Mining Glasgow Norms

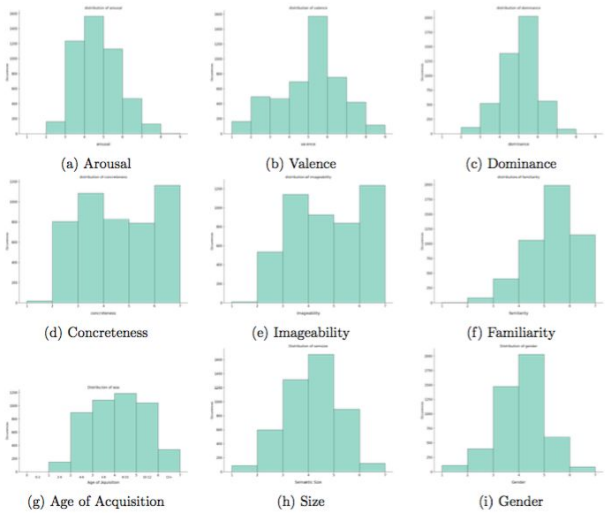
Presentation of the report by Giulio Cordova

Data Understanding & Preparation

Data Semantics & Distribution

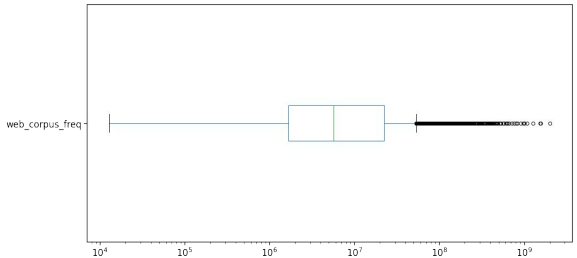
Set of ratings of 4682 English Words based on 12 different numerical features

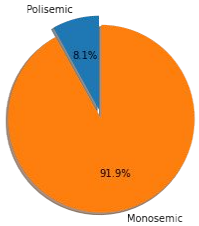
Glasgow Norms per se (continuous)



Other attributes in dataset (discrete)

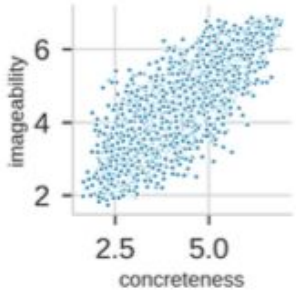
- Length
- Polysemy
- Web Corpus Frequency





Assessing Data Quality & Variable Transformation

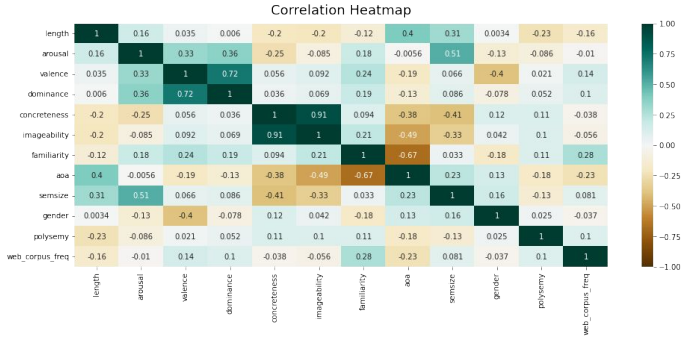
- **Missing Values:** 14 only in Web Corpus Freq, substituted with the mean
- **Outliers:** detected with the boxplots
- **Errors** (syntactic and semantic) not detected
- **Normalization:** features normalized between 0 and 1



- **Gender:** renamed to masculinity
- **Web Corpus Freq:** discretized by taking the logarithm and flooring the result

$$k' = \lfloor \log_{10}(k) \rfloor$$

- **Perceivability:** new variable mediating *concreteness* and *imageability*



Clustering

Preprocessing

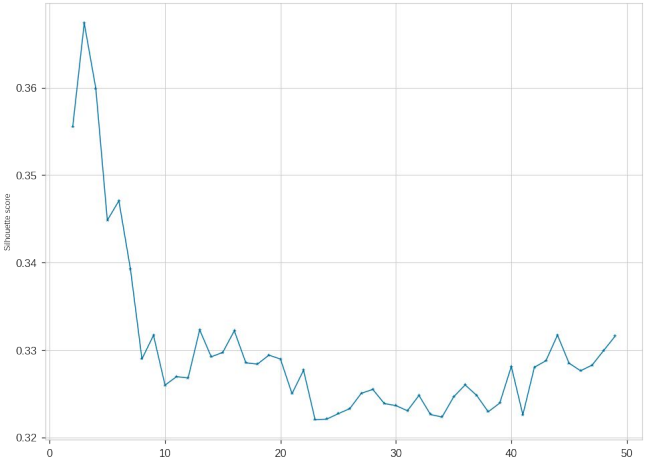
Transformed dataset as exposed in previous section

