Case Study 2

CAB 330

Jack Farrel – N8640866

Anton Polison – N9907637

Dennis Tran – N9937072

2017

Table of Contents

[Task 1: Descriptive Data Mining - Clustering. 2](#_Toc496363742)

[Task 2: Descriptive Mining – Association 8](#_Toc496363743)

[Task 3: Text Mining 11](#_Toc496363744)

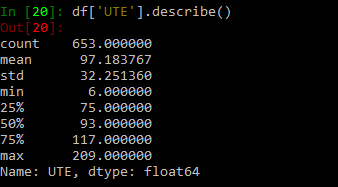
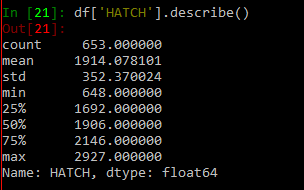
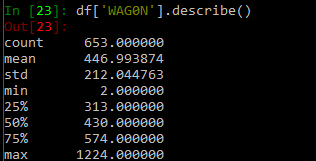
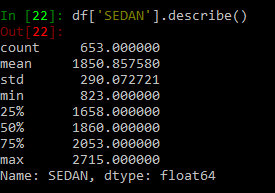
[Task 4: Web Mining 14](#_Toc496363745)

# Task 1: Descriptive Data Mining - Clustering.

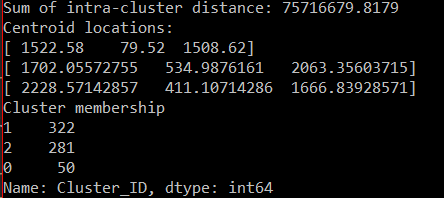
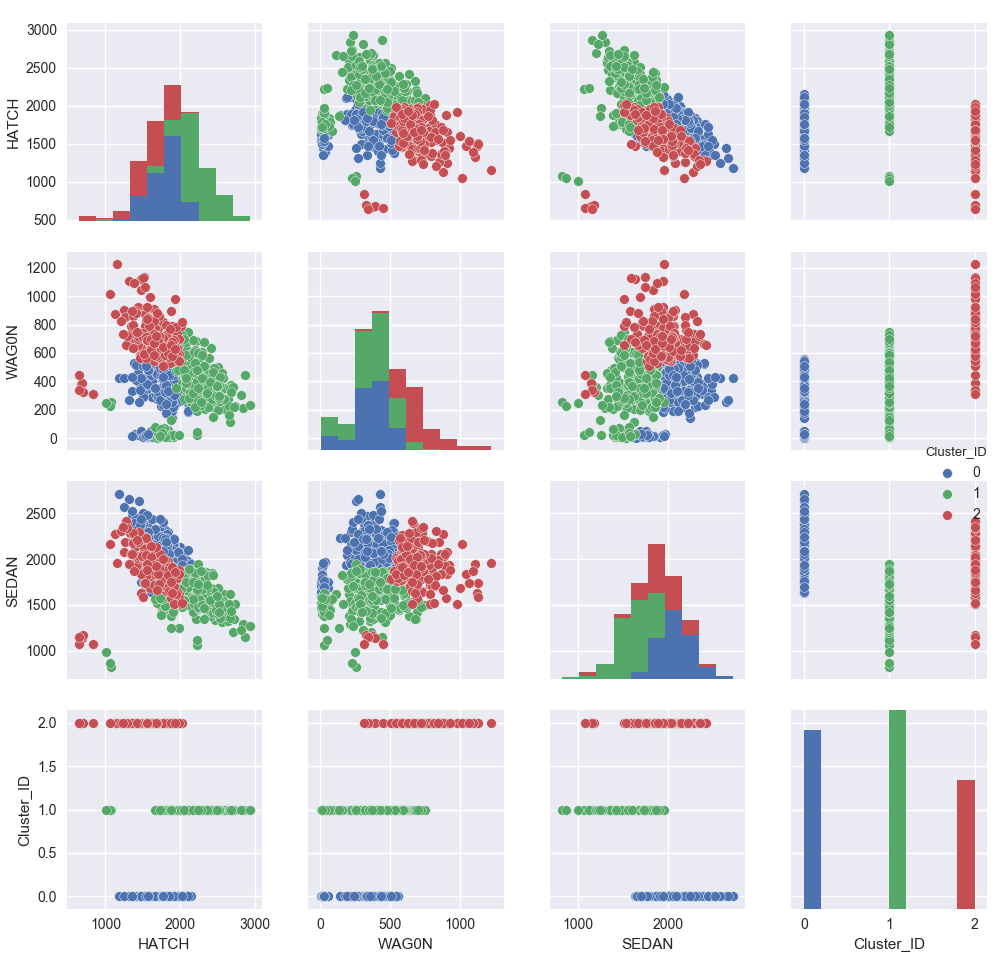
1. There were some changes made to the original dataset to ensure that the information from the analysis was accurate. There are several quality problems with “model\_car\_sales.csv” data set such as missing values and redundant or irrelevant variables
   1. ***Missing Values:*** The dataset contains 22 rows which have empty values. These values were deleted as they are irrelevant for our research.
   2. ***Dropped Variables:*** In total 4 variables were dropped due to various reasons. The below table shows the dropped variables and their corresponding reasons.

