

Abstracts

Here is the list of all the abstracts used in the data file.

Source	Category	Number
https://europepmc.org/abstract/med/7658958	Computer Science	1
https://europepmc.org/abstract/MED/31531532	Computer Science	2
https://europepmc.org/abstract/MED/31489585	Computer Science	3
https://europepmc.org/abstract/MED/30886093	Computer Science	4
https://europepmc.org/abstract/MED/30957009	Computer Science	5
https://europepmc.org/abstract/MED/30205258	Computer Science	6
https://europepmc.org/abstract/PPR/PPR46281	Computer Science	7
https://europepmc.org/abstract/PPR/PPR50534	Computer Science	8
https://europepmc.org/abstract/MED/28652335	Computer Science	9
https://europepmc.org/abstract/MED/29249891	Computer Science	10
https://europepmc.org/abstract/MED/28790936	Computer Science	11
https://europepmc.org/abstract/MED/28487664	Computer Science	12
https://europepmc.org/abstract/MED/28456017	Computer Science	13
https://europepmc.org/abstract/MED/28479315	Computer Science	14
https://europepmc.org/abstract/MED/29367799	Computer Science	15
https://europepmc.org/abstract/MED/25824671	Computer Science	16
https://europepmc.org/abstract/MED/28400991	Computer Science	17
https://europepmc.org/abstract/MED/27227145	Computer Science	18
https://europepmc.org/abstract/MED/26433013	Computer Science	19
https://europepmc.org/abstract/AGR/IND605264253	Computer Science	20
https://europepmc.org/abstract/MED/31605232	Breast Cancer	1
https://europepmc.org/abstract/MED/31605227	Breast Cancer	2
https://europepmc.org/abstract/MED/31577379	Breast Cancer	3
https://europepmc.org/abstract/MED/31605532	Breast Cancer	4
https://europepmc.org/abstract/MED/31512725	Breast Cancer	5
https://europepmc.org/abstract/MED/31395590	Breast Cancer	6
https://europepmc.org/abstract/MED/31456069	Breast Cancer	7
https://europepmc.org/abstract/MED/31496430	Breast Cancer	8

https://europepmc.org/abstract/MED/31590453	Breast Cancer	9
https://europepmc.org/abstract/MED/31490845	Breast Cancer	10
https://europepmc.org/abstract/MED/30884986	Breast Cancer	11
https://europepmc.org/abstract/MED/31295322	Breast Cancer	12
https://europepmc.org/abstract/MED/31581056	Breast Cancer	13
https://europepmc.org/abstract/MED/31527531	Breast Cancer	14
https://europepmc.org/abstract/MED/31601987	Breast Cancer	15
https://europepmc.org/abstract/MED/31366131	Breast Cancer	16
https://europepmc.org/abstract/MED/31576449	Breast Cancer	17
https://europepmc.org/abstract/MED/31464762	Breast Cancer	18
https://europepmc.org/abstract/MED/31312788	Breast Cancer	19
https://europepmc.org/abstract/MED/31175583	Breast Cancer	20
https://europepmc.org/abstract/MED/31611544	Aircraft	1
https://europepmc.org/abstract/MED/31547365	Aircraft	2
https://europepmc.org/abstract/MED/31581502	Aircraft	3
https://europepmc.org/abstract/MED/31510463	Aircraft	4
https://europepmc.org/abstract/MED/31557959	Aircraft	5
https://europepmc.org/abstract/MED/31480420	Aircraft	6
https://europepmc.org/abstract/MED/31558194	Aircraft	7
https://europepmc.org/abstract/MED/31260521	Aircraft	8
https://europepmc.org/abstract/MED/31558286	Aircraft	9
https://europepmc.org/abstract/MED/31488246	Aircraft	10
https://europepmc.org/abstract/MED/31336303	Aircraft	11
https://europepmc.org/abstract/MED/31435938	Aircraft	12
https://europepmc.org/abstract/MED/31563003	Aircraft	13
https://europepmc.org/abstract/MED/31392506	Aircraft	14
https://europepmc.org/abstract/MED/31309420	Aircraft	15
https://europepmc.org/abstract/MED/31261705	Aircraft	16
https://europepmc.org/abstract/MED/30747795	Aircraft	17
https://europepmc.org/abstract/MED/31478471	Aircraft	18
https://europepmc.org/abstract/PPR/PPR82017	Aircraft	19
https://europepmc.org/abstract/MED/31502943	Aircraft	20

Task 1

Selected Words

Stop words:

- “a” (col. 57)
- “and” (col. 121)
- “the” (col.1163)

Rare Words:

- “computer” (col. 279)
- “cancer” (col. 1498)
- “aircraft” (col. 2345)

Other Words:

- “each” (col. 396)
- “study” (col. 1124)
- “technology” (col. 1152)

Data Table (Task 1 - Binary)

No.	domain	a	and	the	computer	cancer	aircraft	each	study	technology
10	Computer-Science	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
30	Breast-Cancer	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
50	Aircraft	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0

Data Table (Task 2 – Term Frequency)

No.	domain	a	and	the	computer	cancer	aircraft	each	study	technology
10	Computer-Science	0	10	7	4	0	0	0	0	0
30	Breast-Cancer	1	5	4	0	5	0	0	1	0
50	Aircraft	2	9	18	0	0	2	0	0	0

Data Table (Task 3 – TF*IDF)

No.	domain	a	and	the	computer	cancer	aircraft	each	study	technology
10	Computer-Science	0	0	0	5.28702336	0	0	0	0	0

30	Breast-Cancer	0.0870 1138	0	0	0	5.4930 6144	0	0	0.798 5077	0
50	Aircraft	0.1740 2275	0	0	0	0	2.197224 58	0	0	0

Data Table Analysis

At first glance, it's quite obvious that the first data table is drastically different in values to the next 2. In the binary data table, the values only range from 0.0-1.0. This makes sense since Weka is determining if the word is in the abstract (1.0), or not in the abstract (0.0). The term frequency table represents the literal counts for the words in the abstract. Because of this, the values for the stop-words are much higher compared to the rest of the words per abstract (constantly seeing "a", "and", and "the" having higher values overall per row). However, in the TF*IDF table, we see the opposite. The stop-words are valued at 0 or close to 0, whereas the rare-words are valued much higher (we see a value of about 5 for "computer" and "cancer" for their respective domains, and then a value of about 2 for "aircraft" in its domain). From these values, we can conclude that the words "a", "and", and "the" have a Inverse Document Frequency (IDF) of close to 0. In the analysis of the screenshots below, we can see that the words "and" and "the" appear in 60/60 of the abstracts, so their IDF would equal 0. Looking at the words "cancer" and "aircraft", both those words only appear 20/60 of the abstracts, and only in those abstracts of their domain. This would rank these words as far more valuable since they do not appear as frequently. The score of 2.19722458 for "aircraft" in the TF*IDF table is most likely due to the value of 2 in its Term Frequency table. Since it only appeared twice in that particular abstract, it will have a lower value, than "cancer", which appeared 5 times for abstract #30. If we were to pick another abstract that had the word "aircraft" more than 2 times in the abstract, we should expect its TF*IDF value to increase. The word "study" is an interesting case as it appears in 27/60 abstracts in the data set and is evenly scattered among the 3 domains. Because of this, the word "study" acts as the middle ground between the very frequent words (the stop-words, "a", "and", "the") and the domain-specific words (rare-words, "computer", "cancer", "aircraft"). Because of this, we can expect the word "study" to have an IDF ranging somewhere between the stop-words, and the rare-words.