	length	arousal	valence	dominance	familiarity	aoa	semsize	masculinity	polysemy	web_corpus_freq	perceivability
count	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000	4682.000000
mean	0.310597	0.428289	0.532598	0.482728	0.684871	0.508419	0.498718	0.519165	0.080948	0.457027	0.554749
std	0.143302	0.179275	0.209314	0.144739	0.174077	0.217797	0.184810	0.152787	0.272785	0.168797	0.266786
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.214286	0.292810	0.405015	0.402488	0.578042	0.329451	0.372584	0.436443	0.000000	0.400000	0.313092
50%	0.285714	0.410784	0.559275	0.494868	0.716364	0.514256	0.507766	0.522693	0.000000	0.400000	0.538709
75%	0.428571	0.549346	0.664041	0.569051	0.816704	0.683762	0.633375	0.612293	0.000000	0.600000	0.813629
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

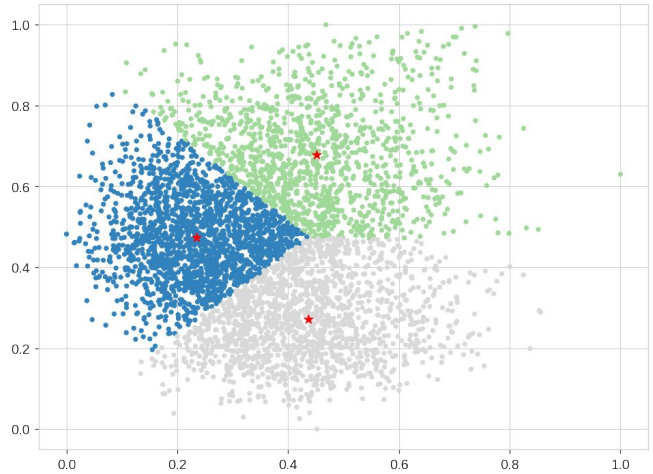
Principal Component Analysis reduce the dataset to a dimension of 4682 x 2 columns

