Krista Miller

Data Mining Assignment 1- part 2

Purpose:

Association rules are descriptive summaries of relationships between attributes. It is a rule-based unsupervised machine learning method for discovering interesting relationships among frequent variables in large datasets.

This data mining project applies association rule mining to three works of English literature, written by English people at roughly the same time:

- Sense and Sensibility, by Jane Austin, published 1811
- A Tale of Two Cities, by Charles Dickens, published 1859
- On The Origin of Species, by Charles Darwin, published 1859 (although written over a 20 year period preceding that date).

The sentences of each book have been processed in a transaction format, where each word is comma-separated. This project explores each text individually and collectively, to see what can be determined about the texts.

Method:

I used the processed text files that contained at most 10 words per line, with punctuation removed and all words converted to lower case. Using the Apriori algorithm, 1, 2, and 3 item-sets were studied for all texts, using a minimum support of 0.003 for both Sense and Sensibility and A Tale of Two Cities. I encountered an error when I applied the same parameters to On the Origin of Species. After several experiments, the most interesting results I found was using a minimum support of 0.005 and a minimum threshold of 0.8.

The lowest minimum support that I could use for On the Origin of Species was 0.008 without encountering an error.

Sense and Sensibility, by Jane Austin, published 1811

Code parameters:

```
freq_items= apriori(ohe_df, min_support=0.003, use_colnames=True, verbose=1)

#print(freq_items.sort_values(by="support", ascending=False))
freq_items['itemlength']=freq_items['itemsets'].apply(len)
freq_items.query('itemlength == 3').sort_values(by="support", ascending=False)
```

```
#Mining Association Rules
#(results show which item is frequently with other items)

rules= association_rules(freq_items, metric="confidence", min_threshold=0.8)
rules.sort_values(by="lift", ascending=False).head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(middleton)	(lady)	0.006709	0.009268	0.005948	0.886598	95.666564	0.005886	8.736459
1	(sir)	(john)	0.007677	0.010512	0.006432	0.837838	79.699324	0.006351	6.101840
2	(brandon)	(colonel)	0.008853	0.011481	0.007815	0.882812	76.895096	0.007714	8.435364
9	(i, sure)	(am)	0.005533	0.017498	0.004565	0.825000	47.148913	0.004468	5.614299
12	(to, jennings)	(mrs)	0.003942	0.034442	0.003873	0.982456	28.524766	0.003737	55.036794
6	(jennings)	(mrs)	0.015354	0.034442	0.014662	0.954955	27.726293	0.014133	21.435383
13	(and, jennings)	(mrs)	0.004634	0.034442	0.004357	0.940299	27.300755	0.004198	16.173093
15	(of, jennings)	(mrs)	0.003320	0.034442	0.003112	0.937500	27.219503	0.002998	15.448925
14	(the, jennings)	(mrs)	0.003942	0.034442	0.003596	0.912281	26.487282	0.003461	11.007359
10	(am, sure)	(i)	0.004703	0.120064	0.004565	0.970588	8.083949	0.004000	29.917837

Sense and Sensibility Analysis:

The association rule {middleton} -> {lady} has the highest lift of 95.66 (minimum confidence threshold= 0.8). 95.66 is the rise in probability of having 'lady' in the same sentence with the knowledge of 'middleton' over the probability of having 'lady' without any knowledge about the presence of 'middleton'. The higher the lift, the greater the chances of seeing 'lady' if we have already seen 'middleton'. Likewise, the association rule {sir} -> {john} has a lift of 79.69 (minimum confidence threshold= 0.8) means that there is a greater chance of 'john' occurring in a sentence if we have already seen 'sir'. The association rule {brandon} -> {colonel} is also an interesting association rule, with a lift of 76.89 and the same minimum confidence threshold.

A Tale of Two Cities, by Charles Dickens, published 1859

Code parameters:

```
freq_items= apriori(ohe_df, min_support=0.003, use_colnames=True, verbose=1)

#print(freq_items.sort_values(by="support", ascending=False))
freq_items['itemlength']=freq_items['itemsets'].apply(len)
freq_items.query('itemlength == 3').sort_values(by="support", ascending=False)
```

