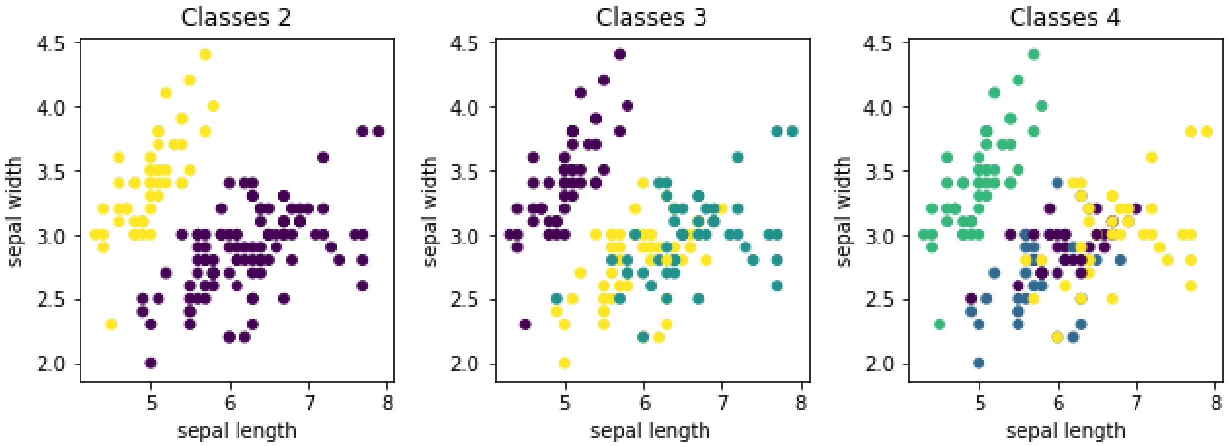
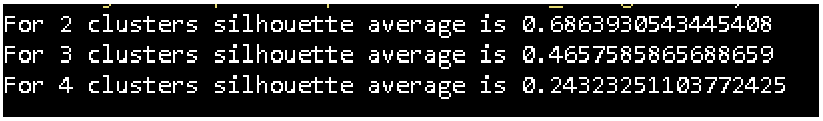
Solutions and Screenshots

1.1



1.2

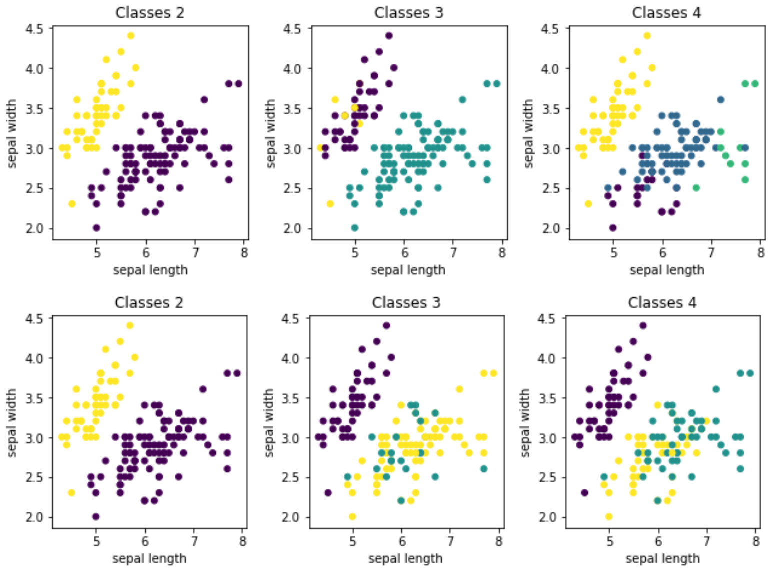
There is an internal criterion which states that high intra-cluster similarity is observed in good clustering. So, in case of **2** classes there is high intra-cluster similarity. This can also be done by calculating silhouette_score. The value near to 1 indicates non-overlapping clusters while negative value states that data assign to wrong cluster. For different clusters silhouette scores are



```
For 2 clusters silhouette average is 0.6863930543445408
For 3 clusters silhouette average is 0.4657585865688659
For 4 clusters silhouette average is 0.24323251103772425
```

This shows for 2 classes it is better.