Task 2

Cluster Summarization Table

	Clustering effectiveness with different vector representations		
	Binary	TF	TF*IDF
Incorrectly clustered (%)	61.6667%	63.3333%	63.3333%

Words to Keep TF*IDF Summarization Table

	Clustering effectiveness with different number of attributes			
Words to Keep	10	40	160	640

#	28	120	505	2972
Attributes				
Incorrectly clustered	15%	31.6667%	48.3333%	63.3333%

Task 2 Analysis

Actually, quite surprised at the outcomes of the clustering. I was expecting the incorrect percentage to go down instead of increased between Binary, TF, and TF*IDF. What was more unexpected was that there seemed to be no improvement between TF and TF*IDF. Most were clustered into Computer-Science with only 2 other abstracts not being categorized as such. Those 2 were then correctly clustered into their domain, leaving 22 correct abstracts. Perhaps many of the words found in Aircraft and Breast-Cancer abstracts seemed to have overlapped with words included in many Computer-Science abstracts.

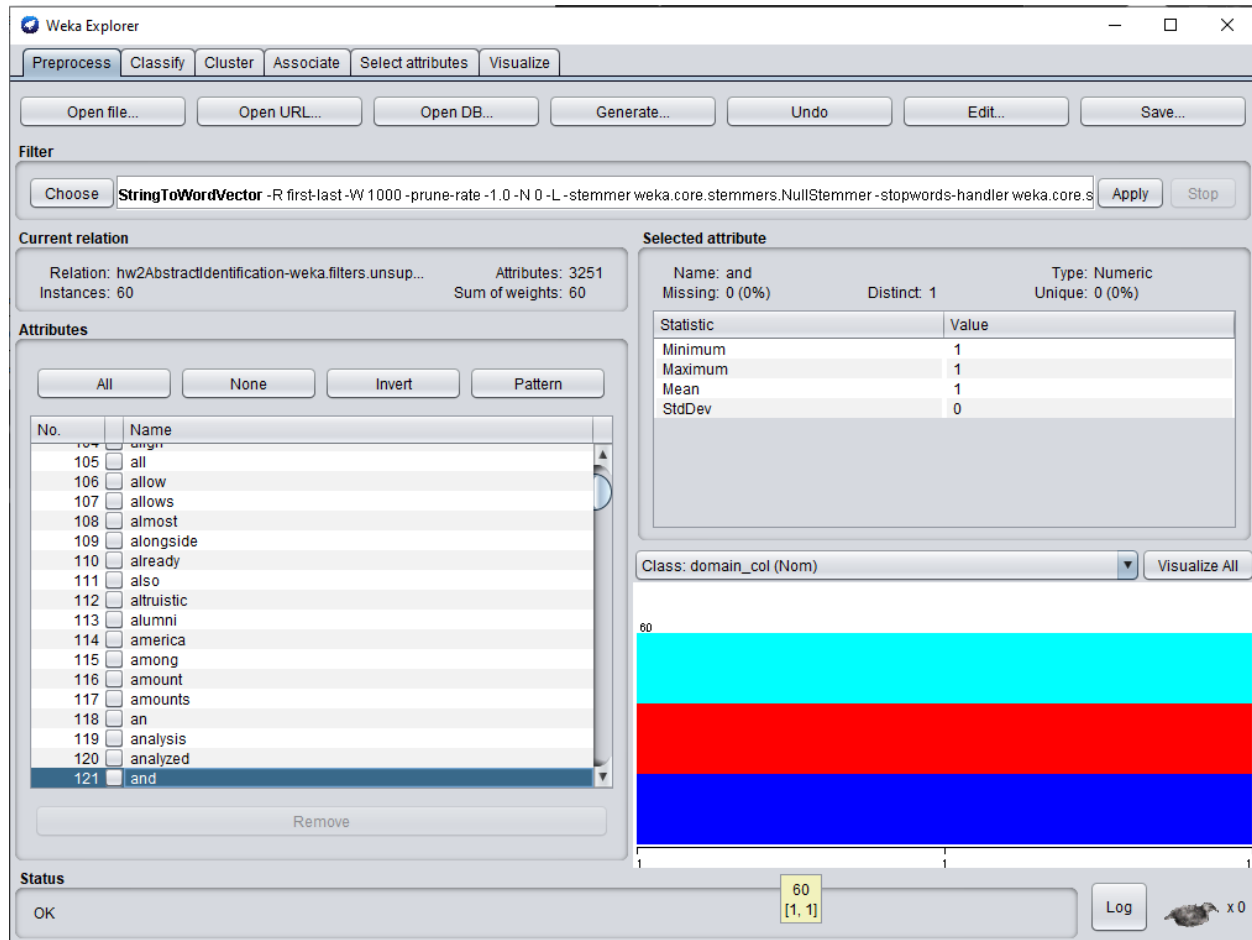
See in the screenshots section below for the comparison between all the Words to Keep variations, as well as the results of all the cluster tests.