K-Means

Finding the value for k: Average Silhouette Score

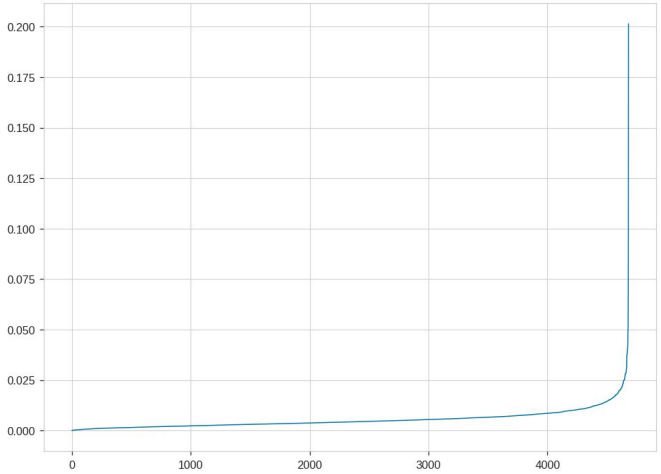


Found 3 Clusters with the relative centroids

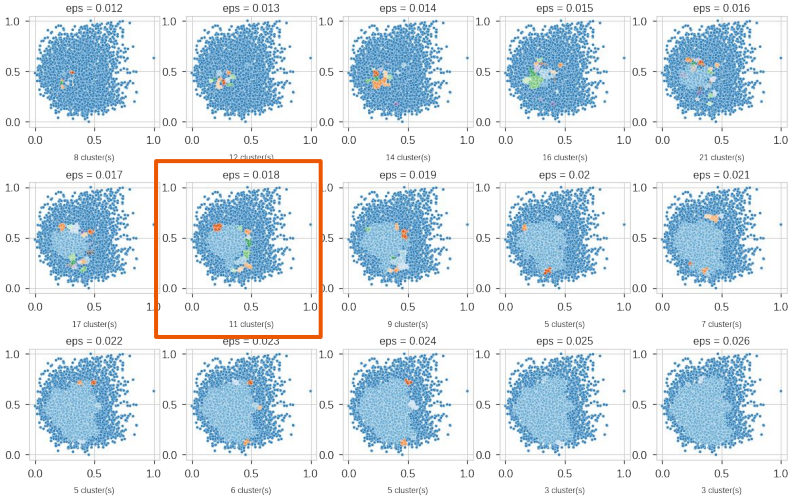


DBSCAN

Elbow Point for finding the right value of ε



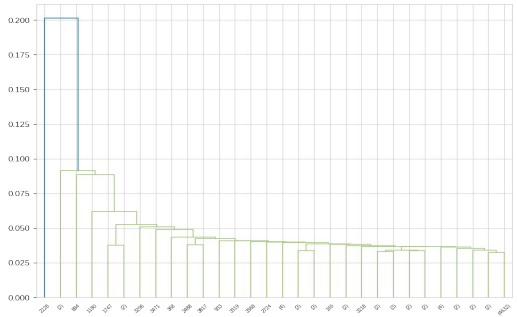
Fine Tuning varying ε between 0.012 and 0.026