|  |  |
| --- | --- |
| *Variable* | *Reason For Drop* |
| *REPORT\_DATE* | Irrelevant. The variable has the date when the report was generated and doesn’t provide any useful information. |
| *DEALER\_CODE* | Redundancy. “LOCATION” variable provides more organised information which is at the same time easier to work with as it is numerical variable. |
| *K\_SALES\_TOT* | Irrelevant. The company wants to find the minimum number of product sale segments and it is not interested in identifying what is a number of total car sales. |

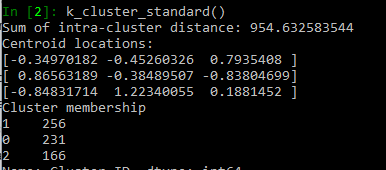
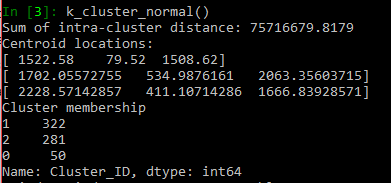
1. After preprocessing and analyzing the dataset it was identified that “UTE” is clearly underperforming in sales in comparison with “HATCH”, “SEDAN” and “WAGON”.

Form the tables above we can determine that “mean” of the variables “HATCH”, “WAGON”, “SEDAN” is significantly higher than of “UTE” variable.

1. “HATCH”, “SEDAN”, “WAGON” and “LOCATION” variables were included in the analyses. These were the variables for the three best performing cars and the locations of the dealerships. These variables were chosen as they best fit the task of determining the optimum number of sale segments. By looking at sales data per location, the analysis can determine the best results.  
     
     
     
   1.   
      From cluster membership table we can identify that 231 records were assigned to **cluster 0**, 256 records to **cluster 1** and 166 records to cluster 2.
   2. The pairplot shows the sales distribution on different variables.

* If we look at the first row and first column we can notice that **cluster 0** has average to high sales. cluster 2 has low to average sales and finally, **cluster 1** has high sales.
* Wagon variable distribution is the following: **cluster 0** has low to average sales, cluster 2 has average to high sales and **cluster 1** has low to average sales with a slight win over **cluster 0.**
* SEDAN variable has the following distribution: **cluster 0** has high sales, **cluster 1** has low to average sales and cluster 2 average to high sales.