Screenshots

Screenshots (Task 1.1 - Binary)

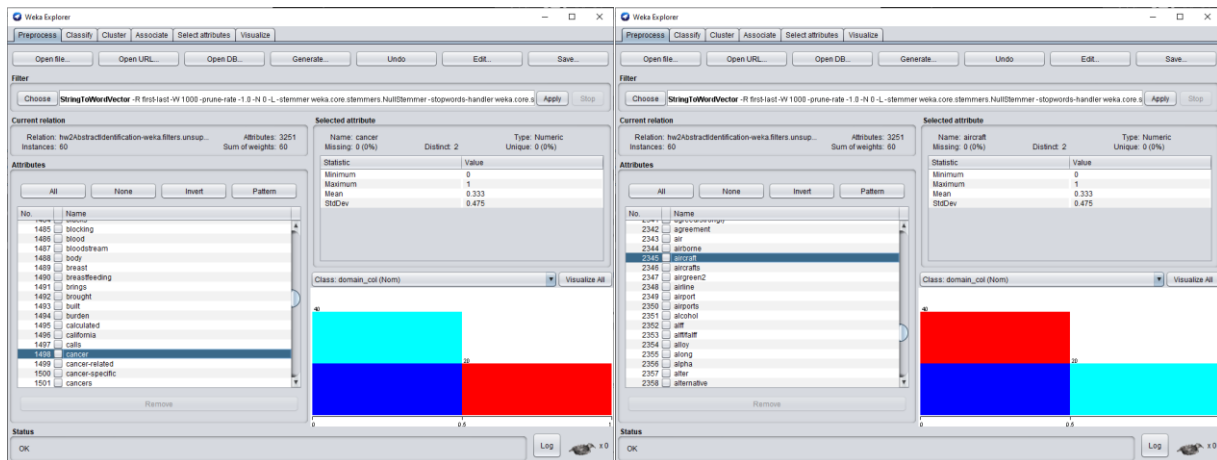
The screenshot shows the Weka Explorer interface. The 'Filter' tab is active, displaying the 'StringToWordVector' filter with parameters: -R first-last -W 1000 -prune-rate -1.0 -N 0 -L -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.s. The 'Current relation' section shows 'Relation: hw2AbstractIdentification-weka.filters.unsup...' with 3251 attributes and 60 instances. The 'Selected attribute' section shows 'Name: computer' with a distinct value of 2. The 'Attributes' list on the left shows various terms like 'complexity', 'component', 'comprehension', etc. The 'Visualize All' button is visible. A bar chart at the bottom right shows the distribution of the selected attribute, with a red bar for 'computer' and a blue bar for 'domain_col (Nom)'.

Word analysis for “computer” showed up in 16 abstracts, 1 time each. Majority of the abstracts being Computer-Science domain (dark blue). Most Breast-Cancer domain (red) and Aircraft domain (light blue) were 0. Upon further analysis, we see that 15 Computer-Science abstracts had the word “computer” in it once, and 1 Aircraft abstract had the word “computer” in it once.

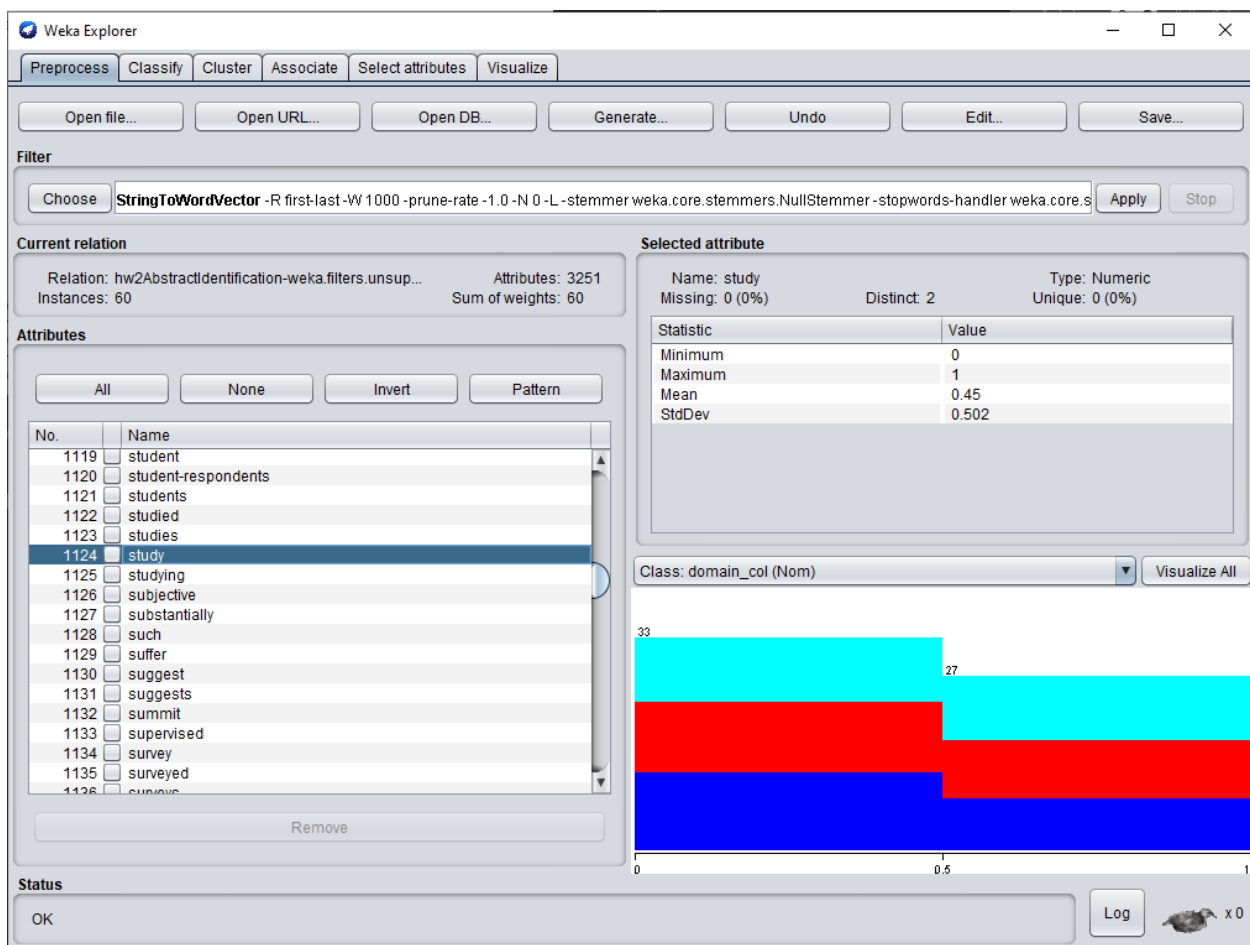


The word “and” seems to appear in 60/60 abstracts presented in the data file. Very interesting... Later analysis showed that this was the same case for the word “the” as well.

