Man Tik Li (ml3546)
INFO 371 – Data Mining Applications
Assignment 2

Abstracts

Abstract.arff

Number (Line)	Category	Source
15-34	AI	Artificial Intelligence: A Modern Approach (Third) – Stuart Russell
35-54	STAT	Introductory Biological Statistics (Fourth) – John E. Havel
55-74	ENVS	New Age Environmental Science – Dr. Y. K. Singh

Task 1

Selected Words

Stop Words:

- "at" (No. 173)
- "like" (No. 832)
- "it" (No. 777)

Rare Words:

- "ai" (No. 101)
- "environment" (No. 516)
- "method" (No. 903)

Other Words:

- "network" (No. 958)
- "developers" (No. 437)
- "project" (No. 1121)

Data Table (Task 1 - Binary)

	Feature/attribute weight								
Paper# (domain)	at	like	it	ai	environments	method	network	developers	computer
2 (AI)	1	1	1	1	1	1	1	1	1
22 (STAT)	1	1	1	0	0	1	0	0	1
42 (ENVS)	1	1	1	0	1	0	1	0	0

Data Table (Task 2 – Term Frequency)

	Feature/attribute weight								
Paper# (domain)	at	like	it	ai	environments	method	network	developers	computer
2 (AI)	10	2	12	9	11	1	4	1	1
22 (STAT)	8	5	12	0	0	2	0	0	1
42 (ENVS)	9	3	13	0	9	0	1	0	0

Data Table (Task 3 – TF*IDF)

	Feature/attribute weight								
Paper# (domain)	at	like	it	ai	environments	method	network	developers	computer
2(AI)	0	0	0	4.294091	1.937004	0.1760913	0.704365	0.4771213	0.1760912
22(STAT)	0	0	0	0	0	0.35218252	0	0	
42(ENVS)	0	0	0	0	1.5848213	0	0.1760913	0	0

Data Table Analysis

Data table 1:

The first table is the binary table, all the values were between 0 and 1. It's used to determine the word in abstract.

Data table 2:

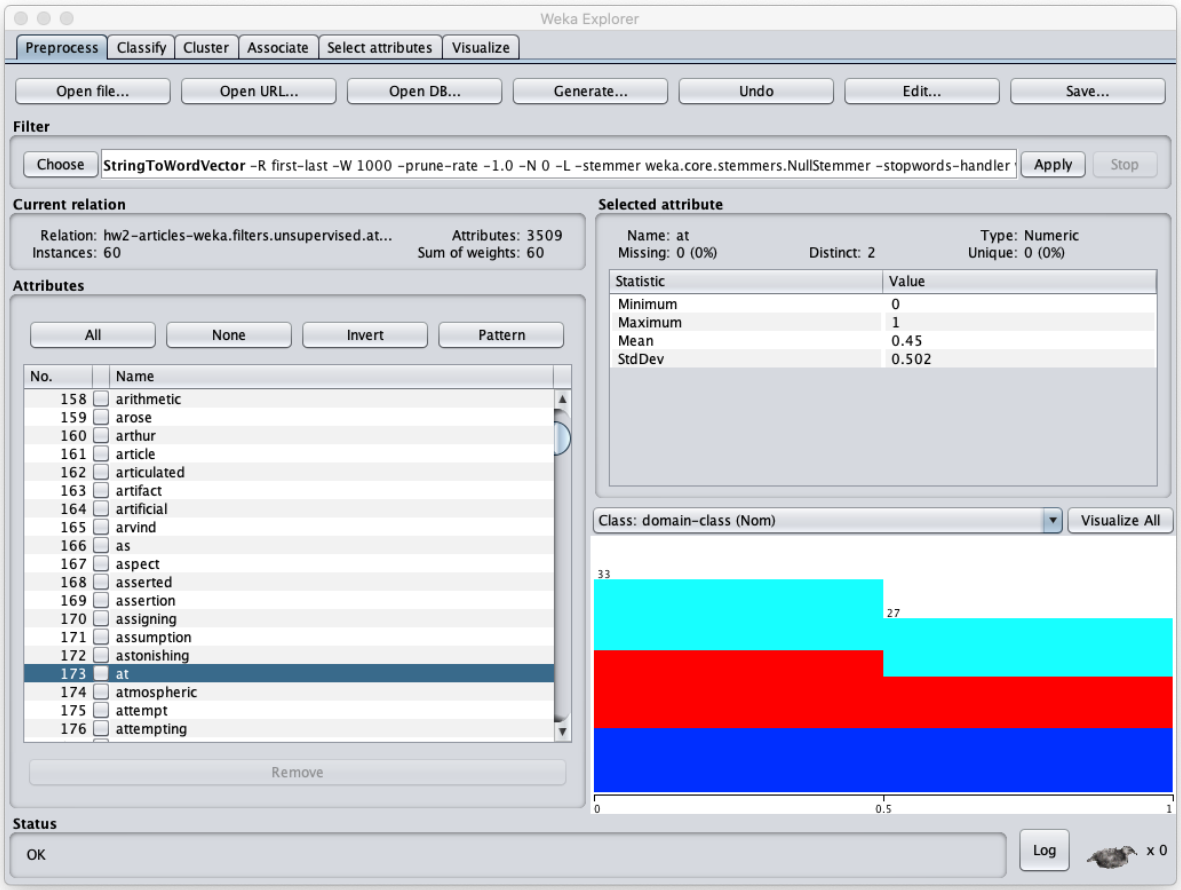
The second table is the term frequency, it used to count for the words in the abstract. As the data shown, the stop-words had significant higher number than the rest of the words. It is because abstract using a lot of stop-words, such as "a", "it", "their", "at", "like", etc.

Data table 3:

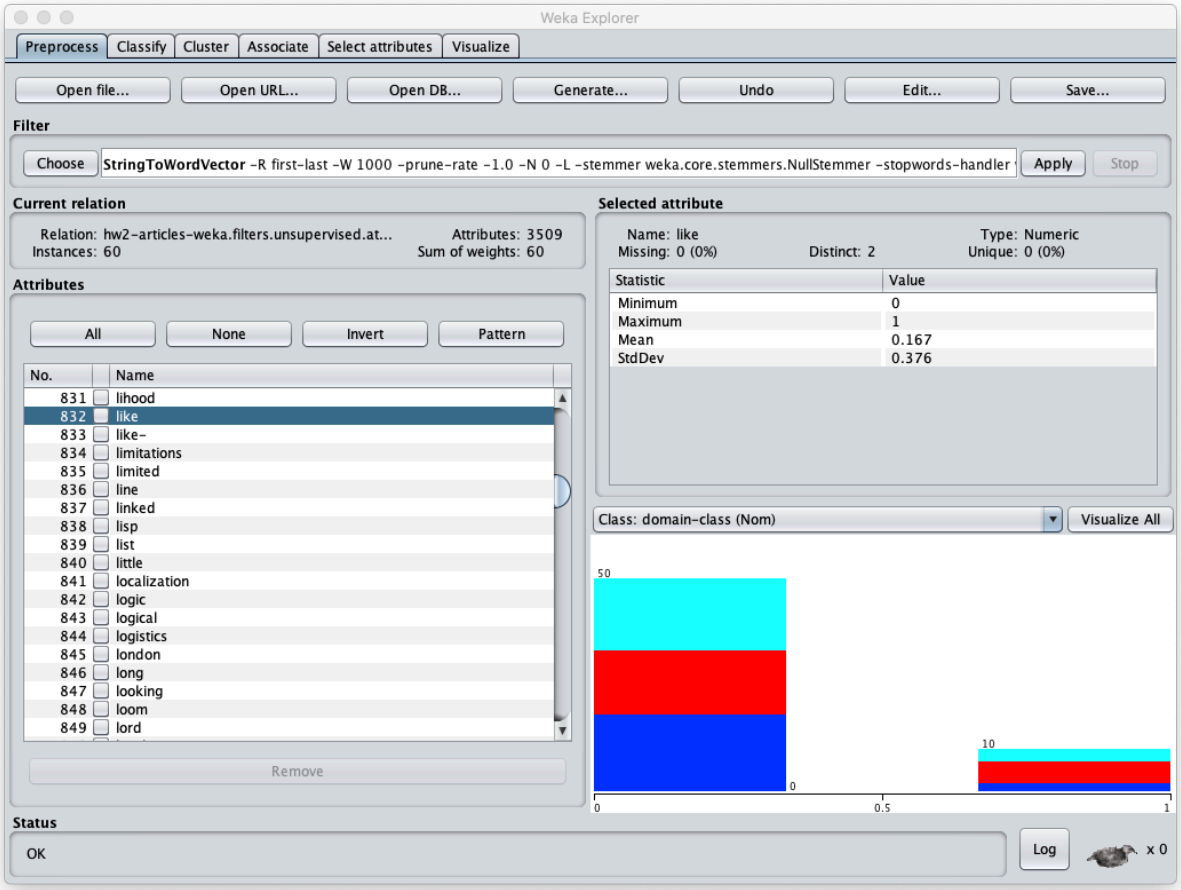
For the TF*IDF, we can see the rarity of the term in the collection. All the stop-words has 0 value, which opposite to the values from term frequency table. It also means the rare words are valued much higher. From the data we can tell that the words "at", "like", and "it" have a inverse document frequency (IDF) of close to zero.

Screenshots:

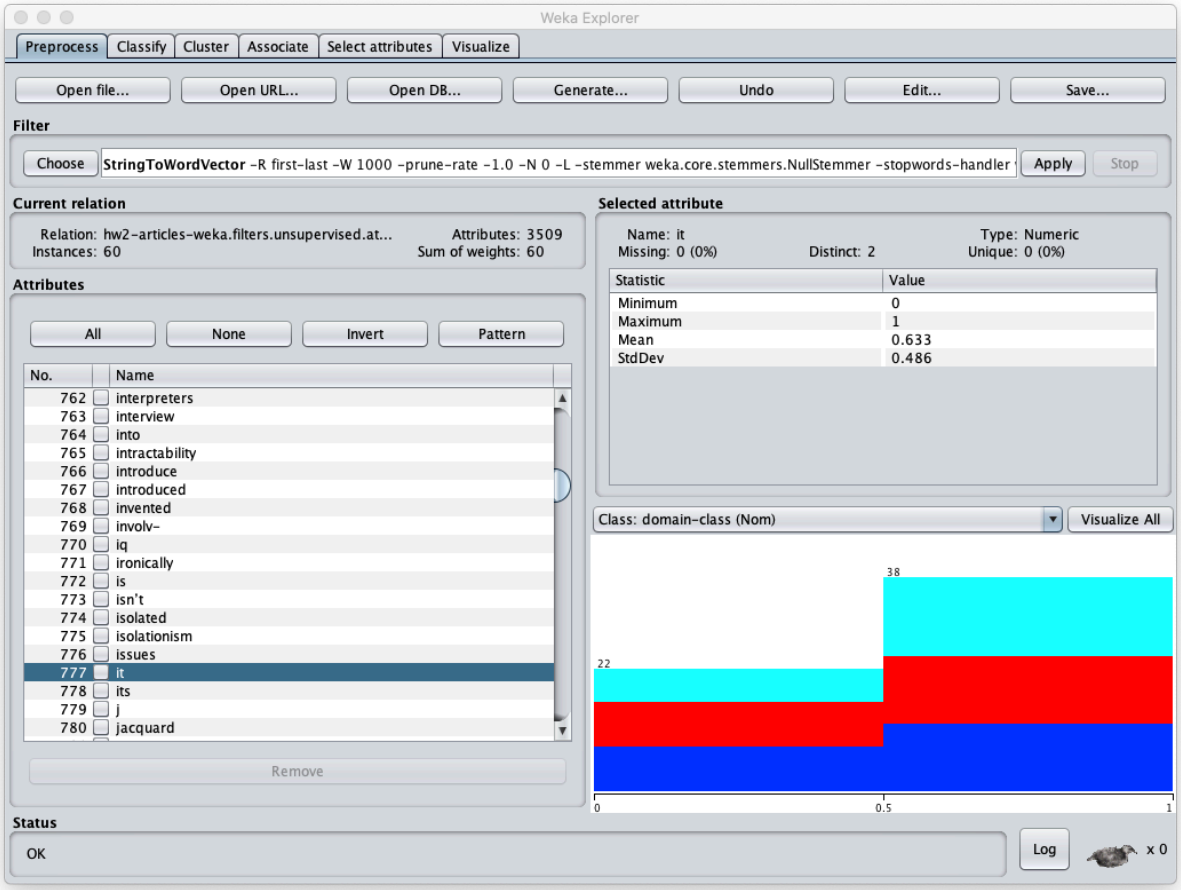
Screenshots



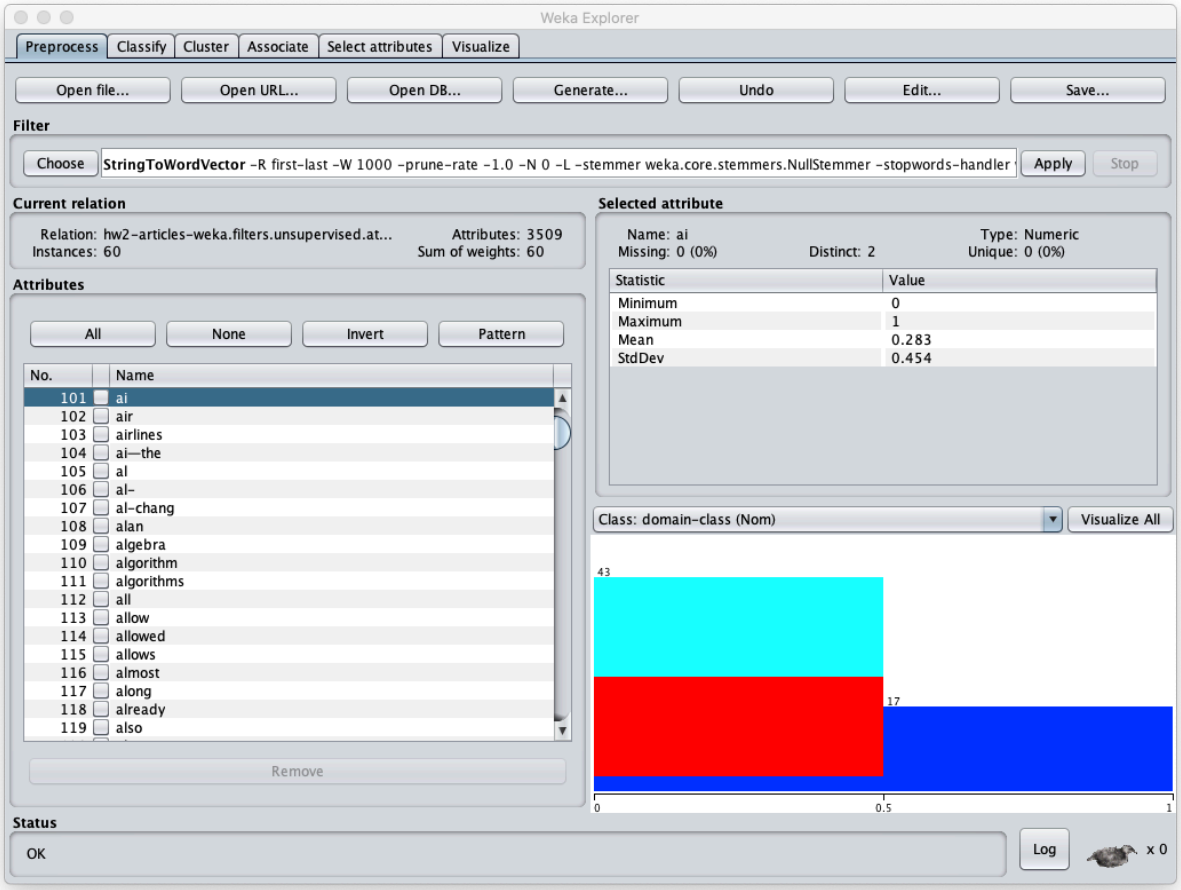
The word “at” appear in most abstract.



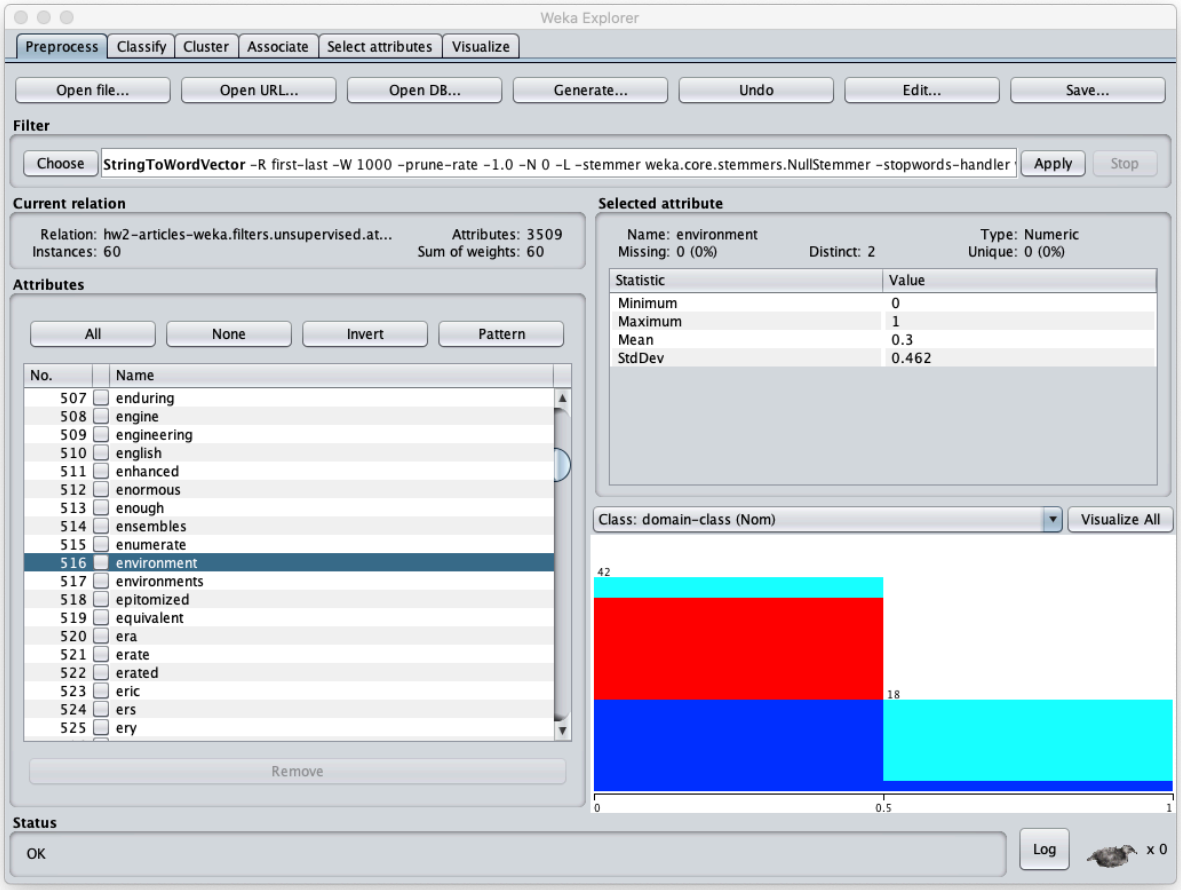
The word “like” appears the most in STAT and only appear few in AI and ENVS.