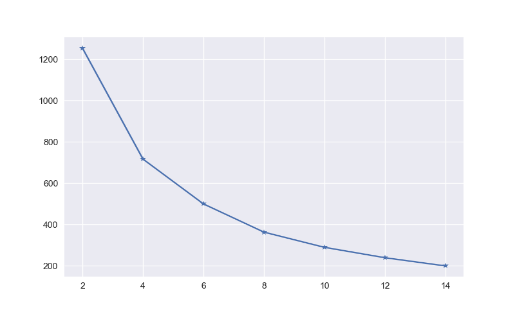
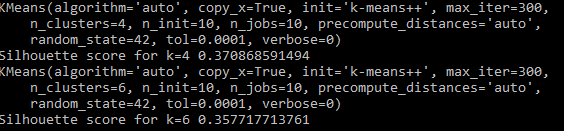
 

Normal distribution Standardization

The standardization made dramatic changes to our model. First and most obvious change is that intra-cluster distance was significantly reduced by standardization which is one of the main purpose of clustering. Furthermore, cluster membership changed as well “cluster 0” 50 =>231, “cluster 1” 322=>256, “cluster 2” 281 => 166. This provides a more “even” distribution.

If we look at the “pairplots” we can see 2 absolutely different results due to increased “cluster 0” membership and changed intra-cluster distance.

As we can see, all the above-mentioned facts proving the fact that standardization brings better clustering solution and better visualization outcome.

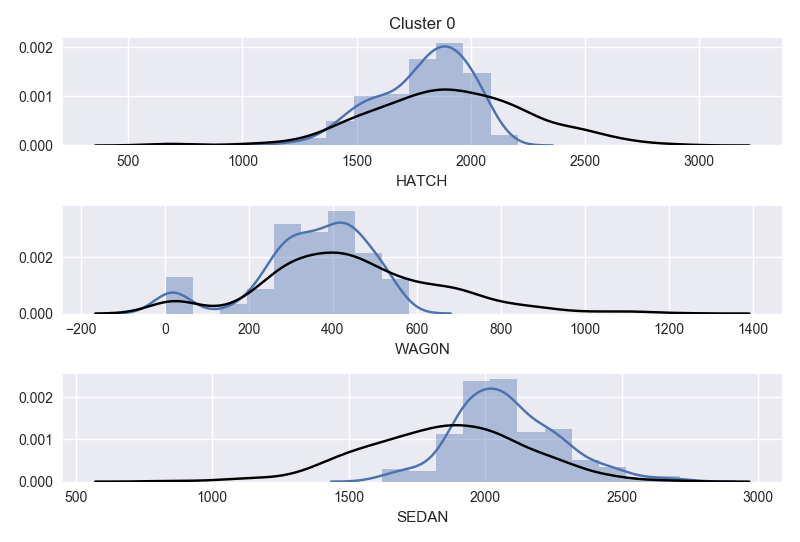
  “Elbow Method” “Silhouette Method”

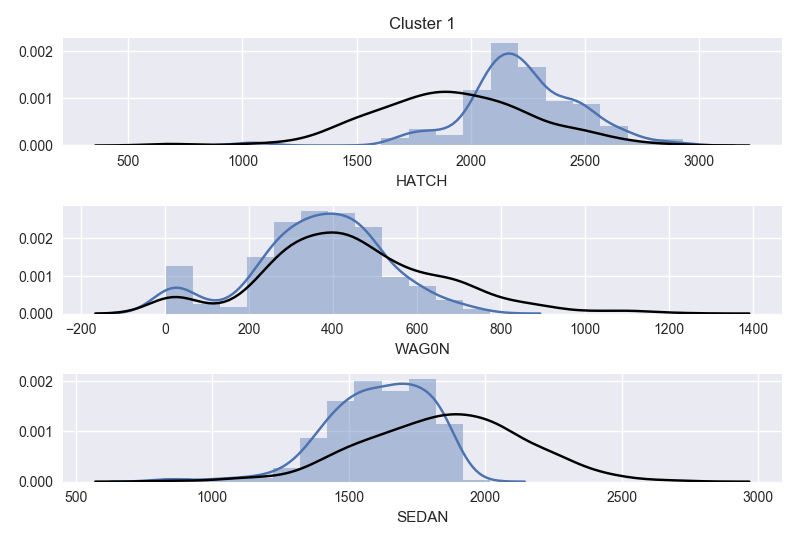
First, we look at the elbow method of determining better K. The graph shows that our best K is between 4 and 6. Though, this method doesn’t always provide the best K.