Abir Razzak (amr439@drexel.edu)
INFO 371 – Data Mining Applications
Assignment 2

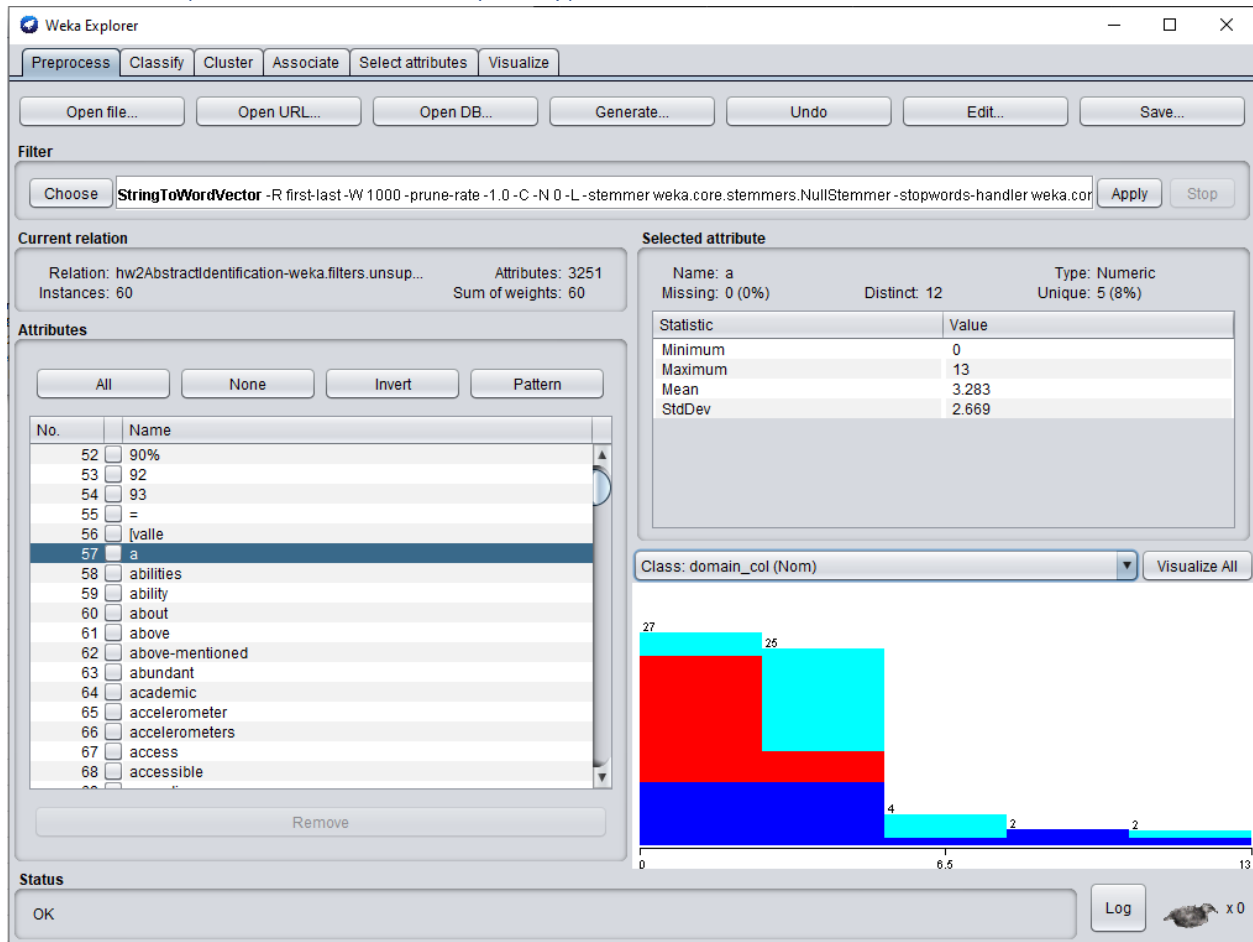


Analysis for the words “cancer” and “aircraft” were clear cut and only appeared in their respective domains. This makes sense for the most part, however it still made for some interesting screenshots.



“Study” appears in 27/60 abstracts, and is almost evenly spread across all 3 domains.

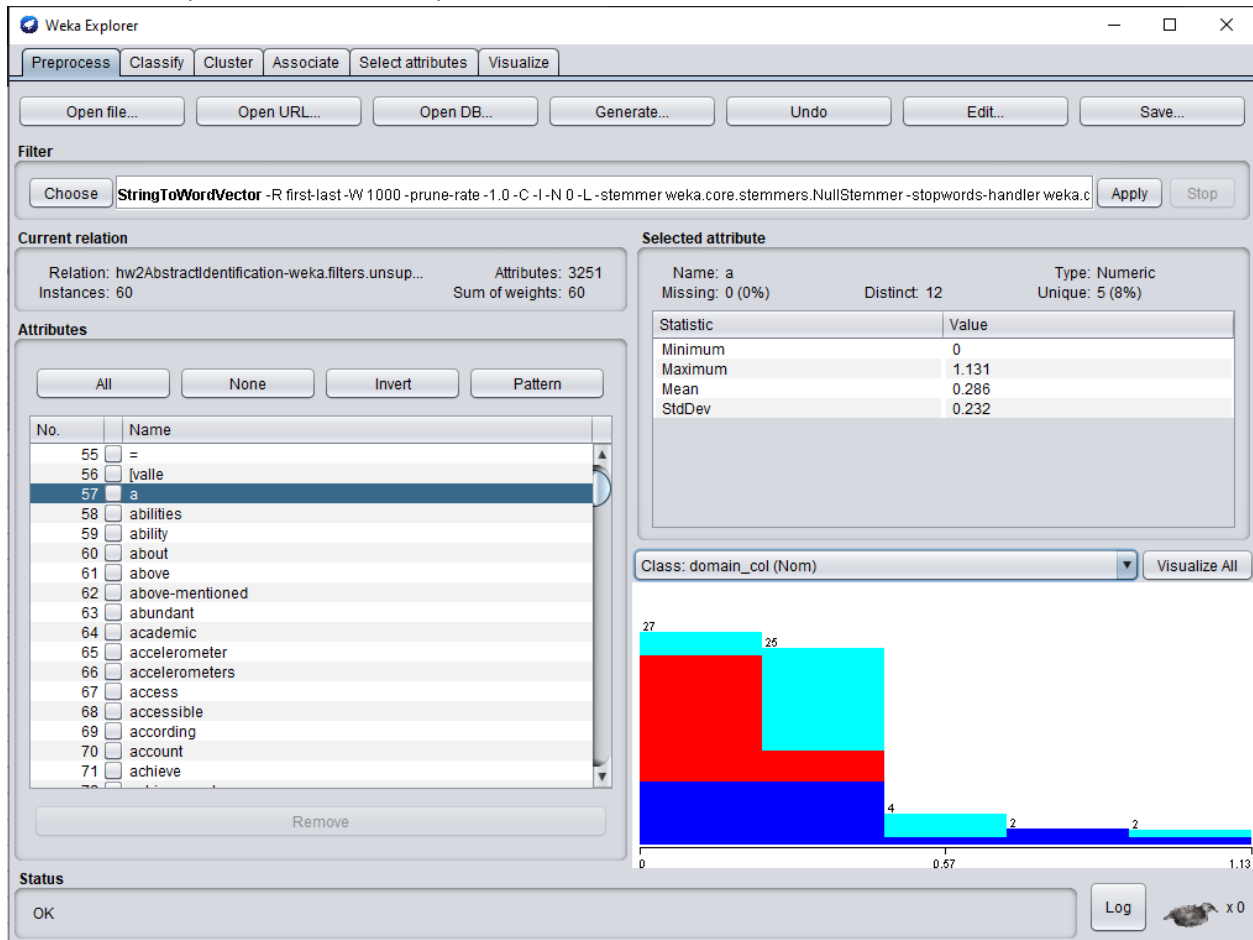
Screenshots (Task 1.2 – Term Frequency)



