Classification of text using Association Rule mining with Critical Relative Support based pruning

ICACCI, Jaipur

by

Saurabh Mathur (VIT, Vellore)

September 23, 2016

Contents

Introduction

Association Rules

```
{Bread, Milk} => {Butter}
```

The Apriori algorithm [1] is a way to generate such rules from transaction data

Problem Statement

Find the minimal set of interesting association rules *from text*

The dataset

BBC Insights dataset [5]

2004-2005	Time period
2,225	Documents
5	Categories

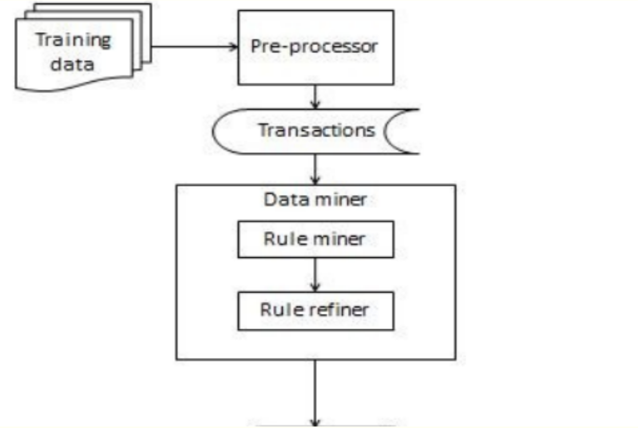
Background

Literature Survey

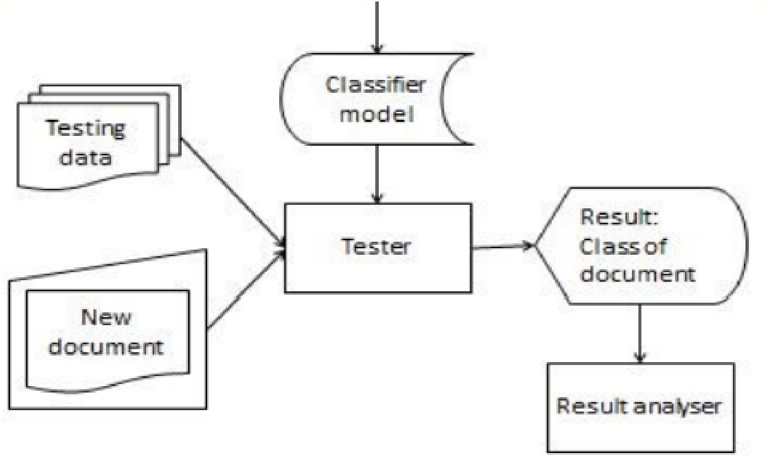
Citation	tl;dr
Rakesh Agrawal and Ramakrishnan Srikant; 1993 [1]	Apriori Algorithm proposed.
Zailani Abdullah, et al.; 2011 [2]	Proposed a metric (CRS) to remove uninteresting rules.
Gayathri, K.; Marimuthu, A.; 2013 [3]	Text classification performance is independent of size of feature space in most cases.
Kulkarni et al. 2012 [6]	Feature co-occurrence and association are can be used for classification
Kadhim, A.I.; Cheah, Yu.-N; 2014 [4]	Term weighing and NLP reduce the dimensionality of the feature space

Proposed Idea

Flowchart - Part 1



Flowchart - Part 2



Pre-Processing

Convert text document to transaction format

Tokenizing. Stop word removal. POS Tagging. Stemming. Tf-Idf.

Find representative features

Features found only in a particular class

Rule mining

$$X \;\; => \;\; Y$$

$Y \in Classes$
Classes = { Business, Entertainment, Politics, Sports, Technology }

Rule Filtering with CRS

Before applying CRS filtering, update support as -

$$Mod_{Supp}(X => Y) = (1 - a) \cdot (1 - b) \cdot (1 - c) \cdot supp(X => Y)$$

Where,

$$a = \frac{\text{Number of documents of class Y}}{\text{Total number of documents}}$$

$$b = \frac{\text{Number of representative features of class Y}}{\text{Total number of documents}}$$

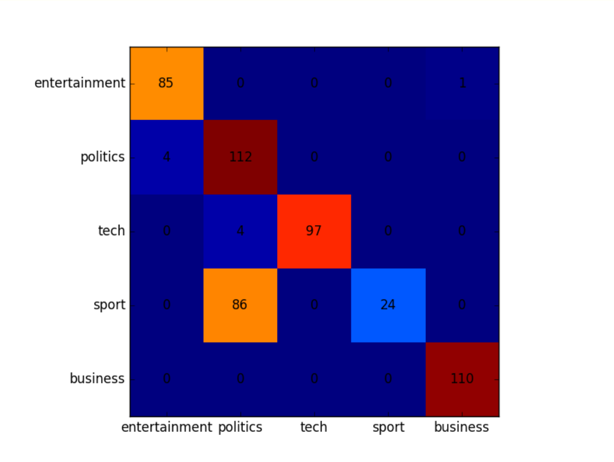
$$c = \frac{\text{Number of rules of class Y}}{\text{Total number of rules}}$$

