Christopher Foley

CAP6776 – Information Retrieval

Assignment 2 – WEKA

Assignment 3

Centroid Based Document Summarization

Assignment

CAP 6776 Information Retrieval (Fall 2017)

Homework # 3

DUE: 11/30/2017 11:59pm otal: 10 points

Implement the centroid based document summarization method.

Data set: docs4sum.txt contains 850 sentences and each line is a sentence. Please generate a short summary of 10 sentences. Download the text file in the "Files" section.

Method: please implement any centroid based summarization method to conduct document summarizaton using given data set.

An example procedure could be:

1. Cluster these sentences into 10 clusters
2. Calculate the sentence which has the highest similarity scores with all of the rest sentences in the same cluster, and include this centroid sentence into the summary.
3. Repeat step 2 for all the 10 clusters and generate the 10-sentence short summary.

You can choose any clustering method in the lecture of document clustering and also any similarity measure you think is the most appropriate. Please indicate the reason why you choose these methods.

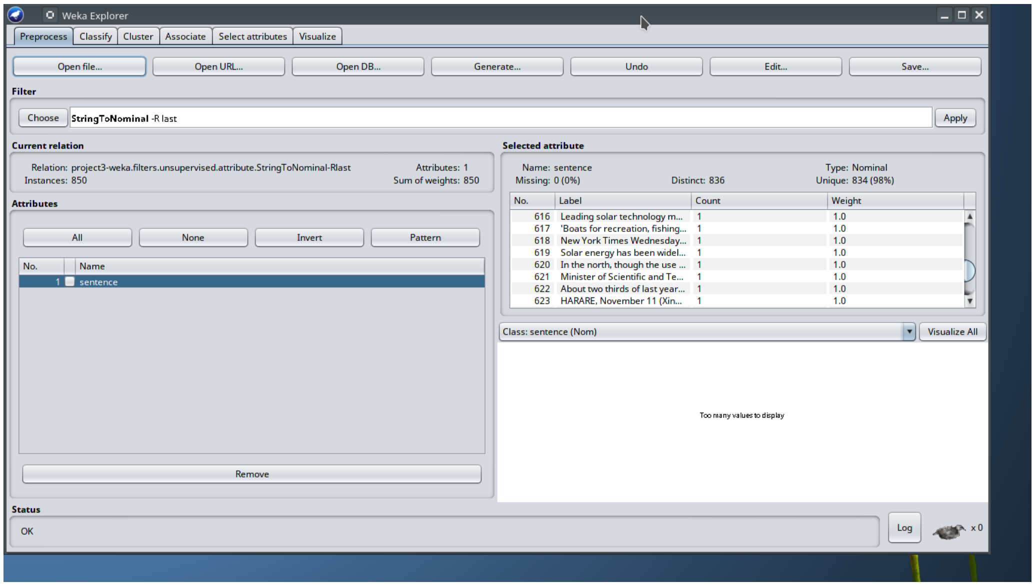
Introduction

The initial attempt was to use the techniques of Assignment 1 to read, stem and tokenize the input data. This provided a vector which could be used for cosine similarity to cluster. The clusters were then grouped and sorted to obtain the vector closest to the centroid. This was not effective because the values closest to the centroid were the vectored terms thus punctuation and numbers rated highest.

WEKA

Multiple approaches were attempted with WEKA. First the input file was edited to conform to ARFF format by adding @relation, @attribute and @data tags. Each line was pre/post fixed with a double quotation mark using VI (due to ease of substitution). This created problems with data when the data contained quotes from individuals. An examination of the data file also showed numerous paired quotes (similar to older typewritten text). The file was then changed to convert pairs of single quotes to double quotes and then the lines appropriately pre/post fixed.