Dark Blue – Computer Science, Red – Breast Cancer, Light Blue – Aircraft

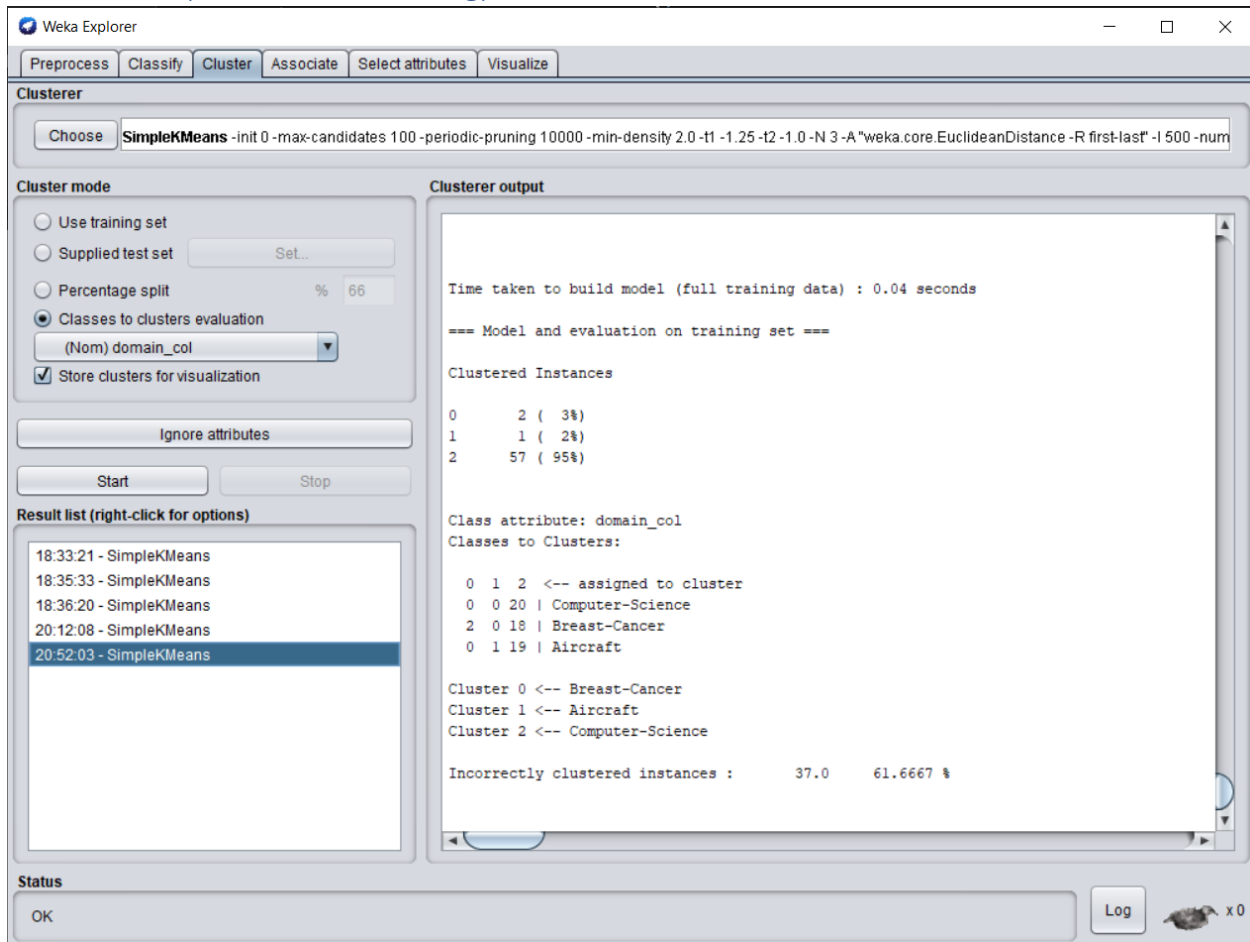
Further analysis of the words from before. This time in term frequency, so the bins are broken down a bit more to show a further breakdown.

Screenshots (Task 1.3 – TF*IDF)