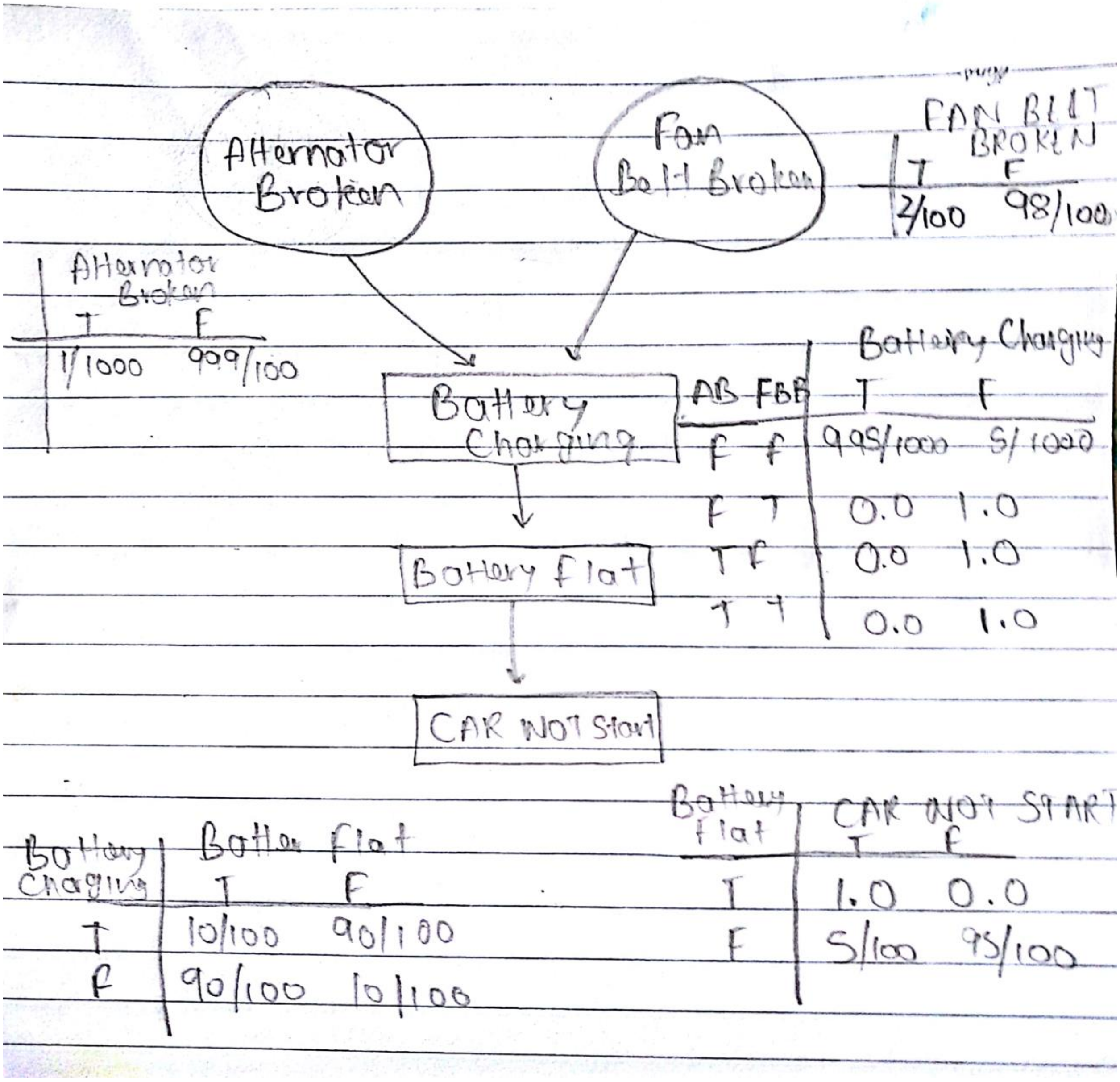
There is a stability based method [3], in which when data repeatedly produce similar clusters there is a high level of agreement. Based on this



As we can see after repeated iteration clustering model (2) produce the cluster with high level of stability and agreement.

2.