Then WEKA was used (third attempt to use WEKA). The input file was edited to convert double qoutes to single quotes, then double qoute marks were added to the beginning and end of all lines. The ARFF headers were added. Using the unsupervised data StringToNominal filter the following was observed:



Filter

StringToNominal -R last

Current relation

Relation: project3-weka.filters.unsupervised.attribute.StringToNominal-Rlast
Attributes: 1
Instances: 850
Sum of weights: 850

Attributes

No.		Name
1	□	sentence

Selected attribute

Name: sentence
Type: Nominal
Missing: 0 (0%)
Distinct: 836
Unique: 834 (98%)

No.	Label	Count	Weight
616	Leading solar technology m...	1	1.0
617	'Boats for recreation, fishing...	1	1.0
618	New York Times Wednesday...	1	1.0
619	Solar energy has been widel...	1	1.0
620	In the north, though the use ...	1	1.0
621	Minister of Scientific and Te...	1	1.0
622	About two thirds of last year...	1	1.0
623	HARARE, November 11 (Xin...	1	1.0

Class: sentence (Nom)

Too many values to display

Status

OK

It should be noted that each sentence was treated as a single entitiy.

Then SimpleKMeans clustering with EuclideanDistance were applied and the following were determined to be the centroids of the clusters and the entire data file:

```
Cluster 0: ?

Cluster 1: 'Said Batchelor: ``I kept thinking, `What if the mosquito already bit me?\' \'\' NYT-08-23-00 1334EDT &QL; '

Cluster 2: 'Charges against the reputed gangsters center around the killing of a mainland Chinese businessman and a Hong Kong resident, armed robberies, smuggling explosives into Hong Kong, and the kidnapping of the two Hong Kong businessmen for more than 1.6 billion Hong Kong dollars (U.S. dlrs 205 million) in ransom. \t '
Cluster 3: '\'\'That allows them to continue to list coho salmon as a threatened
species, which in turn allows them to continue to regulate. '

Cluster 4: 'And Miller says that ``Massachusetts is a special case; the political
atmosphere here is more open. '

Cluster 5: '``The pretravel clinic visit,\'\' Dr. Jong said, ``is an opportunity
to immunize susceptible healthy travelers against chickenpox, which could cause
illness and exclusion from travel before, during or after the trip, and protects
the residents in receiving countries against imported infection.\'\' For adults,
the vaccine requires two injections. '

Cluster 6: '_ Every 20 seconds, someone in the United States is arrested for a
drug violation. '

Cluster 7: 'CONGRESSIONAL NEGOTIATORS AGREE TO CREATE MORE RENT SUBSIDIES (nk) By
DAVID STOUT c.1998 N.Y. '

Cluster 8: 'More than 160 million Hong Kong dollars are restrained under the
order. \t '

Cluster 9: 'People who become ill with a fever during or after tropical travel
should see a doctor and make sure to mention the trip, the FDA advised. '
```

All clusters: ?