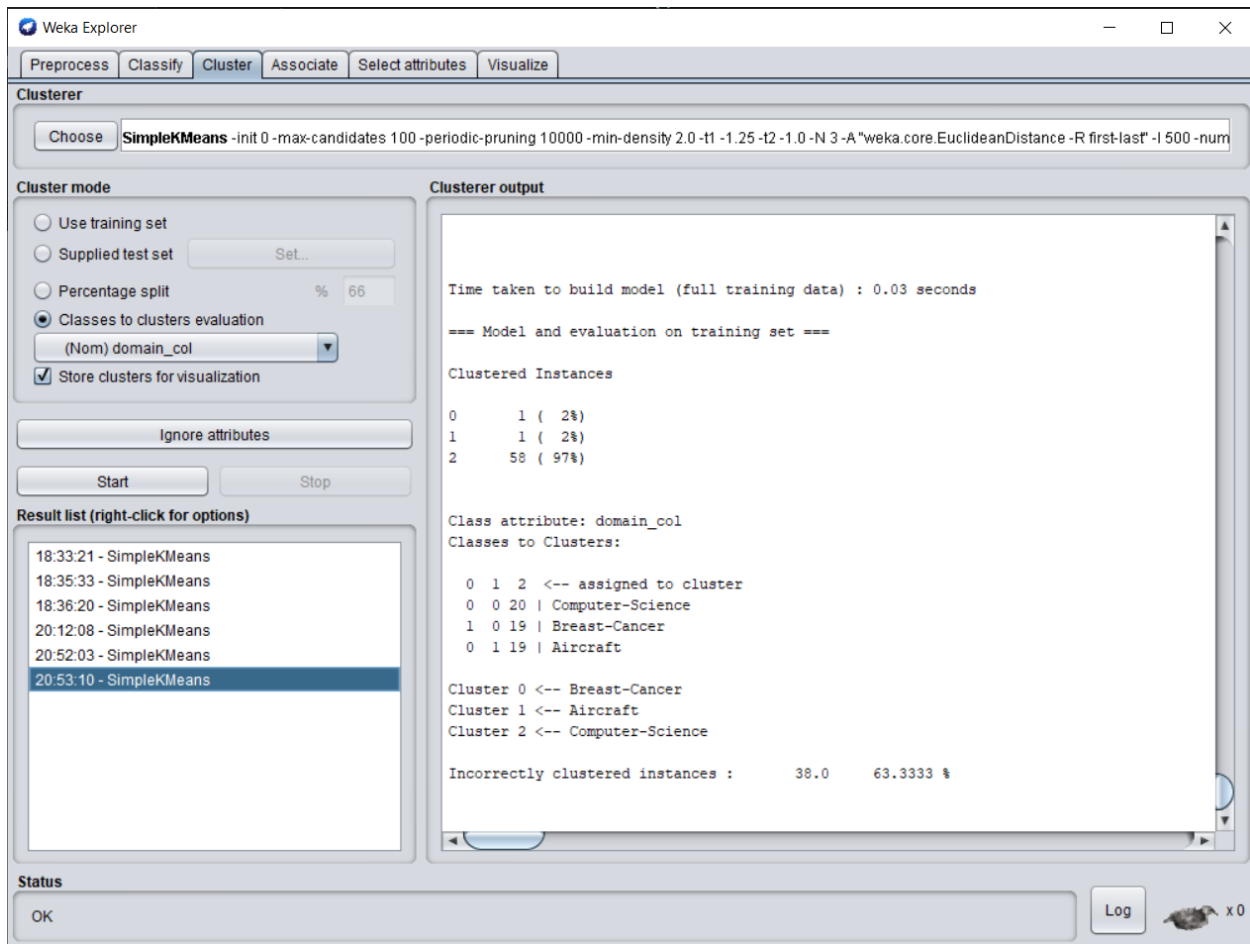
The graph looks identical to the term frequency one for the word “a”.

Screenshots (Task 2.1 – Clustering)



Binary cluster.

Abir Razzak (amr439@drexel.edu)
INFO 371 – Data Mining Applications
Assignment 2



Weka Explorer

Preprocess | **Cluster** | Associate | Select attributes | Visualize

Clusterer

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num

Cluster mode

- ☐ Use training set
- ☐ Supplied test set
- ☐ Percentage split % 66
- ☒ Classes to clusters evaluation
(Nom) domain_col
- ☒ Store clusters for visualization

Result list (right-click for options)

- 18:33:21 - SimpleKMeans
- 18:35:33 - SimpleKMeans
- 18:36:20 - SimpleKMeans
- 20:12:08 - SimpleKMeans
- 20:52:03 - SimpleKMeans
- 20:53:10 - SimpleKMeans**

Cluster output

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

Cluster	Instances	Percentage
0	1	2%
1	1	2%
2	58	97%

Class attribute: domain_col
Classes to Clusters:

Cluster	Instances	Assigned to cluster
0	1	2
0	0	20
1	0	19
0	1	19

Cluster 0 <-- Breast-Cancer
Cluster 1 <-- Aircraft
Cluster 2 <-- Computer-Science

Incorrectly clustered instances : 38.0 63.3333 %

Status

OK x 0

Term frequency cluster.

Abir Razzak (amr439@drexel.edu)
INFO 371 – Data Mining Applications
Assignment 2

The screenshot shows the Weka Explorer Clusterer window. The 'Clusterer' tab is selected, and 'SimpleKMeans' is chosen. The 'Cluster mode' section shows 'Classes to clusters evaluation' selected with 'domain_col' as the class attribute. The 'Cluster output' pane displays the following results:

```
Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1 ( 2%)
1      1 ( 2%)
2     58 ( 97%)

Class attribute: domain_col
Classes to Clusters:

0 1 2 <-- assigned to cluster
0 0 20 | Computer-Science
1 0 19 | Breast-Cancer
0 1 19 | Aircraft

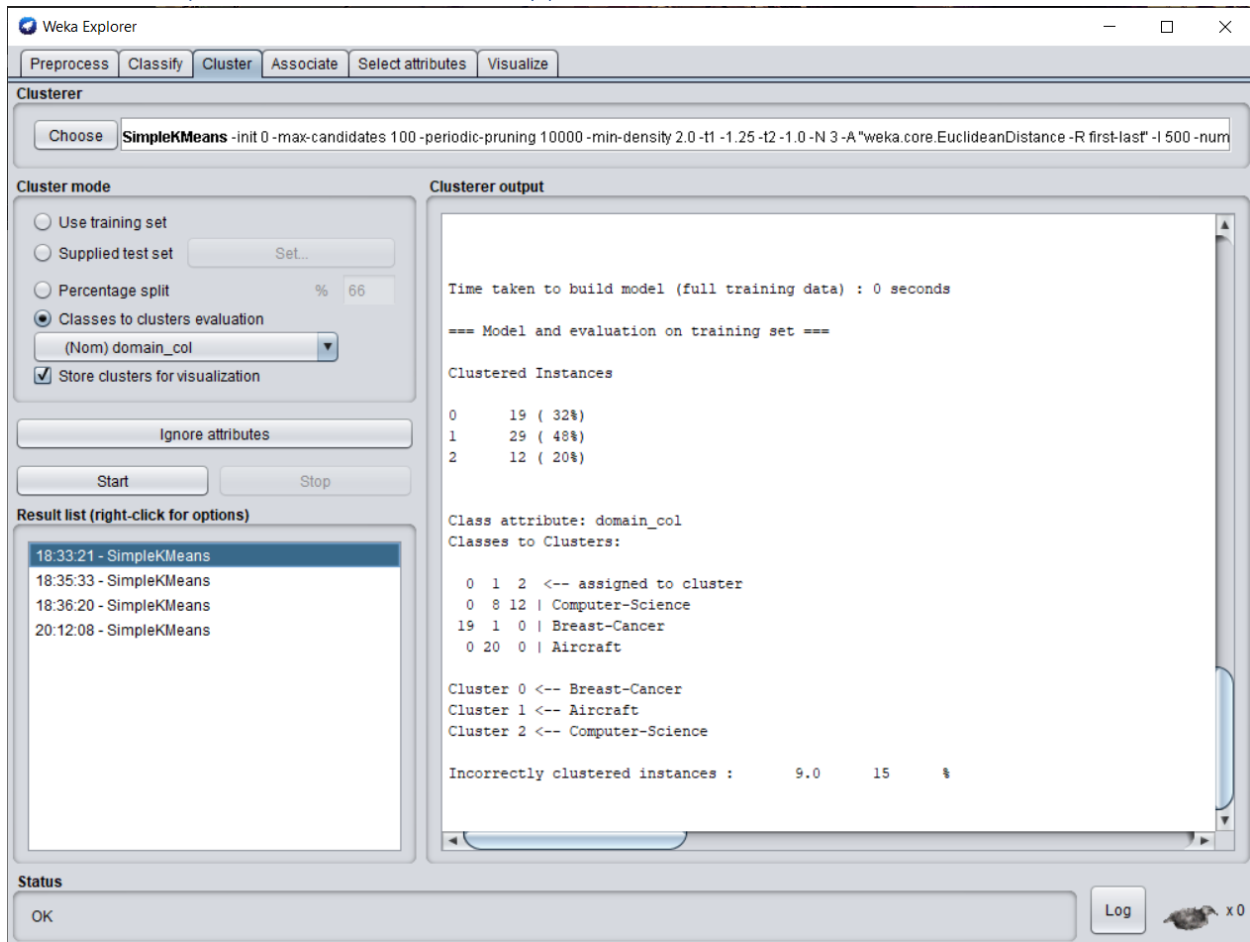
Cluster 0 <-- Breast-Cancer
Cluster 1 <-- Aircraft
Cluster 2 <-- Computer-Science

Incorrectly clustered instances :      38.0      63.3333 %
```