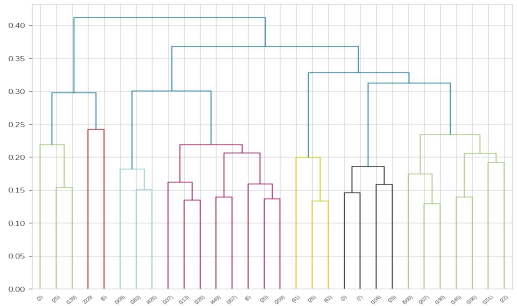
Hierarchical Clustering

The Dendograms

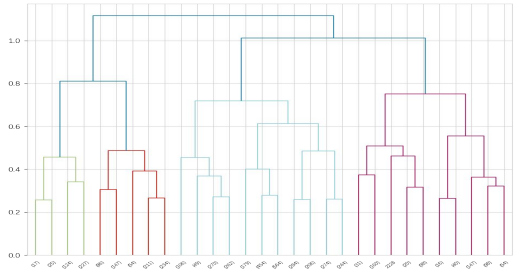
Single Link



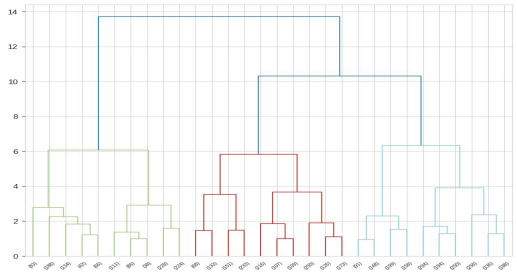
Group Average



Complete Link

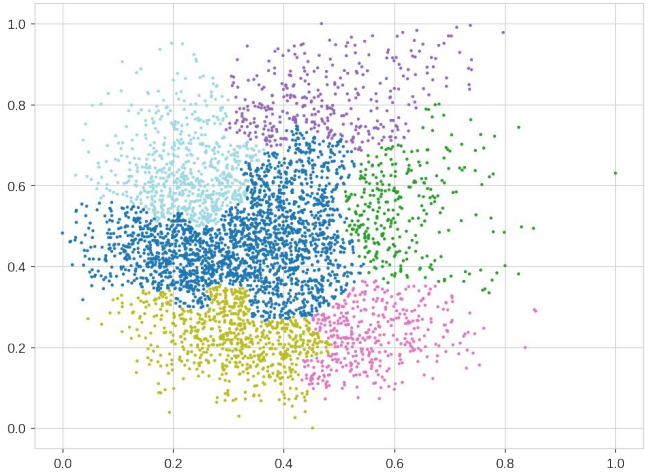


Ward's Method



Hierarchical Clustering

Complete

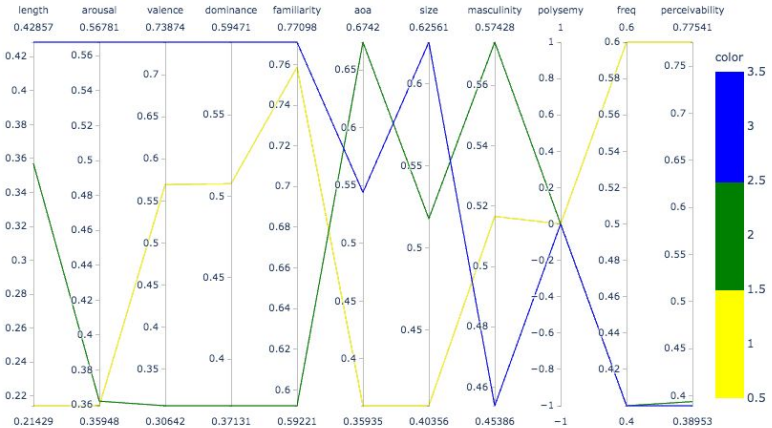


Average

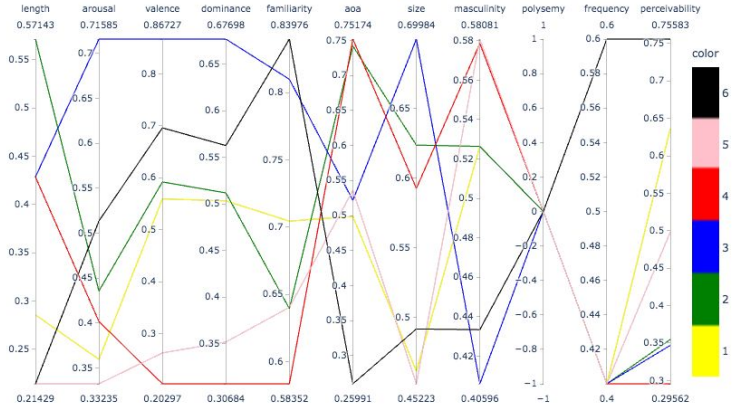
Evaluation of Clustering

	K-Means	DBSCAN	Hierarchical
Silhouette	0.36	-0.34	0.21
Calinski-Harabasz	3576	59.95	2075

K-Means



Hierarchical Clustering



Classification

Preprocessing

Trees don’t need a normalized datasets

Same preprocess as clustering but

- Length discrete
- Web Corpus Freq discrete
- Dataset is split in training (70%)
- and testing (30%)

Target Variable discretized

Binary division:

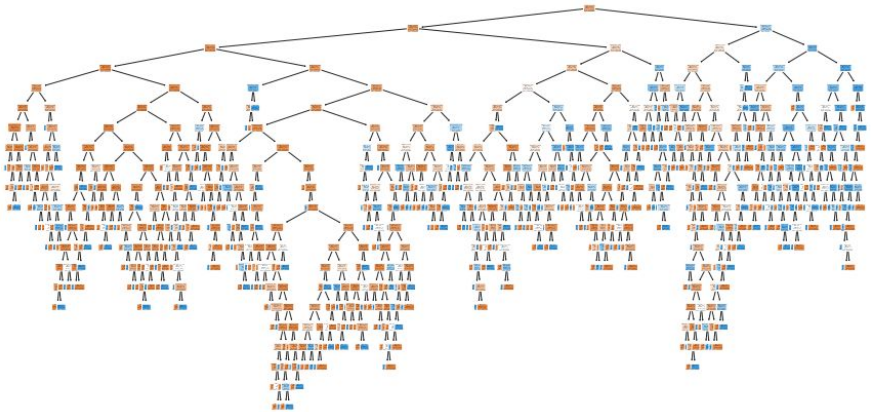
- Value < Threshold → Value=0
- Value > Threshold → Value=1

Multisplit:

- Web Corpus Freq
- Age of Acquisition

Decision Trees

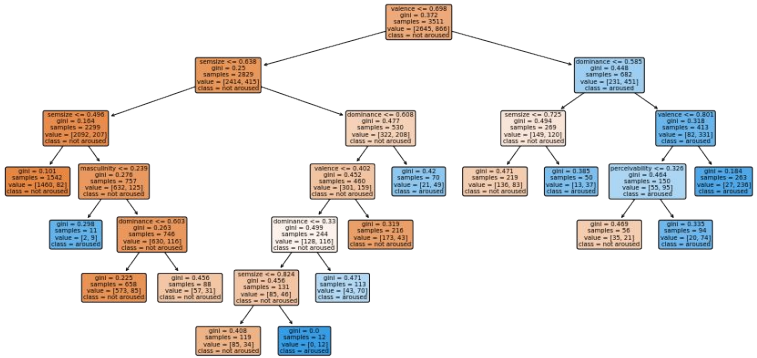
useful tool for classification but tend to overfit