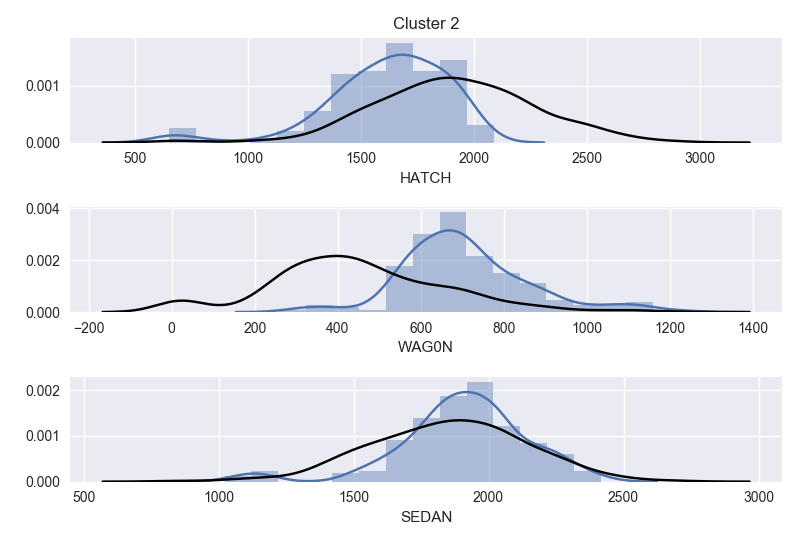
“Silhouette Method” is a more precise technique in identifying best K. It measures how close each point in a cluster is to the point in the neighborhoodlike cluster. Silhouette value lie in the range [-1, 1]. One of the main goals of clustering is to minimise inter-cluster similarity. Therefore, higher the value better the cluster configuration. In our case, k=4 has a better inter-cluster similarity (~0.3709) in comparison with k=6 which has (~0.3577). Based on this method we can state that k=4 is the best model.

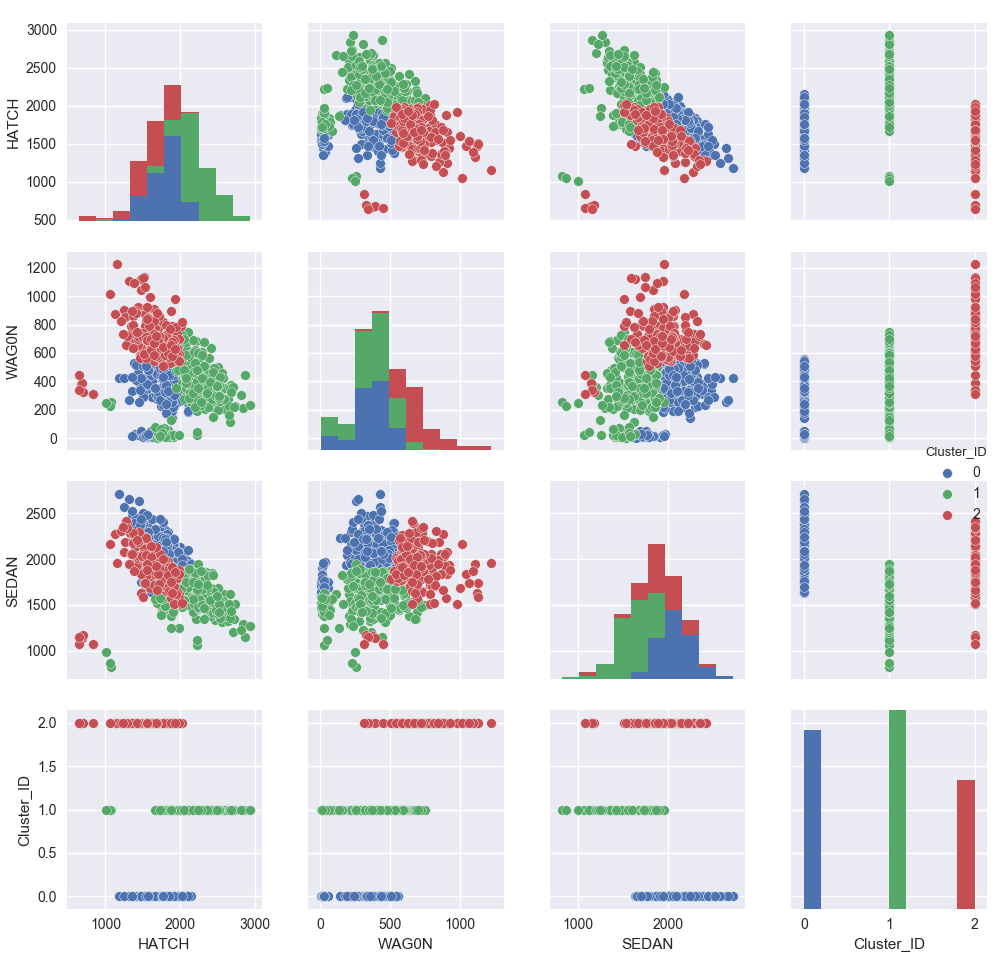
However, even though K=4 is the best model for this task we will use K=3 in our research as it’s easier to interpret and provide more distribution understandably.