Comments

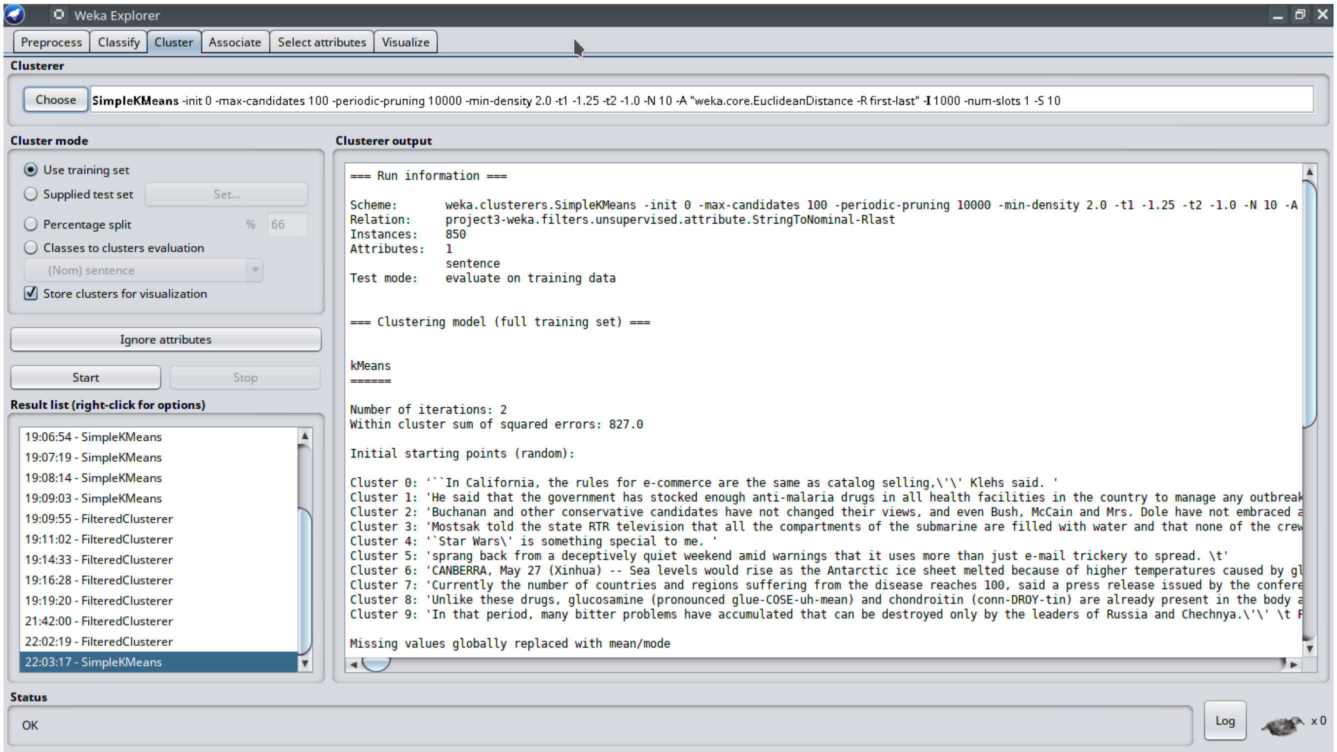
Attempts to analyze the data using the techniques of assignment 1 would not produce sentences since the strings were stemmed and tokenized. The computation of distance from the centroid then created a lexicographic output that did nor form complete sentences. The terms closest to the centroid could be selected and then statistically compared to the input data to find the sentences most likely to match, but that would be beyond the scope of the assignment.

As stated in the problem definition there were 850 inputs and 850 instances were found. Of concern, for possible review, would be that of the 850 only 836 were recognized as unique with 2 duplicates.

WEKA Console output

For unknown reasons, all data appeared to cluster into cluster 1.

The following was the output from WEKA:



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 1000 -num-slots 1 -S 10

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) sentence
☑ Store clusters for visualization

19:06:54 - SimpleKMeans
19:07:19 - SimpleKMeans
19:08:14 - SimpleKMeans
19:09:03 - SimpleKMeans
19:09:55 - FilteredClusterer
19:11:02 - FilteredClusterer
19:14:33 - FilteredClusterer
19:16:28 - FilteredClusterer
19:19:20 - FilteredClusterer
21:42:00 - FilteredClusterer
22:02:19 - FilteredClusterer
22:03:17 - SimpleKMeans

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A
Relation:     project3-weka.filters.unsupervised.attribute.StringToNominal-Rlast
Instances:    850
Attributes:   1
              sentence
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 827.0

Initial starting points (random):

Cluster 0: '``In California, the rules for e-commerce are the same as catalog selling,\'\' Klehs said. '
Cluster 1: 'He said that the government has stocked enough anti-malaria drugs in all health facilities in the country to manage any outbreak
Cluster 2: 'Buchanan and other conservative candidates have not changed their views, and even Bush, McCain and Mrs. Dole have not embraced a
Cluster 3: 'Mostsak told the state RTR television that all the compartments of the submarine are filled with water and that none of the crew
Cluster 4: '`Star Wars\' is something special to me. '
Cluster 5: 'sprang back from a deceptively quiet weekend amid warnings that it uses more than just e-mail trickery to spread. \t'
Cluster 6: 'CANBERRA, May 27 (Xinhua) -- Sea levels would rise as the Antarctic ice sheet melted because of higher temperatures caused by gl
Cluster 7: 'Currently the number of countries and regions suffering from the disease reaches 100, said a press release issued by the confere
Cluster 8: 'Unlike these drugs, glucosamine (pronounced glue-COSE-uh-mean) and chondroitin (conn-DROY-tin) are already present in the body a
Cluster 9: 'In that period, many bitter problems have accumulated that can be destroyed only by the leaders of Russia and Chechnya.\'\' \t F

Missing values globally replaced with mean/mode
```

OK