Pruning techniques:

- Grid Search for best parameters (Max Depth, Min Sample Leaf…)
- **Cost Complexity Pruning:**

$$Score = Index + \alpha \cdot T(\#leaves)$$



10-fold Cross Validation is needed to check results

Random Forest & KNN

Decision Trees don't classify well results because of the overfitting.

Idea: create multiple trees (aka a **Forest**)

- Create different bootstrapped datasets
- For each bootstrapped dataset create a tree

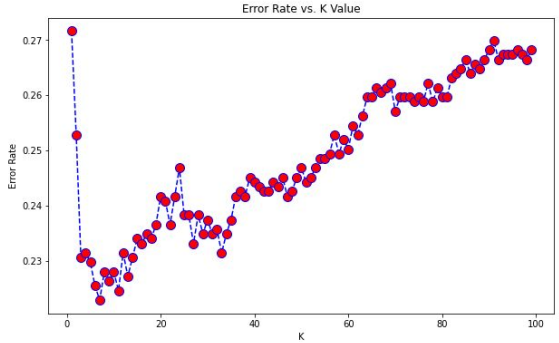
A new item is runned through the whole forest and classified according to the majority of trees

- **Out of Bag error:** misclassification error on the items not included it in the bootstrapped datasets

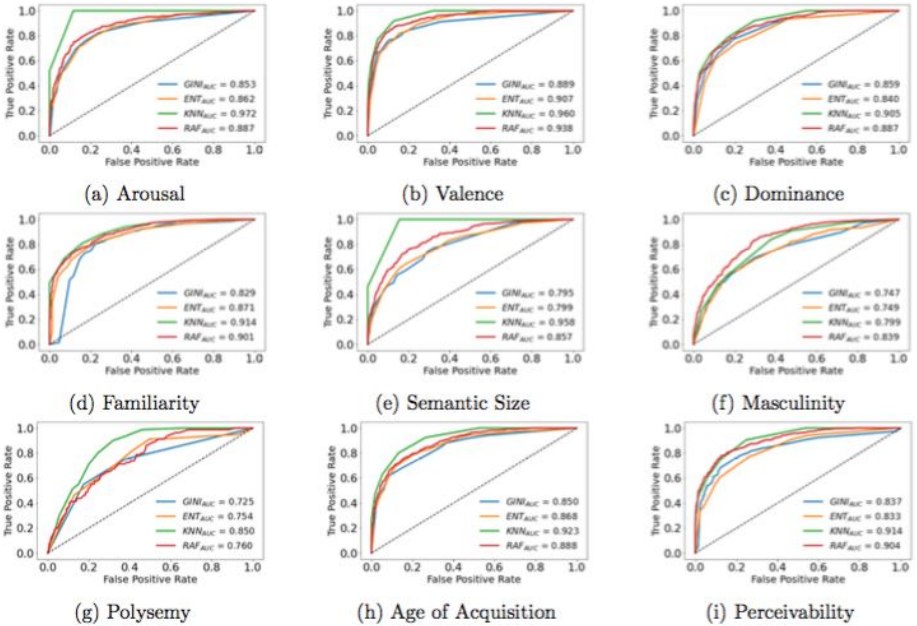
Idea: look at values with similar features

A new item is classified according to the majority of votes of its **k^{th} neighbors**

Problem: how do i choose k?
Answer: try different values and pick the one that minimize the error rate



Evaluation of models and target variables



Which **model** is to prefer?

- KNN and Random Forest perform better
- Entropy measure better than Gini

Which **variable** is to prefer?