After careful examination of all distplots and pairplot with K=3 we can sum up our findings.

**Cluster 0**: has mostly average sales for HATCH. At the same time WAGON has low to medium sales. On the other hand, SEDAN has average to high sales.

**Cluster 1**: showing that sales for HATCH is medium to high, WAGON has low to medium sales and SEDAN is sold on an average level.

**Cluster 2**: sales on HATCH are from low to medium, WAGON has medium to high sales and SEDAN has medium to high sales.

# Task 2: Descriptive Mining – Association

|  |  |
| --- | --- |
| *Variable* | *Importance* |
| *Location* | Irrelevant. The variable has no use to our research |
| *Transactin\_Date* | Irrelevant. The transaction date’s values are similar and of no use. |
| *Quantity* | Irrelevant. Each row of Quantity variable has a value of “1”. |

1. “pos\_transaction.csv” doesn’t have obvious erroneous or missing values. Some of the variables were dropped due to irrelevancy in our research.

1. “Transaction\_Id” and “Product\_Name” were the 2 most valuable variables which we used to perform association mining. The “Transaction\_Id” variable includes the transaction ids which group the products under the particular transaction. The “Product\_Name” attribute has the list of products. By seeing which products were bought together during transactions, association rules can be determined. Thus, these to variables are chosen
   1. The table below shows the rules sorted according to lift in descending order. It can be seen that two rules share the highest lift value. These rules are “**Perfume => Toothbrush**” and its mirror “**Toothbrush => Perfume**” with a lift value of 3.601370. This followed by another pair in “**Toothbrush =>Bow**” and “**Bow => Toothbrush**” with lift level of 3.081236  
      
   2. The Table below shows the rules sorted by confidence in descending order. From the table it can be seen that the rule with the highest confidence is “**Toothpaste, Pencils=> Candy Bar**” with a confidence of 0.463762. This is closely followed by the rule “**, Magazine, Greeting Cards => Candy Bar**” with a confidence of 0.458649  
      
   3. The full pairplot based on the number of items in a rule comparing support, confidence and lift can be found in appendix A. When looking at the plot many things can be seen.
      1. When looking at the Items Vs. Support plot we can see that 1 item rules have a varied support, with the most supported falling in this category. The plot also shows that as more items are added to the rule the lower the support. This relationship also works in inverse. This makes sense as 1 item is more likely to occur than two or three. The more items to a rule the less spread the support index is
      2. Looking at the Items Vs. Confidence plot we can see that the lower the number of items in a rule, the confidence tends to be lower. This relationship also works in inverse.
      3. Looking at the Items Vs. Lift plot we can say that the more items to a rule the higher a lift value is likely to be (excluding outliers). This relationship also works in inverse. It is also clear that all 1 item rules have a lift of 1.
      4. Looking at Lift Vs. Confidence we can see that an increase in lift value tends to correspond with an increase in confidence.