Results & Conclusion

Accuracy

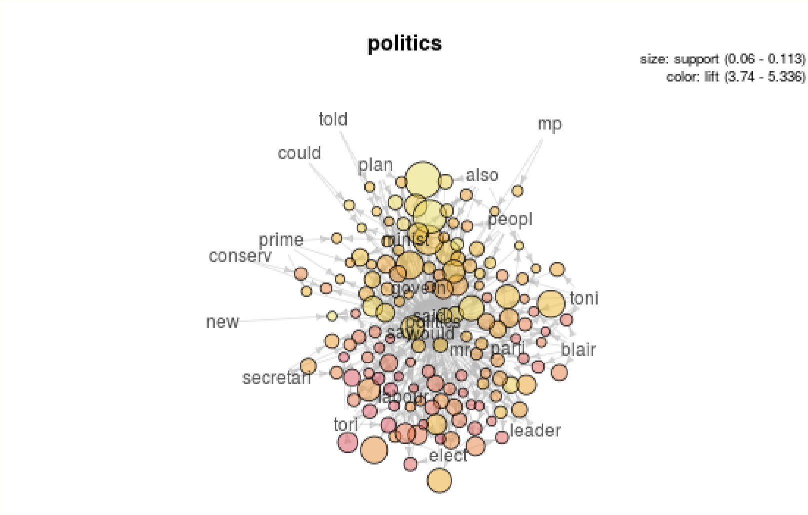
81 %

Confusion Matrix

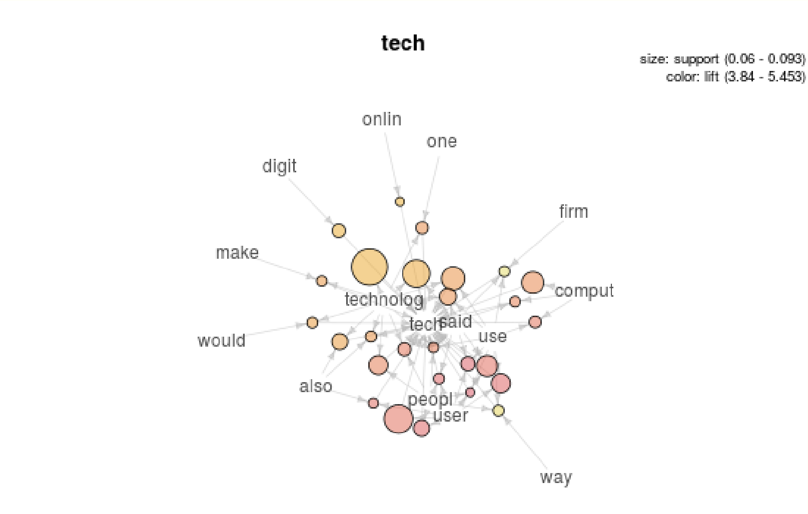


Predicted categories v/s Actual categories

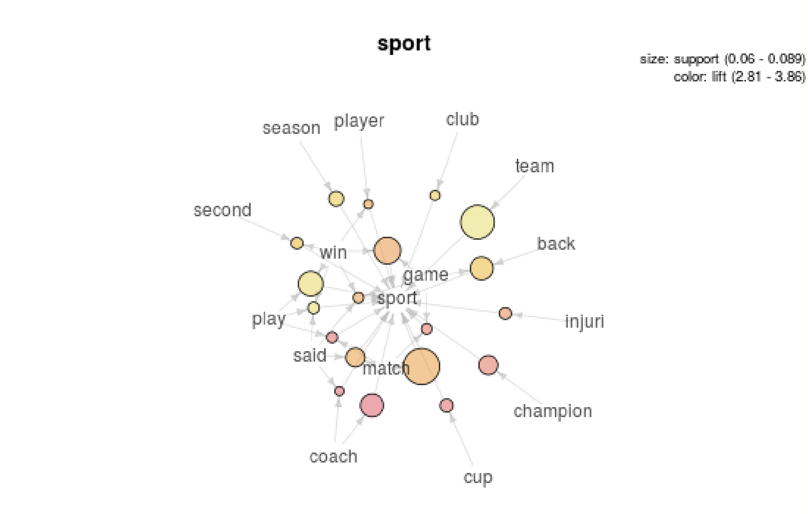
Rules - politics



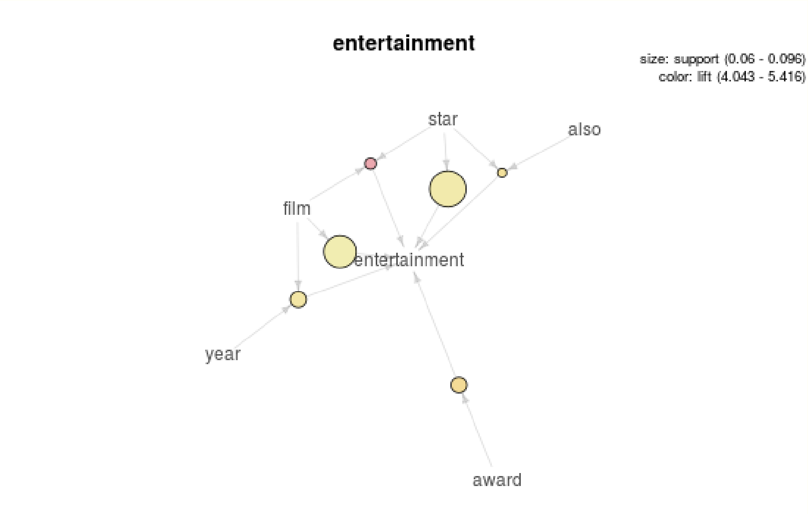
Rules - tech


tech
size: support (0.06 - 0.093)
color: lift (3.84 - 5.453)
onlin
one
digit
firm
make
comput
technolog
tech
said
would
use
also
peopl
user
way

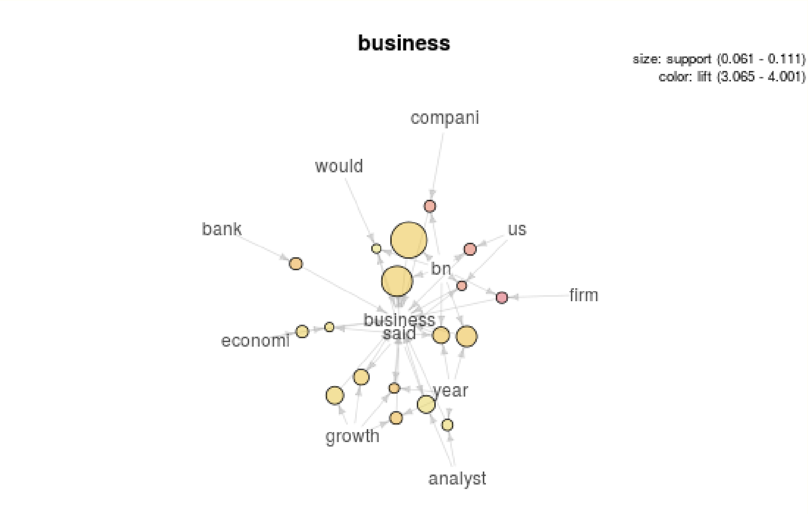
Rules - sport


sport
size: support (0.06 - 0.089)
color: lift (2.81 - 3.86)
season
player
club
team
second
win
back
game
sport
play
injuri
said
match
champion
coach
cup

Rules - tech


entertainment
size: support (0.06 - 0.096)
color: lift (4.043 - 5.416)
star
also
film
entertainment
year
award

Rules - business



Pros

- Uses existing tools (Apriori Algorithm).
- 50% to 67% faster than Apriori.
- More transparent than other text classification Algorithms.
- Allows domain experts to tune the model.

Cons

- Ignores the relative order of terms.
- While used for pre-processing, term frequency is not considered for classification.
- Parameters (Minimum support, Minimum confidence, Critical Relative support threshold) need to be tuned carefully.

References

1. Rakesh Agrawal and Ramakrishnan Srikant, ”Fast algorithms for mining association rules in large databases, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB”, pages 487-499, Santiago, Chile, September 1994.
2. Zailani Abdullah, Tutut Herawan, Noraziah Ahmad and, Mustafa Mat Deris, ”Mining significant association rules from educational data using critical relative support approach, Procedia - Social and Behavioral Sciences”, 28 (2011) 97 – 101.
3. Gayathri, K.; Marimuthu, A., ”Text document pre-processing with the KNN for classification using the SVM,” in Intelligent Systems and Control (ISCO), 2013 7th International Conference on, vol., no., pp.453-457, 4-5 Jan. 2013.

References

4. Kadhim, A.I.; Cheah, Yu.-N.; Ahamed, N.H., ”Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering,” in Artificial Intelligence with Applications in Engineering and Technology (ICAIET), 2014 4th International Conference on , vol., no., pp.69-73, 3-5 Dec. 2014.
5. D. Greene and P. Cunningham. ”Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering”, Proc. ICML 2006
6. Kulkarni, A.R.; Tokekar, V.; Kulkarni, P., ”Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining,” in Software Engineering (CONSEG), 2012 CSI Sixth International Conference”, Sept.2012.