F1 Score represents a tradeoff between precision and recall (harmonic mean)

- **Valence** has the best F1 and Accuracy Score
- **Polysemy** is already a boolean variable (no bias)
- **Age of Acquisition** can be compared to the multi-split analysis

Confusion Matrices for variables of interest

Binary Split Analysis

	Predicted 1	Predicted 0	Predicted 1	Predicted 0	Predicted 1	Predicted 0	
True 1	203	104	0	80	279	159	Decision
True 0	41	823	0	1091	70	663	Tree
True 1	207	100	1	79	299	139	Random
True 0	36	828	2	1080	73	660	Forest
True 1	178	129	30	50	276	162	
True 0	15	849	13	1078	39	694	KNN
	Valence		Polysemy		Age of Acquisition		

Behavior of ROC confirmed.

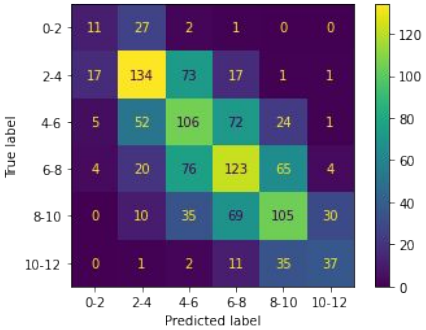
To notice: KNN is the only model that predicts some value of polysemic words

Multi Split Analysis

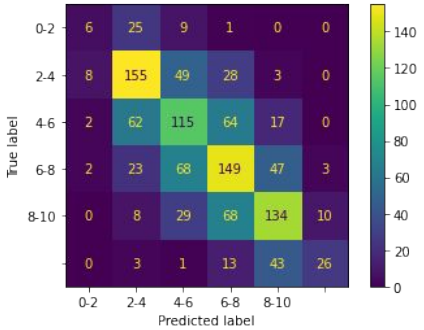
This analysis does not produce optimal results

KNN performs better also with the Multi Split

DT (Entropy)



KNN



Pattern Mining

Preprocess

1. Discretize the features in four intervals
    a. Normalize dataset between 0 and 4
    b. Floor the results
2. Discard the feature *polysemy*

    due to the fact that this attribute produce noise without contribute to the analysis

3. Add a label to each number (to recognize the variable)

Web Corpus Freq is already discretized between 4 and 9

Frequent Itemsets

Three parameters to set:

- Max number of items in set: None

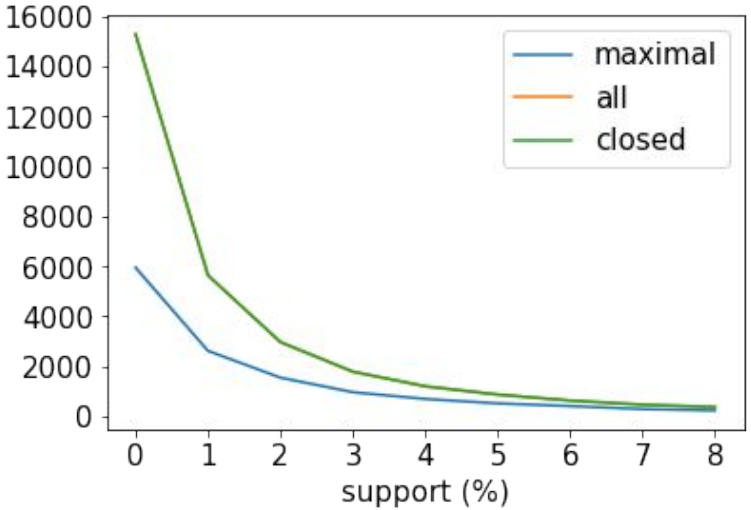
There is no reason to set a limit.

- Min number of items in set: 1

Cannot exclude a priori one-to-one rules. The irrelevant will be pruned by other thresholds

- Support: 6.5%

Number of rules do not decrease after this value



All itemsets are closed (orange line not visible)

Association Rules

Two other parameters to set:

- Confidence: 75%

Look for tradeoff between high value and good number of rules found

- Lift: 1.6

Not really a parameter: determines if a rule indicates correlation between antecedents and target.