```
#Mining Association Rules
#(results show which item is frequently with other items)

rules= association_rules(freq_items, metric="confidence", min_threshold=0.8)
rules.sort_values(by="lift", ascending=False).head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(pross)	(miss)	0.008257	0.012083	0.007468	0.904412	74.852572	0.007368	10.335136
1	(don)	(t)	0.007954	0.012629	0.007286	0.916031	72.533764	0.007186	11.758691
7	(don, i)	(t)	0.003522	0.012629	0.003218	0.913793	72.356598	0.003173	11.453503
9	(said, lorry)	(mr)	0.003522	0.033151	0.003339	0.948276	28.604585	0.003223	18.692410
11	(his, lorry)	(mr)	0.003643	0.033151	0.003218	0.883333	26.645604	0.003097	8.287276
8	(and, lorry)	(mr)	0.004129	0.033151	0.003522	0.852941	25.728830	0.003385	6.574572
4	(lorry)	(mr)	0.019672	0.033151	0.016758	0.851852	25.695971	0.016106	6.526230
10	(to, lorry)	(mr)	0.003886	0.033151	0.003218	0.828125	24.980254	0.003089	5.625302
12	(the, lorry)	(mr)	0.005586	0.033151	0.004554	0.815217	24.590898	0.004369	5.232358
2	(am)	(i)	0.013054	0.103279	0.012447	0.953488	9.232189	0.011099	19.279508

A Tale of Two Cities Analysis:

The association rule with the highest lift equal to 74.85 is {pross} -> {miss} (minimum confidence threshold = 0.8) means there is a greater chance of 'miss' showing up in a sentence if we have already seen 'pross'. The next highest association rule is don ->t, suggesting that the csv parsing interpreted 'don't' as two separate words. Therefore, don -> t isn't an intuitively interesting rule. Then, there are several rules with the word 'lorry' and 'mr'. The words {said,lorry} -> {mr} means there is a greater change of 'mr' showing up in a sentence if we also have the words 'said' and 'lorry'.

Similar to Sense and Sensibility, several English titles are associated with character names. This suggests that, during this time period, it was very common to use honorifics to convey respect or refer to people. For example, we see this with Mr. Lorry, Miss Pross, Lady Middleton, Sir John, Colonel Brandon, and Mrs. Jennings.

On The Origin of Species, by Charles Darwin, published 1859

Code parameters:

```
#Applying Apriori:
#min_support: floating point between 0 and 1 , (# observation with item)/(total observation)

freq_items= apriori(ohe_df, min_support=0.005, use_colnames=True, verbose=1)

#print(freq_items.sort_values(by="support", ascending=False))
freq_items['itemlength']=freq_items['itemsets'].apply(len)
freq_items.query('itemlength == 2').sort_values(by="support", ascending=False)
```

```
#Mining Association Rules
#(results show which item is frequently with other items)

rules= association_rules(freq_items, metric="confidence", min_threshold=0.8)
rules.sort_values(by="lift", ascending=False).head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(descended)	(from)	0.005757	0.059248	0.005142	0.893204	15.075765	0.004801	8.808863
2	(amount)	(of)	0.006204	0.355765	0.005645	0.909910	2.557612	0.003438	7.151003
12	(the, life)	(of)	0.008328	0.355765	0.007210	0.865772	2.433547	0.004247	4.799547
11	(inhabitants, the)	(of)	0.006540	0.355765	0.005645	0.863248	2.426452	0.003319	4.710965
13	(of, same, species)	(the)	0.006204	0.440445	0.006148	0.990991	2.249977	0.003416	62.110614
9	(of, same)	(the)	0.017439	0.440445	0.016712	0.958333	2.175830	0.009031	13.429322
8	(same, species)	(the)	0.008775	0.440445	0.008384	0.955414	2.169202	0.004519	12.550021
10	(of, inhabitants)	(the)	0.005981	0.440445	0.005645	0.943925	2.143118	0.003011	9.978732
6	(in, same)	(the)	0.011850	0.440445	0.010899	0.919811	2.088369	0.005680	6.977981
1	(same)	(the)	0.038399	0.440445	0.034263	0.892285	2.025873	0.017350	5.194788

On The Origin of Species Analysis:

The association rule with the highest lift equal to 15.07 is {descended} -> {from} (minimum confidence threshold = 0.8) means there is a greater chance of 'from' showing up in a sentence if we have already seen 'descended'. The next three association rules have the same consequents 'of'. {Amount} -> {of} (lift= 2.55), {the, life} -> {of} (lift = 2.43), {inhabitants, the} -> {of} (lift = 2.42).

Compared to Sense and Sensibility and A Tale of Two Cities, On the Origin of Species is a non-fiction scientific work rather than fiction. The association rules in Darwin's work are scientific words (species, inhabitants, descended, amount, life) and in contrast to character names seen in the association rules from Tale of Two Cities and Sense and Sensibility.