The 'Result list' on the left shows a list of SimpleKMeans runs, with the most recent one (20:53:38) selected. The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

TF*IDF cluster.

Screenshots (Task 2.2 – Words to Keep)



10 words to keep. For the most part very accurate. Computer-Science seemed to be the hardest to categorize here.

Abir Razzak (amr439@drexel.edu)
INFO 371 – Data Mining Applications
Assignment 2

The screenshot shows the Weka Explorer Clusterer window. The 'Clusterer' tab is selected, and 'SimpleKMeans' is chosen. The 'Cluster mode' section shows 'Classes to clusters evaluation' selected with '(Nom) domain_col' as the evaluation attribute. The 'Cluster output' pane displays the following results:

```
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      8 ( 13%)
1     13 ( 22%)
2     39 ( 65%)

Class attribute: domain_col
Classes to Clusters:

0 1 2 <-- assigned to cluster
0 0 20 | Computer-Science
8 0 12 | Breast-Cancer
0 13 7 | Aircraft

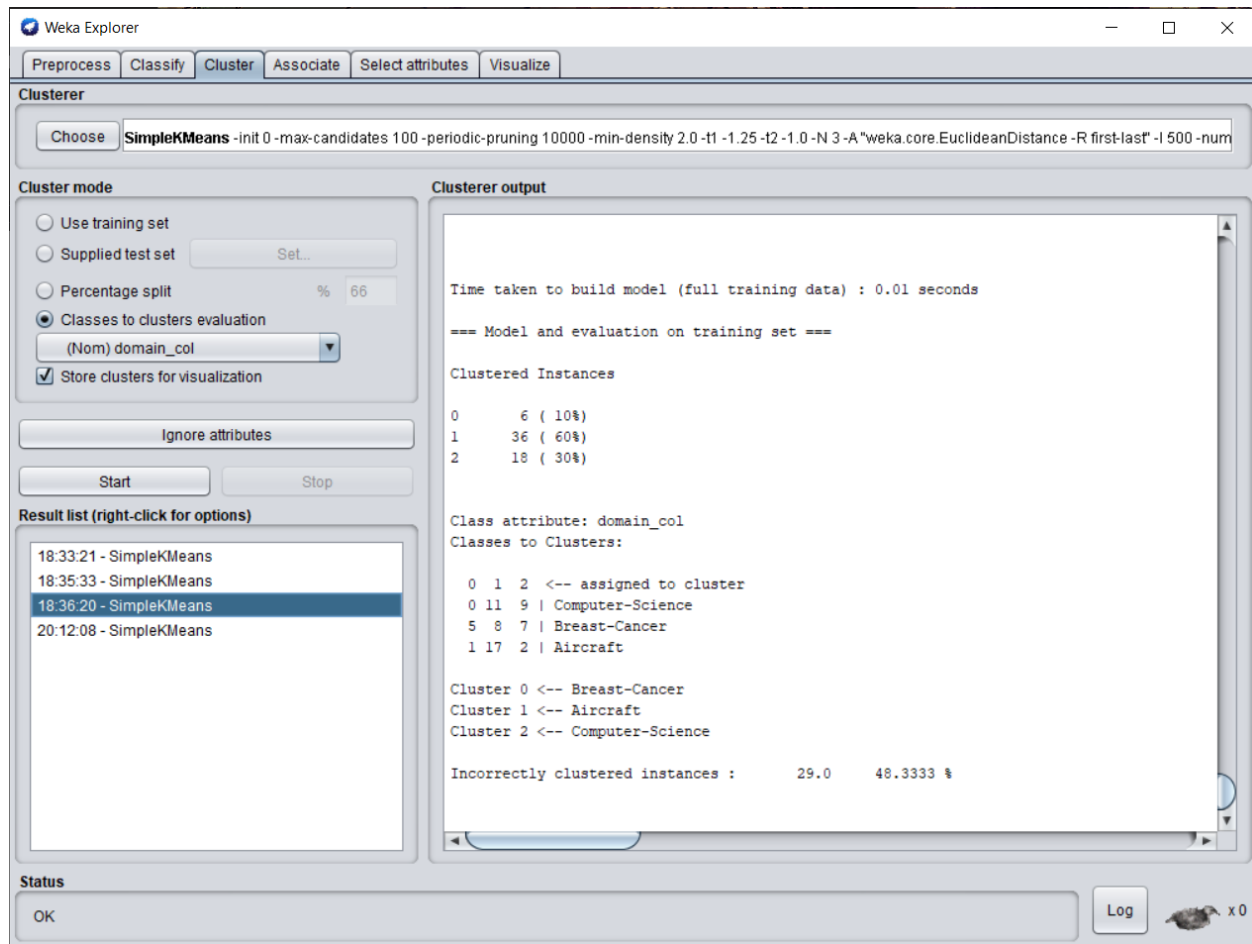
Cluster 0 <-- Breast-Cancer
Cluster 1 <-- Aircraft
Cluster 2 <-- Computer-Science

Incorrectly clustered instances :      19.0      31.6667 %
```

The 'Result list' on the left shows four entries for SimpleKMeans, with the second entry (18:35:33) selected. The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

40 words to keep. Computer-Science all of a sudden has a 100% accuracy, but Breast-Cancer's accuracy is thrown out the window, only correctly being clustered for 8/20 abstracts. Aircraft isn't looking as bad, with a correct cluster of 13/20, however in both cases of Breast-Cancer and Aircraft, both seem to be losing accuracy by incorrectly being identified as Computer-Science. The difference in words between 10 WtK (words to keep) and 40 WtK seemed to have swung into Computer-Science's favor. Perhaps many of the words found in Aircraft and Breast-Cancer abstracts seemed to have overlapped with words included in many Computer-Science abstracts.