We decided to look only for positive correlated rules

	Antecedent	Target	Lift	Conf	Supp (%)	Supp
1	(0.0_SemSize,)	3.0_Perceivability	2.45	0.76	7.78	363
2	(0.0_Age_of_Acquisition, 7.0_Web_Corpus_Freq)	3.0_Familiarity	2.23	0.93	7.26	339
3	(0.0_Age_of_Acquisition, 2.0_Dominance)	3.0_Familiarity	2.16	0.90	8.53	398
4	(0.0_Age_of_Acquisition, 1.0_Masculinity)	3.0_Familiarity	2.16	0.90	7.46	348
5	(0.0_Age_of_Acquisition, 0.0_Length)	3.0_Familiarity	2.11	0.88	10.00	467
6	(0.0_Age_of_Acquisition,)	3.0_Familiarity	2.08	0.87	14.29	667
7	(0.0_Age_of_Acquisition, 2.0_Valence)	3.0_Familiarity	2.07	0.87	8.78	410
8	(0.0_Age_of_Acquisition, 3.0_Perceivability)	3.0_Familiarity	2.05	0.86	8.27	386
9	(0.0_Age_of_Acquisition, 1.0_Arousal)	3.0_Familiarity	2.03	0.85	6.79	317
10	(3.0_Perceivability, 2.0_Dominance, 1.0_Arousal)	2.0_Valence	1.77	0.90	8.05	376
11	(7.0_Web_Corpus_Freq, 2.0_Dominance, 1.0_Arousal)	2.0_Valence	1.74	0.88	6.75	315
12	(3.0_Perceivability, 3.0_Familiarity, 1.0_Arou...	2.0_Valence	1.68	0.85	6.51	304
13	(3.0_Familiarity, 2.0_Dominance, 1.0_Arousal)	2.0_Valence	1.68	0.85	8.63	403
14	(1.0_SemSize, 2.0_Dominance, 1.0_Arousal)	2.0_Valence	1.67	0.84	8.40	392
15	(3.0_Perceivability, 1.0_Masculinity)	2.0_Valence	1.64	0.83	8.61	402
16	(1.0_Masculinity, 2.0_Dominance, 1.0_Arousal)	2.0_Valence	1.64	0.83	6.62	309
17	(2.0_Dominance, 1.0_Arousal, 2.0_Masculinity)	2.0_Valence	1.63	0.82	9.06	423
18	(3.0_Perceivability, 3.0_Familiarity, 2.0_Domi...	2.0_Valence	1.62	0.82	8.20	383
19	(2.0_Dominance, 1.0_Arousal, 1.0_Length)	2.0_Valence	1.61	0.82	8.80	411
20	(1.0_SemSize, 3.0_Familiarity, 2.0_Dominance)	2.0_Valence	1.61	0.81	6.79	317
21	(3.0_Perceivability, 2.0_Dominance, 1.0_Length)	2.0_Valence	1.60	0.81	6.68	312
22	(3.0_Valence, 2.0_SemSize)	2.0_Dominance	1.69	0.76	6.90	322

The majority of the rules concerns the feature *Valence* but the ones with the highest lift apply to *Familiarity*.

Replacing Missing Values

Only variable with missing values in our dataset is Web Corpus Frequency, but no rules with the chosen cuts concern this variable

1. Loose the parameters: supp>2%, conf>50%, lift>1.3, zmin=2
2. Pick only the ones that predict Web Corpus Frequency
3. Confront the rules found with the attributes of missing values and keep only the ones that concern the rows with missing values
4. Count the results

Word	Target Variable	Counts
Christmas	7.0 Web Corpus Freq	2
Dad	7.0 Web Corpus Freq	6
FALSE	6.0 Web Corpus Freq	3
FALSE	7.0 Web Corpus Freq	1
Mom	7.0 Web Corpus Freq	7
Mum	7.0 Web Corpus Freq	3
Mummy	7.0 Web Corpus Freq	2
TRUE	7.0 Web Corpus Freq	1
TV	7.0 Web Corpus Freq	10
Twitter	7.0 Web Corpus Freq	1
yo-yo	6.0 Web Corpus Freq	6

With a tighter cut of parameters (supp>4, conf>60,lift>1.4) one find only one rule that concern yo-yo

Final Discussion

Results according to analyzed techniques

The analysis resulted in some trends:

- *Valence* good target variable
  - high F1 Score, Precision and Accuracy in Classification
  - Large number of extracted rules
- Strong association between low values of *Age of Acquisition* and high values of *Familiarity*
  - First Cluster in K-Means
  - Sixth Cluster of Hierarchical Clustering
  - Root and recurrent split in decision tree
  - Large number of extracted rules with high lift
- High *valence*, high *dominance*, large *size*
  - Third Cluster of Hierarchical Clustering
  - One extracted rule