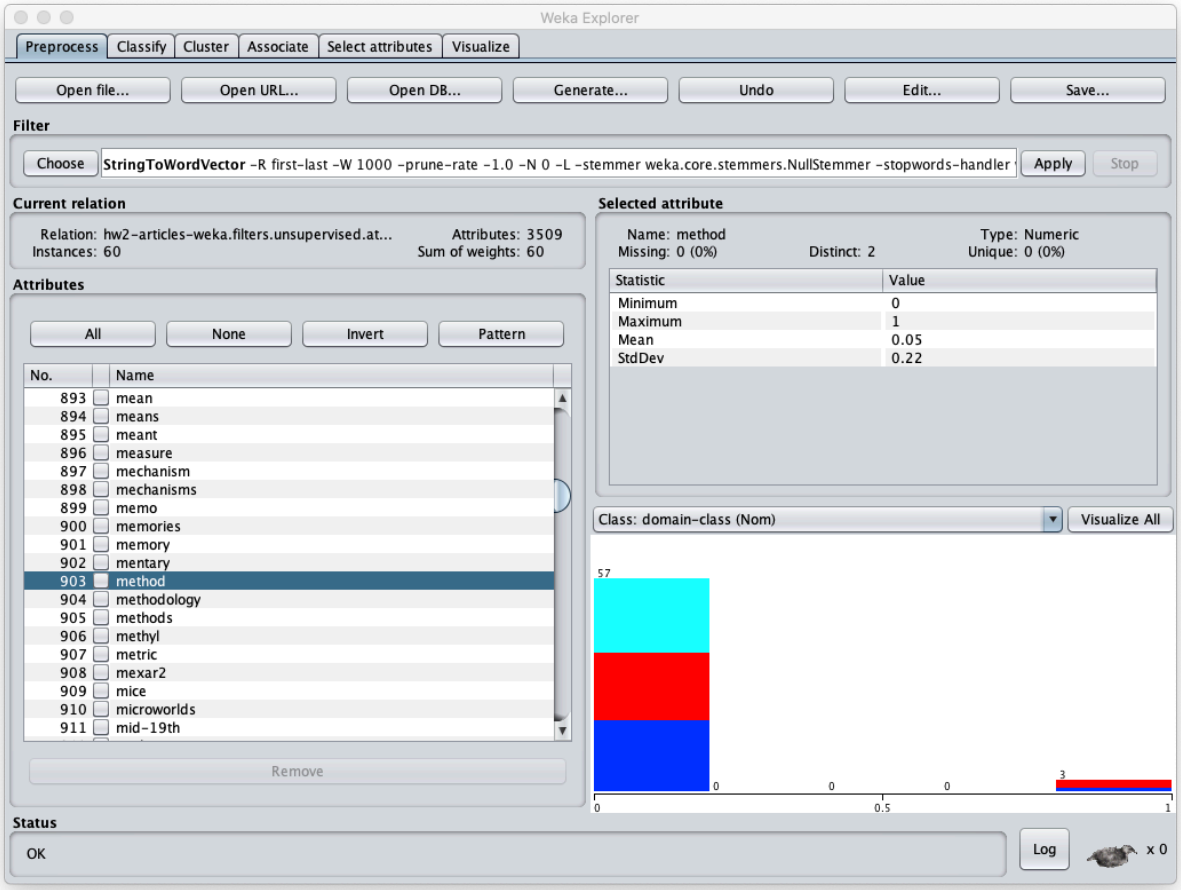
The word “it” appears in most of the domains.



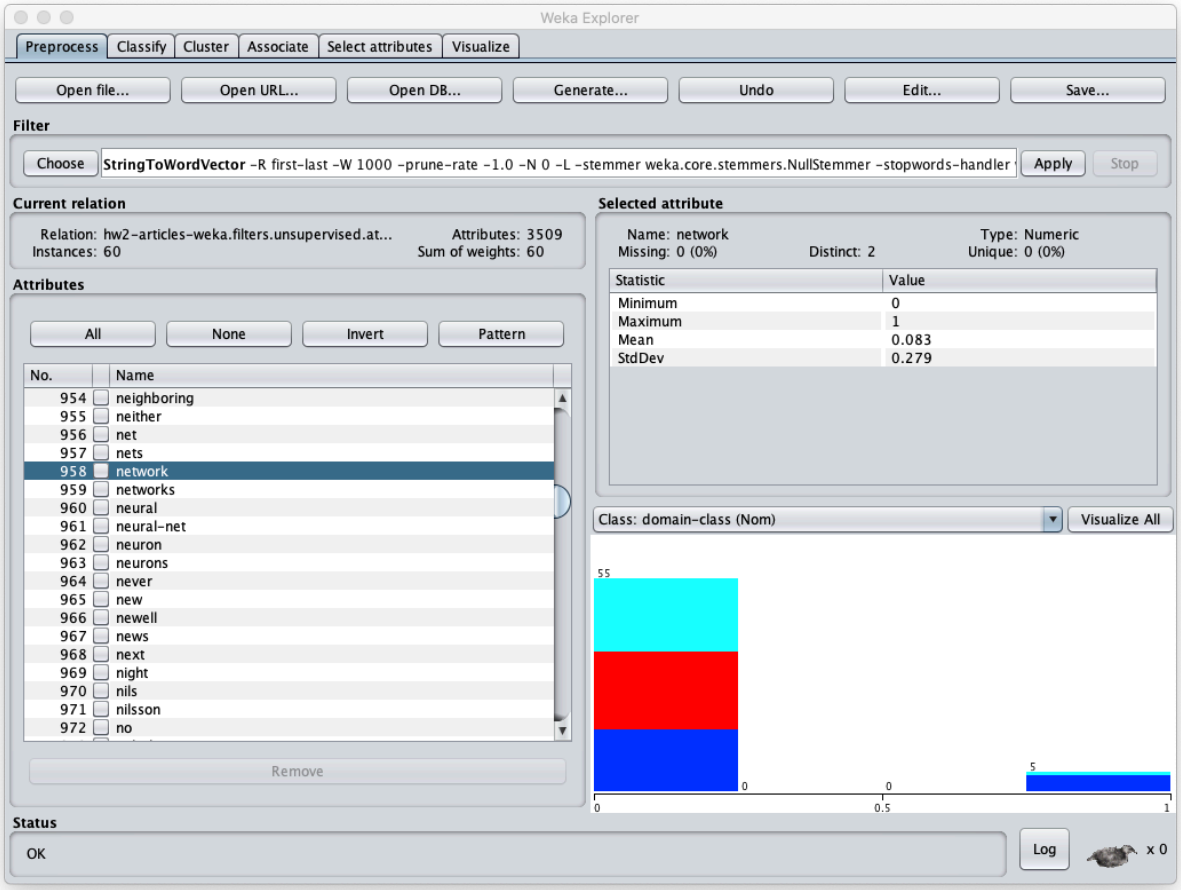
The word “ai” only appears in AI domain.