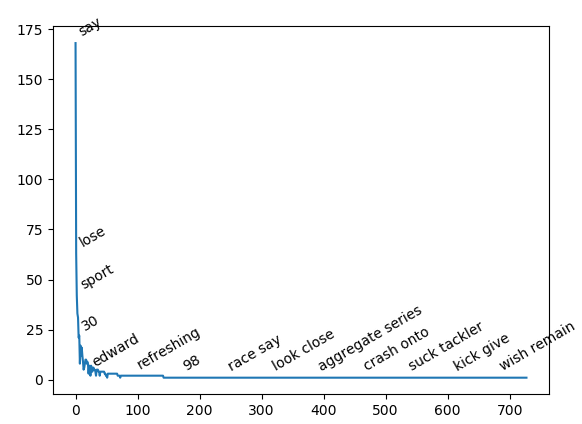
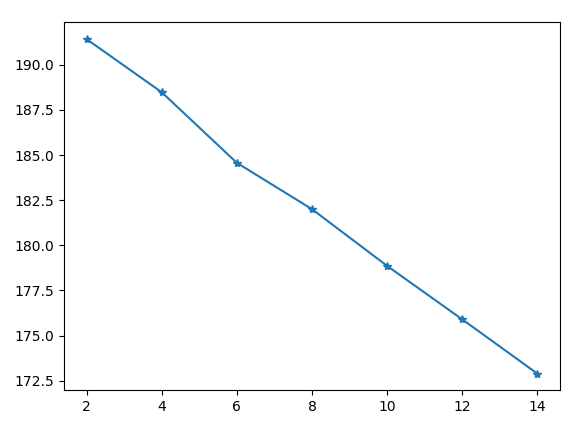
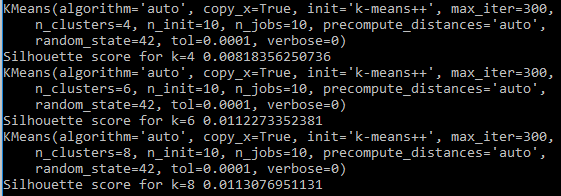
From these observations we can make a statement such as: The more items to a rule, the more likely the lift and confidence is to be higher than corresponding rules, and the lower the support will be.

* 1. By printing the full results list to a CSV file, it can be seen that there a total of 15 rules in which a candy bar is purchased. This includes groups such as “**Toothpaste, Candy Bar”.** When not considering groups, there are 5 rules which involve the purchase of a candy bar. All the rules involving candy bars can be seen in the table below sorted in descending order by lift value:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Left\_side | Right\_side | Support | Confidence | Lift |
| 52 | Candy Bar,Magazine | Greeting Cards | 0.016665 | 0.411126 | 2.798966 |
| 55 | Toothpaste,Candy Bar | Greeting Cards | 0.013175 | 0.331197 | 2.254802 |
| 57 | Candy Bar,Magazine | Pencils | 0.012005 | 0.296164 | 2.195025 |
| 63 | Candy Bar,Pencils | Toothpaste | 0.01139 | 0.344995 | 2.150505 |
| 64 | Toothpaste,Candy Bar | Pencils | 0.01139 | 0.286325 | 2.122103 |
| 60 | Candy Bar,Magazine | Toothpaste | 0.01372 | 0.338473 | 2.109851 |
| 54 | Greeting Cards,Candy Bar | Toothpaste | 0.013175 | 0.301764 | 1.881026 |
| 17 | Candy Bar | Greeting Cards | 0.04366 | 0.255314 | 1.738191 |
| 51 | Greeting Cards,Candy Bar | Magazine | 0.016665 | 0.381699 | 1.581813 |
| 58 | Candy Bar,Pencils | Magazine | 0.012005 | 0.363623 | 1.5069 |
| 25 | Candy Bar | Toothpaste | 0.03978 | 0.232625 | 1.450053 |
| 21 | Candy Bar | Pencils | 0.033015 | 0.193065 | 1.430903 |
| 61 | Toothpaste,Candy Bar | Magazine | 0.01372 | 0.344897 | 1.429299 |
| 19 | Candy Bar | Magazine | 0.040535 | 0.23704 | 0.982325 |
| 23 | Candy Bar | Pens | 0.012265 | 0.071723 | 0.499551 |

