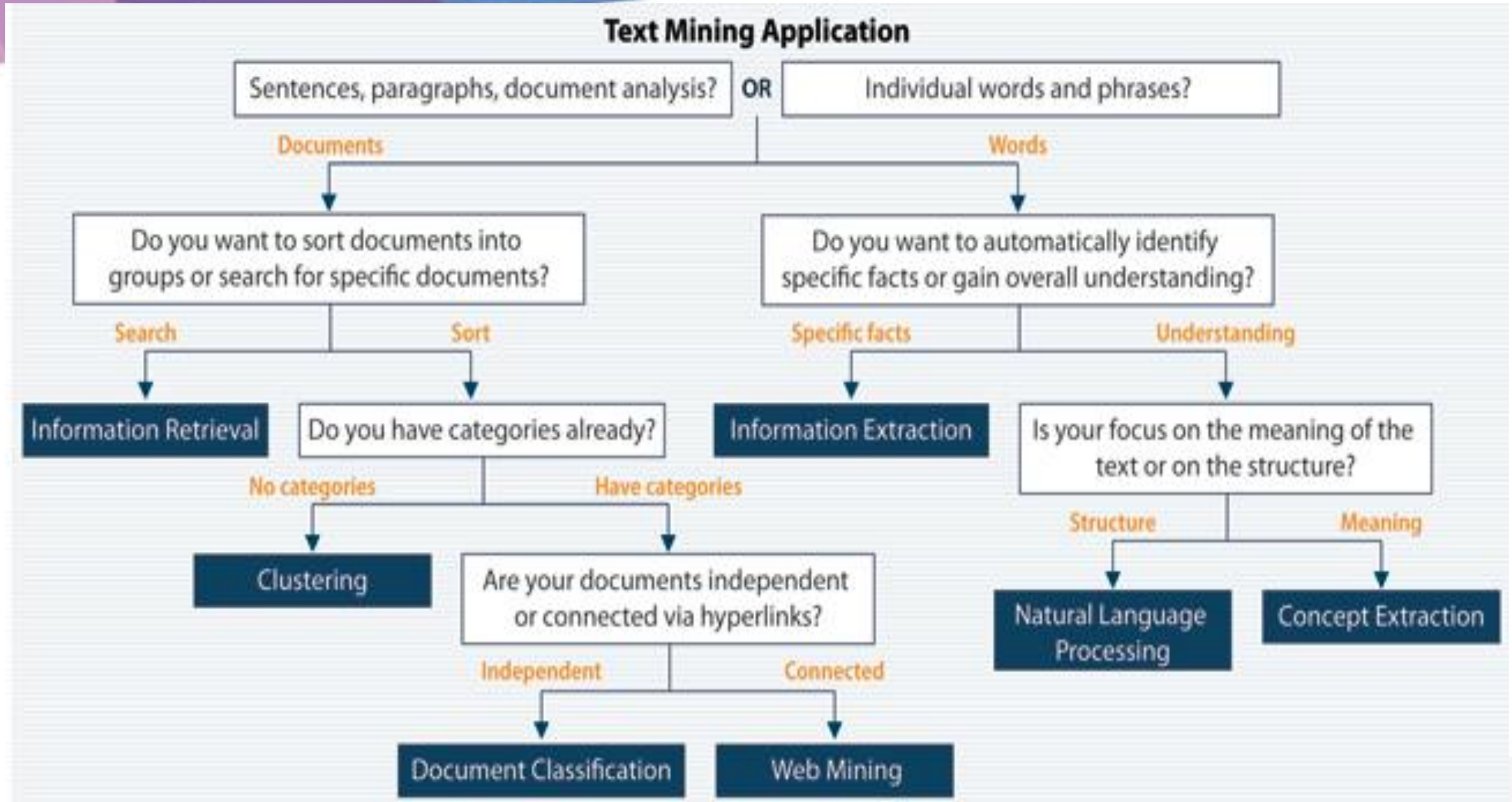
Text mining

Unstructured data: Images, Text, Sounds, HTML, pdf etc.



My Attempt on text mining





http://analytics-magazine.org/text-analytics-deriving-value-from-tweets-blog-posts-and-call-center-logs/

Suggestions

- Look at SR records > group them based on similarity -Jaccard similarity> analysis (e.g. common problems faced by user, did number of problems increase/decrease with the introduction of digital tokens…)
- Create SR buckets that are better representation of the underlying issues : Clustering and automating classification

  Example: *Cluster 4 words: insurances, at, am, at, doing, want*

  - Cluster 4 includes general conversations regarding insurances

- Customer segmentation: more personalised service, identify interesting facets that we might be able to exploit
- Track Customer sentiment over time to track performance
- Identify key promoter and detractors - certain aspect of a product may be hurting the KPI

Text clustering

- Transformations on raw stream of free flow text
- Creation of Term Document Matrix
- TF-IDF (Term Frequency – Inverse Document Frequency) Normalization
- K-Means Clustering using Euclidean Distances
- Auto-Tagging based on Cluster Centers
- https://www.kdnuggets.com/2017/06/text-clustering-unstructured-data.html


The End