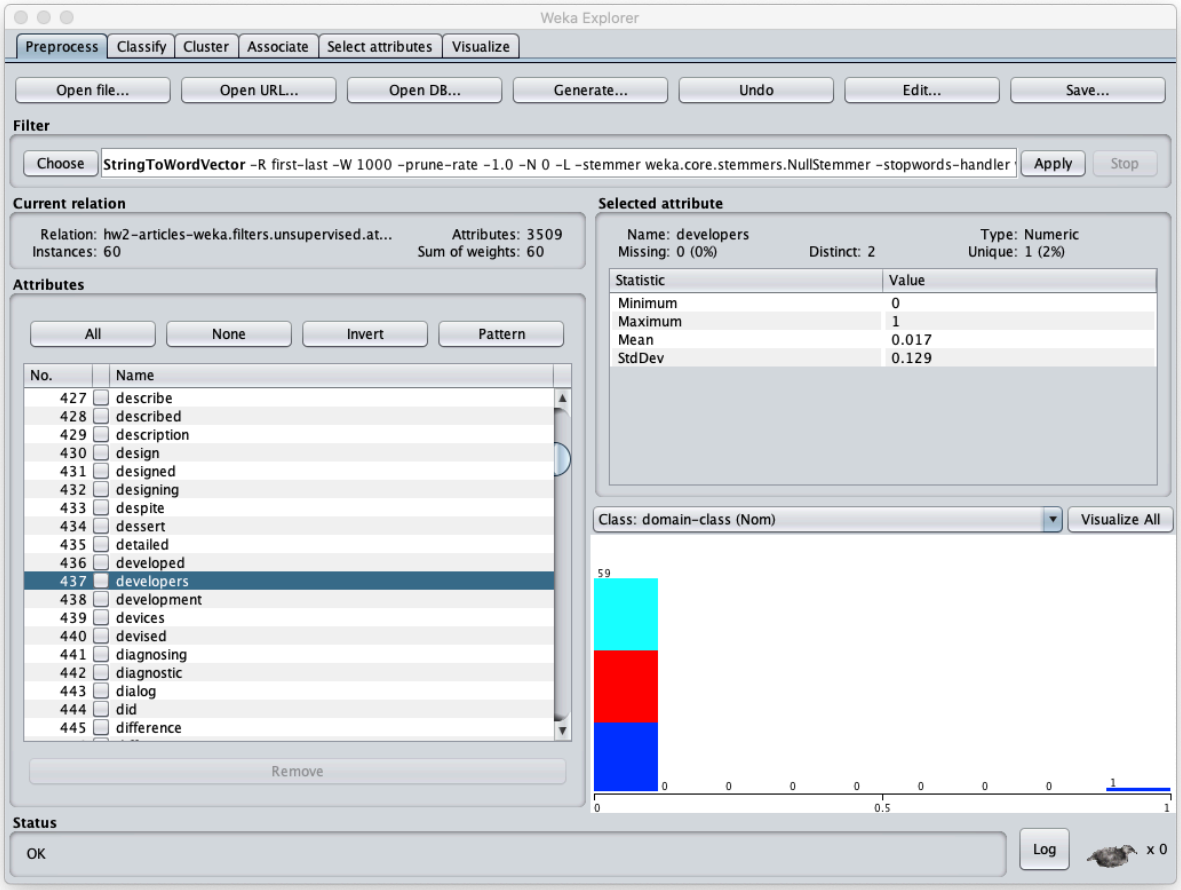
The word “environment” appears once in AI and most in ENVS.



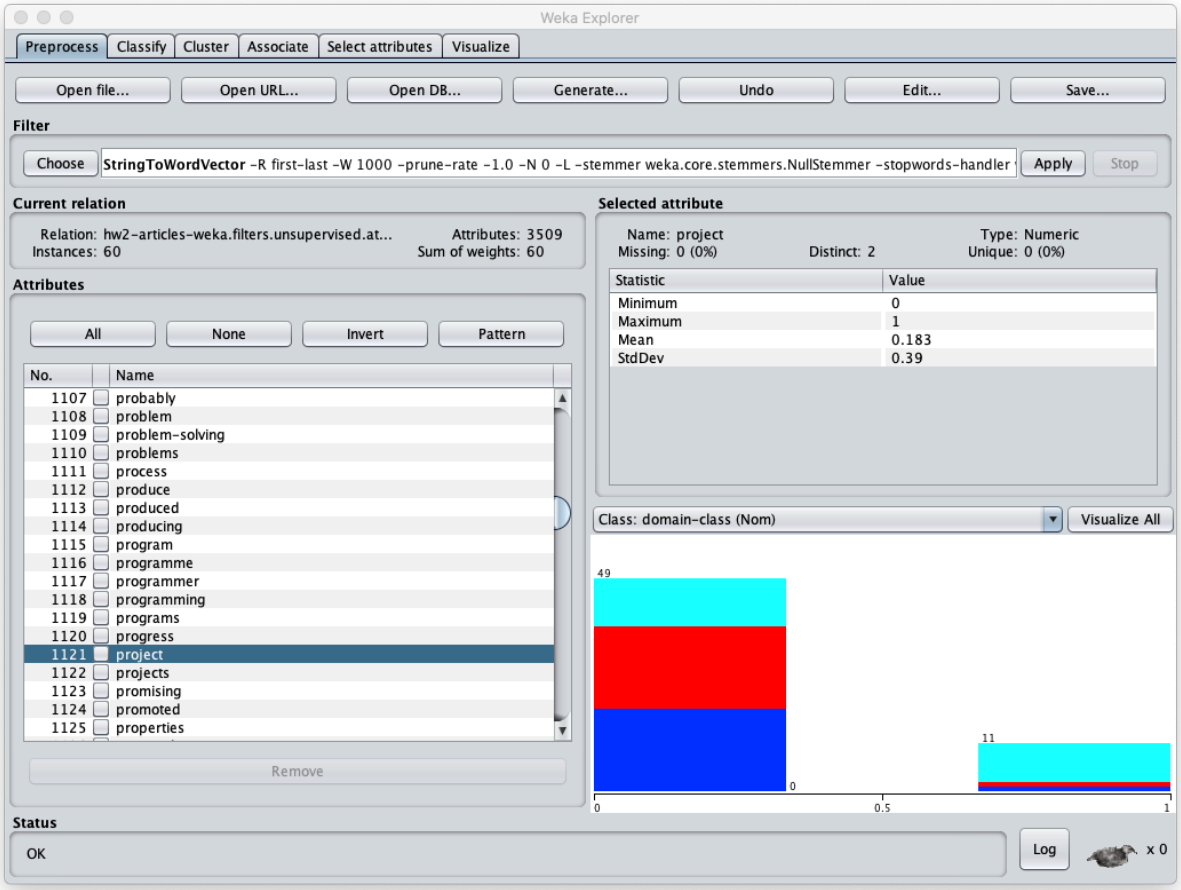
The words “method” only appears 3 times. 2 in STAT and 1 in AI.



The word “network” appears in most in AI domains.



The word “developers” on appears one in AI.



The word “project” appears the most in ENVS.