2.1 Casual Model



Alternator Broken	
T	F
1/1000	999/100

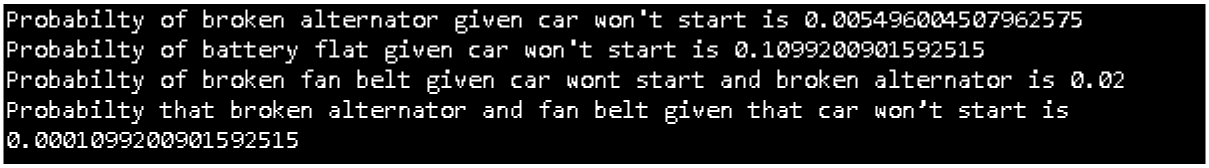
FAN BELT BROKEN	
T	F
2/100	98/100

		Battery Charging	
AB	FBB	T	F
F	F	995/1000	5/1000
F	T	0.0	1.0
T	F	0.0	1.0
T	T	0.0	1.0

Battery Charging	Battery flat	
	T	F
T	10/100	90/100
F	90/100	10/100

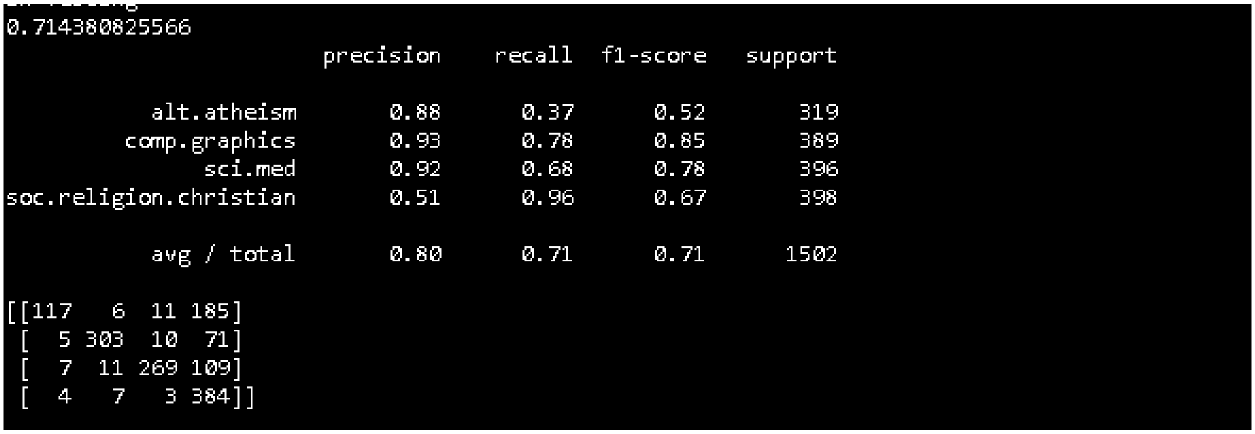
Battery flat	CAR NOT START	
	T	F
T	1.0	0.0
F	5/100	95/100

2.2 – 2.5



```
Probabilty of broken alternator given car won't start is 0.005496004507962575
Probabilty of battery flat given car won't start is 0.1099200901592515
Probabilty of broken fan belt given car wont start and broken alternator is 0.02
Probabilty that broken alternator and fan belt given that car won’t start is
0.0001099200901592515
```

3. Model designed only for 4 categories namely - alt.atheism, soc.religion.christian, comp.graphics, sci.med as discussed with the TA. Accuracy score, classification report and confusion matrix given below.



```
0.714380825566
                        precision    recall  f1-score   support

           alt.atheism       0.88      0.37      0.52       319
         comp.graphics       0.93      0.78      0.85       389
               sci.med       0.92      0.68      0.78       396
soc.religion.christian       0.51      0.96      0.67       398

           avg / total       0.80      0.71      0.71      1502

[[117   6  11 185]
 [  5 303  10  71]
 [  7  11 269 109]
 [  4   7   3 384]]
```

References

[1] 2017. [Online]. Available: http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf. [Accessed: 16- Oct- 2017].

[2] "sklearn.metrics.silhouette_score — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. [Accessed: 16- Oct- 2017].

[3] 2017. [Online]. Available: http://web.engr.oregonstate.edu/~xfern/classes/cs534/notes/Unsupervised-model-11.pdf. [Accessed: 16- Oct- 2017].

[4] "piedenis / Lea — Bitbucket", Bitbucket.org, 2017. [Online]. Available: https://bitbucket.org/piedenis/lea. [Accessed: 16- Oct- 2017].

[5] "Bayesian network", En.wikipedia.org, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Bayesian_network. [Accessed: 16- Oct- 2017].

[6] "Working With Text Data — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html. [Accessed: 16- Oct- 2017].

[7] "The Bernoulli model", Nlp.stanford.edu, 2017. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html. [Accessed: 16- Oct- 2017].

[8] "Text Classification Using Naive Bayes", YouTube, 2017. [Online]. Available: https://www.youtube.com/watch?v=EGKeC2S44Rs. [Accessed: 16- Oct- 2017].