* 1. If we look at the above table we can see the items people are likely to also purchase alongside a candy bar. When looking at combinations (the top 6 lift values) we can see that “**Greeting Cards”**, **“Pencils”**, and **“Toothpaste”** are twice as likely to be bought with a candy bar involved than when picking a random transaction as all have a lift value of higher than 2. Also **“Magazines”** are 1.5 times more likely with a lift of 1.5069 and 1.429299 depending on the combination.  
     When looking at purely candy bar purchases **“Greeting Cards”** have a high lift value of 1.738191. **“Pencils”** and **“Toothpaste”**  are also bought with candy bars with a lift of over 1.4.

# Task 3: Text Mining

1. The **TEXT** variable was the only variable used for the text mining analysis. This variable was chosen as it was the only one relevant to the information trying to be gathered. The goal is to determine patterns in text to allow for groupings of similar articles. With that in mind the **TEXT** variable has all the required information, and thus it was chosen for the analysis
2. From the ZIPF plot, which can be seen below, ten terms can be seen which would not be useful for clustering  
     
   These ten terms are:  
   - wish remain  
   - kick give  
   - suck tackler  
   - crash onto  
   - aggregate series  
   - look close  
   - race say  
   - 98  
   - refreshing  
   - say  
   The first 9 terms listed were chosen due to low frequency occurrences (less than 2), making these not useful to the analysis. “say” was chosen due to its high occurrence. The phrase appears too many times to be useful in clustering
3. “Say” was disregarded as it occurred in nearly 90% of documents. This high frequency occurrence makes it non-beneficial to clustering as it could occur in all clusters. The next highest frequency occurred in less than half the documents, therefore it was decided to keep this term.
4. The feature weightings where as follows:  
   **min\_df** = 2,  
   **max\_df** = 0.8  
   **min\_df** was set as 2 as the lowest occurring terms appeared in less than 2 documents. This eliminates any phrases occurring once. While **max\_df** was set as 0.8 to ensure the highest occurring phrase was eliminated. This means that any occurrence of over 80% will be unused, as “say” was nearly 90% it will be eliminated.
5. After feature selection was performed based on the above information, the number of unique tokens available for clustering is 6917.
6. The graph below shows the output of the elbow method:   
    From the graph it is hard to determine the best k value due to its smoothness. However, it looks as though it falls somewhere between 4-8. With this information the silhouette method was applied todetermine the best k value, with the results seen below:   
     
     
   From the data it can be seen that the best K value would be 8. It has a slightly higher silhouette value than 6. Therefore the number of clusters was 6 and they can be seen below with appropriate names,

|  |  |
| --- | --- |
| **Cluster 0 (Countries)** | england, ireland, wale, nation, scotland, |
| **Cluster 1 (Club News)** | club, want, player, season, manager, |
| **Cluster 2 (Australian Open)** | open, australian, australian open, hewitt, year, |
| **Cluster 3 (Tennis)** | 6, roddick, 7, moya, spain, |
| **Cluster 4 (Athletics)** | world, race, year, athletics, athens, |
| **Cluster 5 (Cricket)** | pakistan, sri, test, wicket, cricket, |
| **Cluster 6 (Cricket)** | south, south africa, africa, england, vaughan, |
| **Cluster 7 (Football)** | chelsea, arsenal, henry, newcastle, real, |