Task 2

Cluster Summarization Table

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

Words to Keep TF*IDF Summarization Table

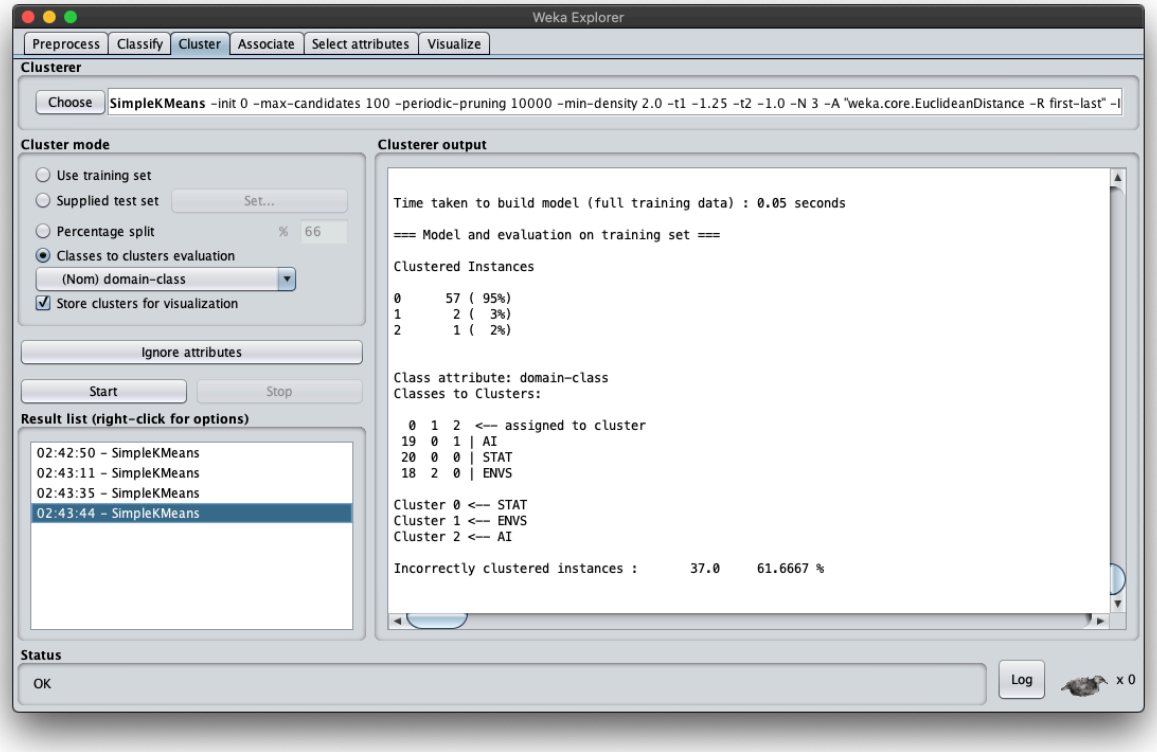
	Clustering effectiveness with different number of attributes			
Words to Keep	10	40	160	640
# Attributes	32	120	500	2968
Incorrectly clustered	25%	31.6667%	51.6667%	63.3333%

Task 2 Analysis

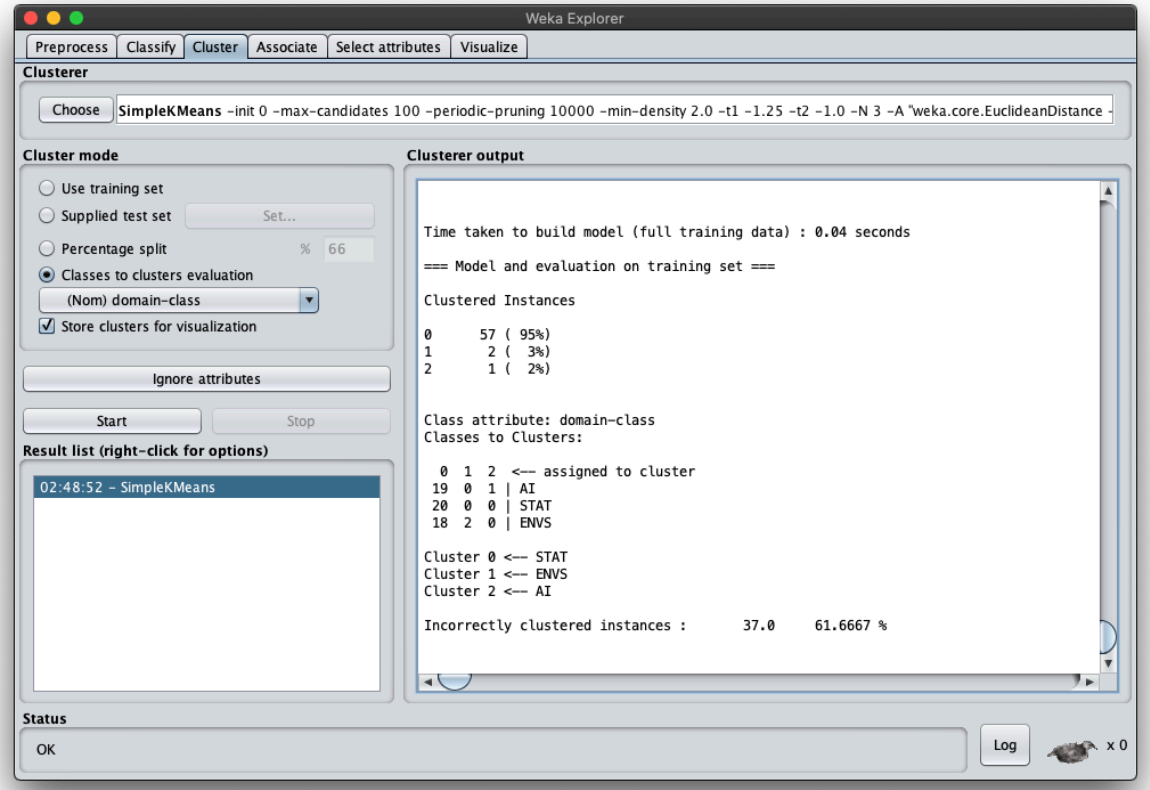
The outcomes of the binary, TF, and TF*IDF were unexpected. The incorrectly percentage get increasing when the number of attributes get higher. Moreover, most of the matrix were clustered into STAT, 2 other abstracts didn't be categorized.