Abir Razzak (amr439@drexel.edu)
INFO 371 – Data Mining Applications
Assignment 2



The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' section shows 'SimpleKMeans' as the chosen algorithm. The 'Cluster mode' section has 'Classes to clusters evaluation' selected, with '(Nom) domain_col' as the class attribute. The 'Cluster output' section shows the results of the clustering process.

Cluster mode

- ☐ Use training set
- ☐ Supplied test set
- ☐ Percentage split % 66
- ☒ Classes to clusters evaluation
- ☐ Store clusters for visualization

Cluster output

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      6 ( 10%)
1     36 ( 60%)
2     18 ( 30%)

Class attribute: domain_col
Classes to Clusters:

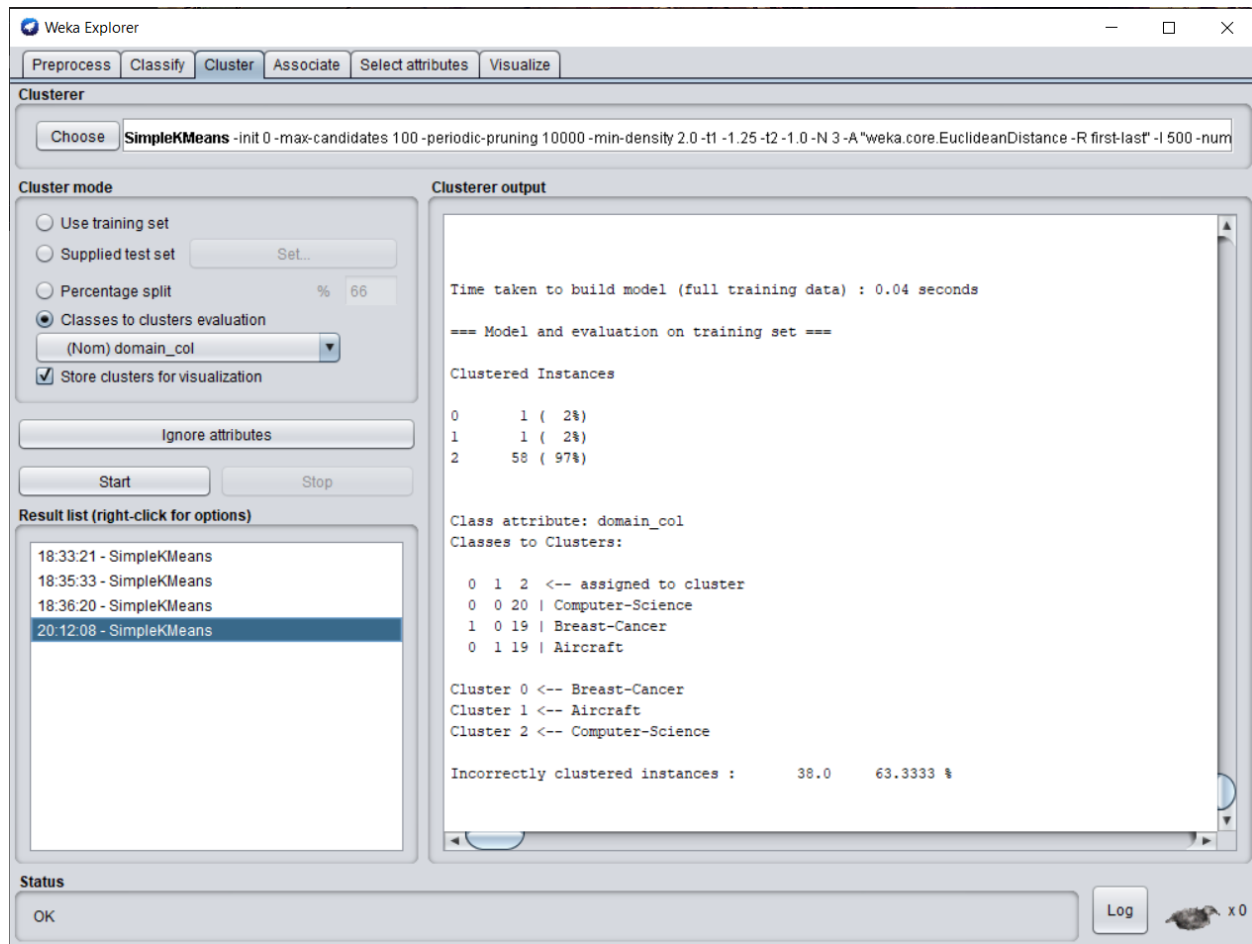
0 1 2 <-- assigned to cluster
0 11 9 | Computer-Science
5 8 7 | Breast-Cancer
1 17 2 | Aircraft

Cluster 0 <-- Breast-Cancer
Cluster 1 <-- Aircraft
Cluster 2 <-- Computer-Science

Incorrectly clustered instances :      29.0      48.3333 %
```

160 WtK. Breast-Cancer is getting more difficult to cluster, as seen in the screenshot Weka had a very difficult time choosing where the abstracts belonged to. We see a spread of 5/8/7, where the clusters are Breast-Cancer/Aircraft/Computer-Science. We actually see an improvement in Aircraft from 40WtK, going from 13/20 to 17/20 correctly identified. We see a big shift in the Computer-Science go towards the Aircraft domain, where its spread was 0/11/9. Seems to be the phenomena we observed in 40WtK except, this time its Aircraft that is drawing in more abstracts towards it.

Abir Razzak (amr439@drexel.edu)
INFO 371 – Data Mining Applications
Assignment 2



640 WtK. We observed here what we saw in the 40WtK, however in a more potent form. All but 2 abstracts were clustered into Computer-Science. However, the 2 that were not clustered into Computer-Science were correctly clustered into the domains they belonged in. This resulted in 22 correct instances for the abstracts, or 38 incorrect instances, giving us a rate of 63.33% incorrectness. Like mentioned previously in the 40WtK screenshot, only explanation for this is that Computer-Science had a lot of overlap with the other two domains, where Breast-Cancer and Aircrafts did not overlap with each other quite as much. This must have caused confusion in the machine learning of Weka and caused most of the abstracts to be analyzed as Computer-Science.