The name Olympics was chosen for cluster 6 based on the countries listed and the fact Michael Vaughan is an ex England cricketer. As can be seen two clusters have been named cricket, meaning there may be some room for refinement to reduce inter-cluster similarities,

1. By applying LSA/SVD the clustering is based on document concept similarity, rather than just word similarity. This can allow for a more accurate clustering, as the documents are similar. This is done by reducing the components significantly, in this case to 100. This helps to significantly reduce the time taken for clustering when compared to the previous method. Below is a table showing the new results of clusters and there possible names:

|  |  |
| --- | --- |
| **Cluster 0 (Rugby)** | kafer, saracen, club, head coach, sinderberry, |
| **Cluster 1 (Cricket)** | cricket, play, england, player, year, |
| **Cluster 2 (Drug Bans in Athletics)** | athletics, ban, drug, athens, year, |
| **Cluster 3 (Tennis)** | 6, roddick, 7, moya, open, |
| **Cluster 4 (Liverpool News)** | club, liverpool, want, play, game, |
| **Cluster 5 (Rugby)** | england, ireland, kick, goal, hodgson,, |
| **Cluster 6 (Olympics)** | indoor, world, olympic, record, birmingham, |
| **Cluster 7 (English Cricket)** | ball, vaughan, wicket, day, england, |

Cluster 0 was named as Rugby as the cluster lists Rugby players and clubs. Cluster 5 was named rugby due to Hodgson being a rugby player and references to kicking goals. Cluster 7 was named English cricket as Vaughan was an ex England international player.  
The clusters from SVD appear to give a better grouping. The terms within the clusters appear to have more relevance to each other, with the exception of cluster three. It can be seen that the clusters from this method do share some similarities, such as tennis and cricket remaining essentially the same. Other clusters such as the Six Nation Cup cluster have been given more relevant terms, making them more recognisable. Six Nation Cup was previously countries.

1. These clusters can be useful to the planned goal of an online personalised news service. With this information articles of a similar nature can be grouped together. For example, if a user was to read an article on a recent Premier League game, the service could then offer up other articles that have been classified in the same cluster. This would offer the user articles targeted at their particular interest.  
   The information could be used in a more direct way by asking the reader to select what types of news they would like to see. Once selected news from clusters relating to their choice can be delivered to them, meaning they obtain only the news they want.

# Task 4: Web Mining

1. Association Mining was chosen as the web mining technique to use for this task. Association mining is used for establishing links between variables, based on commonality, from a set of records within a data set. The goal of this task is to determine user browsing patterns to ensure the most efficient website design. With this in mind, establishing links between pages of the site visited based on sessions would be effective in determining browsing patterns.   
     
   We want to discover what pages are likely to be visited given another page is visited first. This type of analysis can be seen in the association rule (item set A) => (item set B). In this case, item set is web page. With this information the association technique was chosen.
2. For the task, the variables chosen were **user\_id** and **request**.   
   **user\_id:-** provides the information for the session. This variable is the same as the session variable, thus either could have been used. The reason this variable was chosen of IP is due to the fact an IP could be obtain via a proxy, making this information unreliable. As well as this, the user\_id only provides information on a session, where as the same IP could be used for multiple sessions. This could skew the data as a session on a different day could have no relevance to a previous session. Therefore it was decided that individual session data should be used to ensure browsing habits are determined more effectively.  
   **request:-** provides information on links clicked within the webpage. This variable was chosen as requests can be catalogued per session. This information can be used to determine the likelihood of certain links being requested if another link has been requested in a given session. This is the only variable that provides information on the pages visited on the website, making it important to determining browsing behaviour.  
   For the apriori function, the arguments ***transaction***and ***min\_support***were used.   
   ***transaction:-***this is a list of the transactions (in this case, requestions made). It is a list of lists of requests. The data used here was the request lists  
   ***min\_support:-*** This argument is to determine the minimum support of the relationship. This was set as 0.04, as it was found to give the best results.
3. PLACE HOLDER

# Appendix A