Screenshots

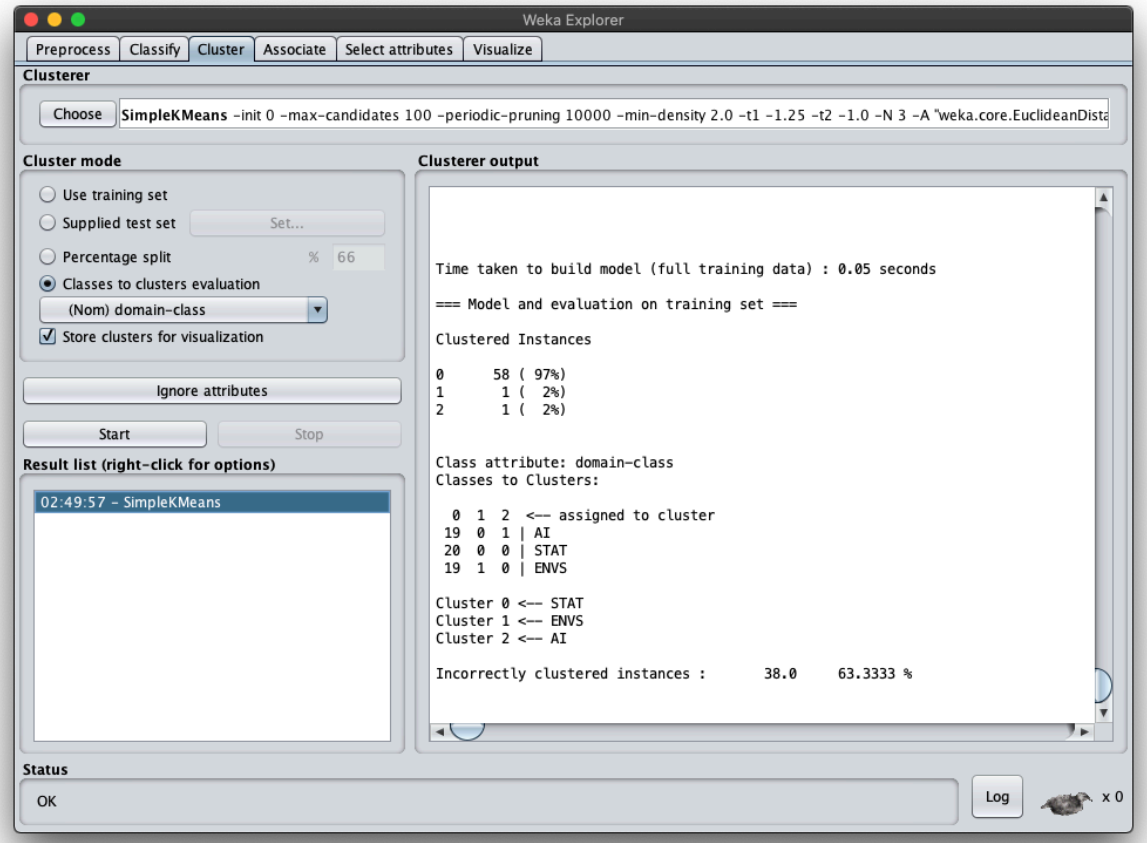
Task 2.1



Binary Cluster

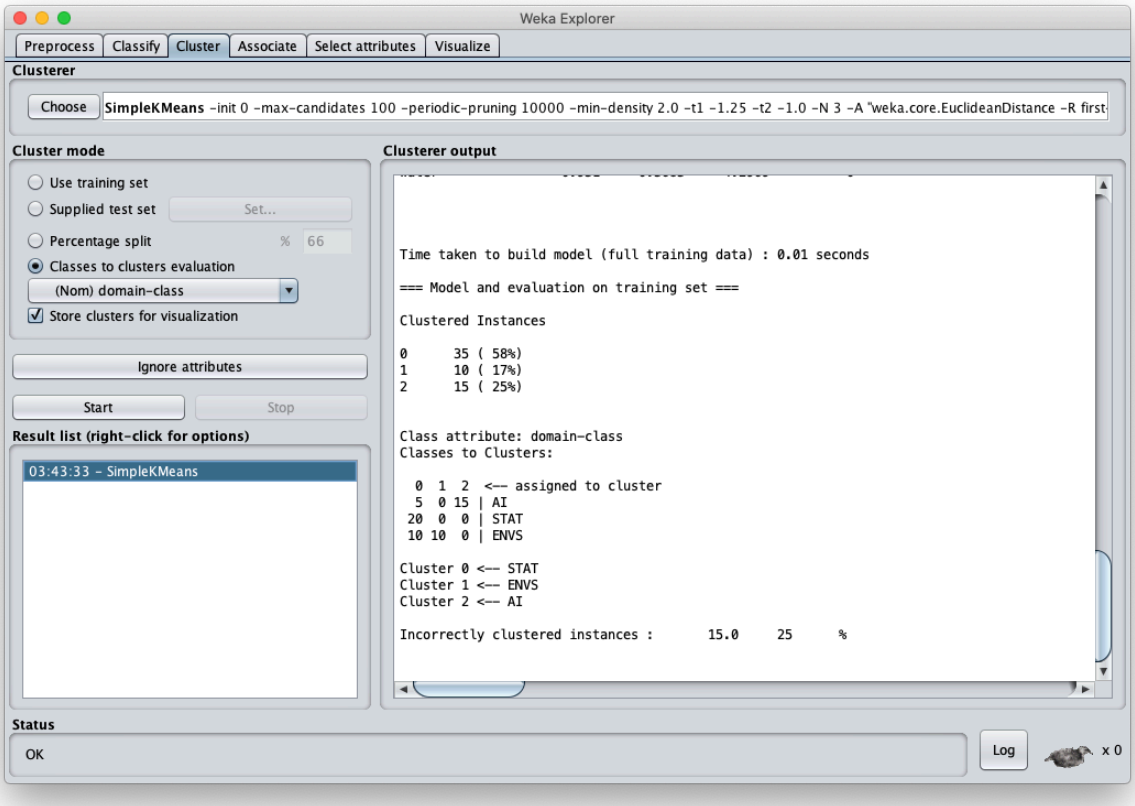


Term Frequency Cluster

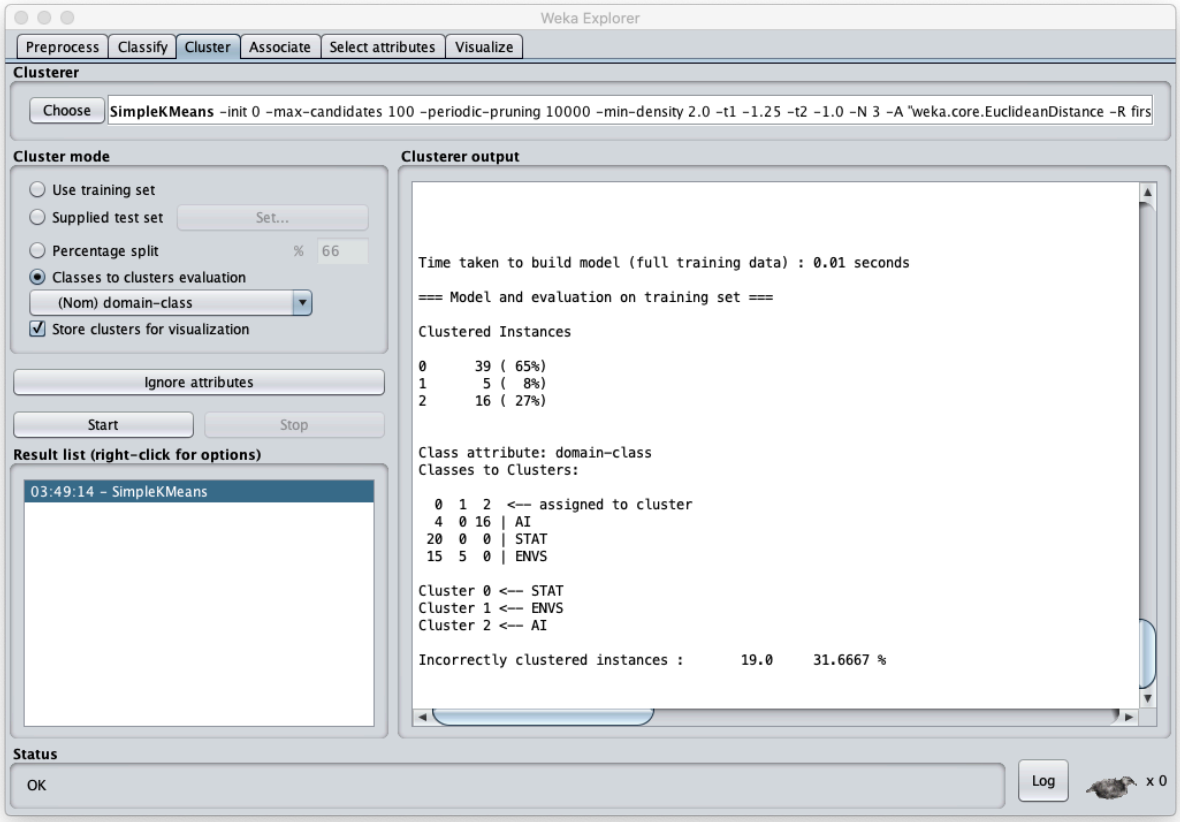


TF*IDF Cluster

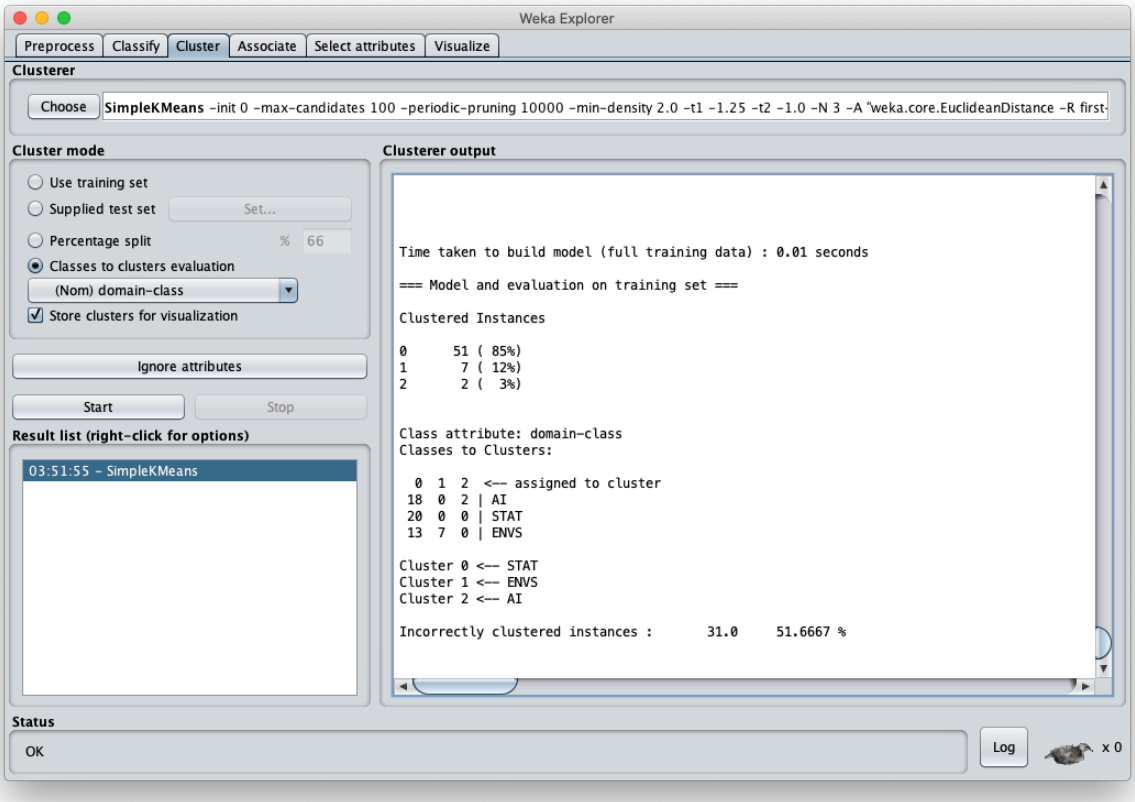
Task 2.2



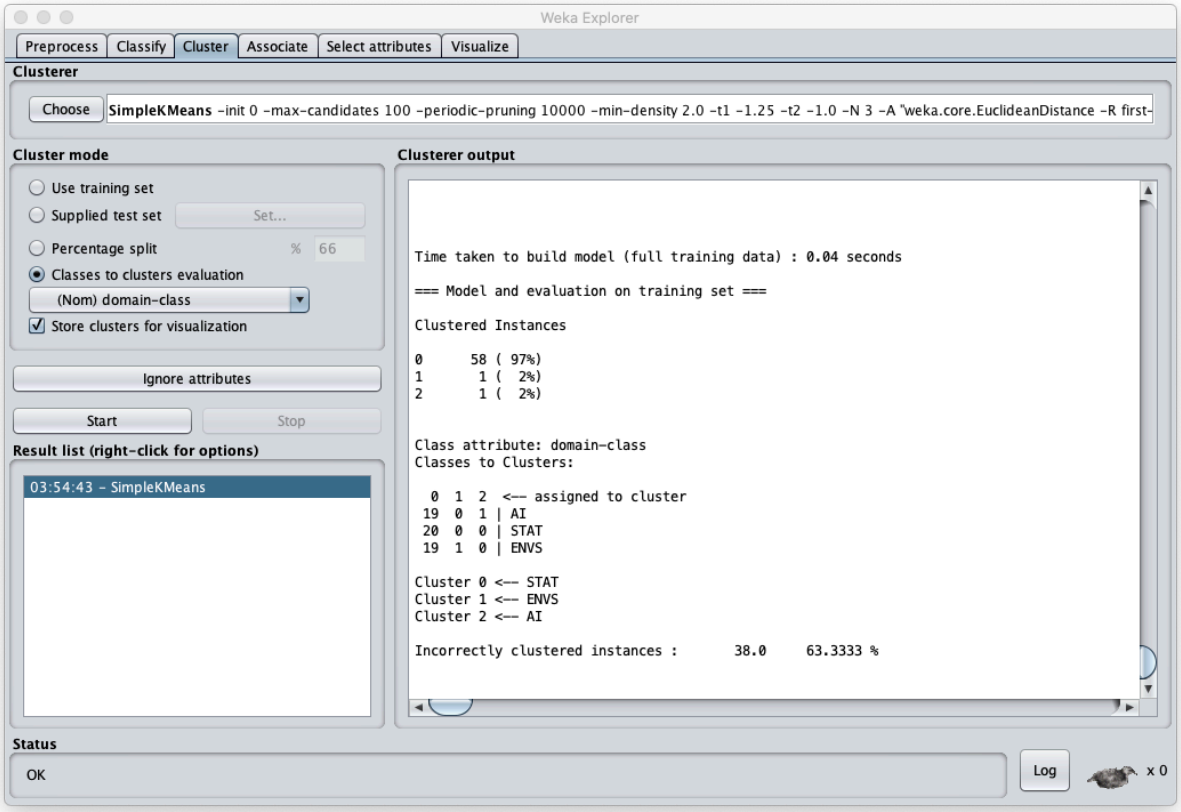
10 words to keep. It is accurate at this point. ENVS is the hardest to categorize here.



40 words to keep. STAT has a 100% accuracy, but ENVS accuracy only being clustered for 5/20. AI has 16/20, which has good number of accuracy. However, the difference in words between 10 and 40 seemed to have affect into ENVS.



160 words to keep. AI and ENVS getting more difficult to cluster. It has 2/20 and 7/20 correctly identified. The STAT maintains the 20/20 correctly percentage.



640 words to keep. Two of the abstracts were clustered into STAT. On the other side, two of the abstracts were not clustered into AI and ENS domains. The incorrectly clustered instances were 63.33%, which means many abstracts overlap with the other two domains, where AI and ENVS did not overlap each other. The problem may be caused by confusion in Weka's machine learning and drove most of the abstracts to be analyzed as STAT.