BAN 5743: Exercise 6 (10 Points)

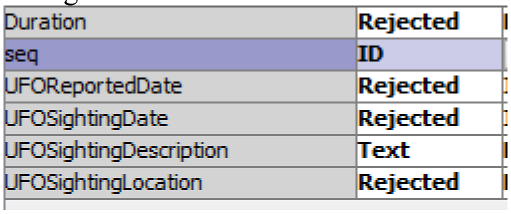
Text Analytics Solution

Problem Introduction and Data Description

This exercise is intended for you to get familiar with Text Topics and Text Clustering techniques. (**Please use VMWare View Client to Access SAS EM 14.1 for this exercise. If you use other versions of SAS software, your results may be different**)

Are we alone in this universe? This is a question that undoubtedly passes through every mind several times during a lifetime. We often hear a lot of stories about close encounters, Unidentified Flying Object (UFO) sightings and other mysterious things, but we lack the documented evidence for analysis on this topic. UFOs have been a matter of interest in the public for a long time. The objective of this exercise is to analyze one database that has a collection of documented reports of UFO sightings to uncover any fascinating story related to the data.

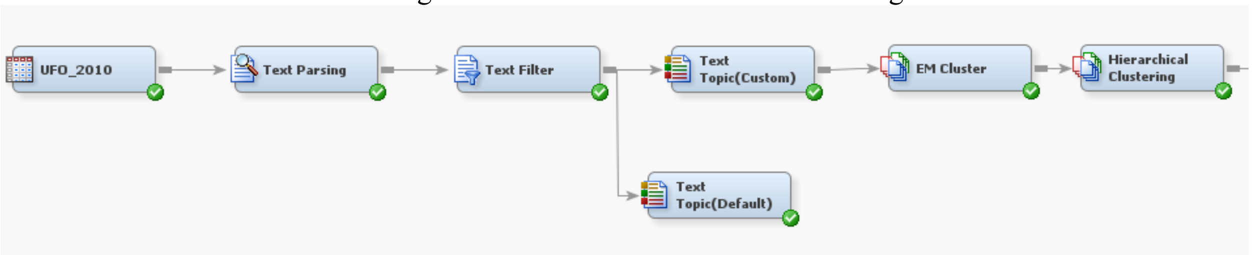
✓ Use the data set UFO_2010. This data set has comments of UFO reports recorded in 2010. Create the data source in SAS EM and assign the roles to variables as follows :



Duration	Rejected
seq	ID
UFOReportedDate	Rejected
UFOSightingDate	Rejected
UFOSightingDescription	Text
UFOSightingLocation	Rejected

✓ Create a new diagram and drag the data set onto the diagram space.

✓ Connect the data set with nodes as shown below. The nodes with Default in their names are run with default settings in their properties panel. Note that the EM Cluster node is a Text Cluster node with default settings but renamed as EM Cluster. Also, note that the Hierarchical clustering node is a Text Cluster node with default settings but renamed as Hierarchical clustering.



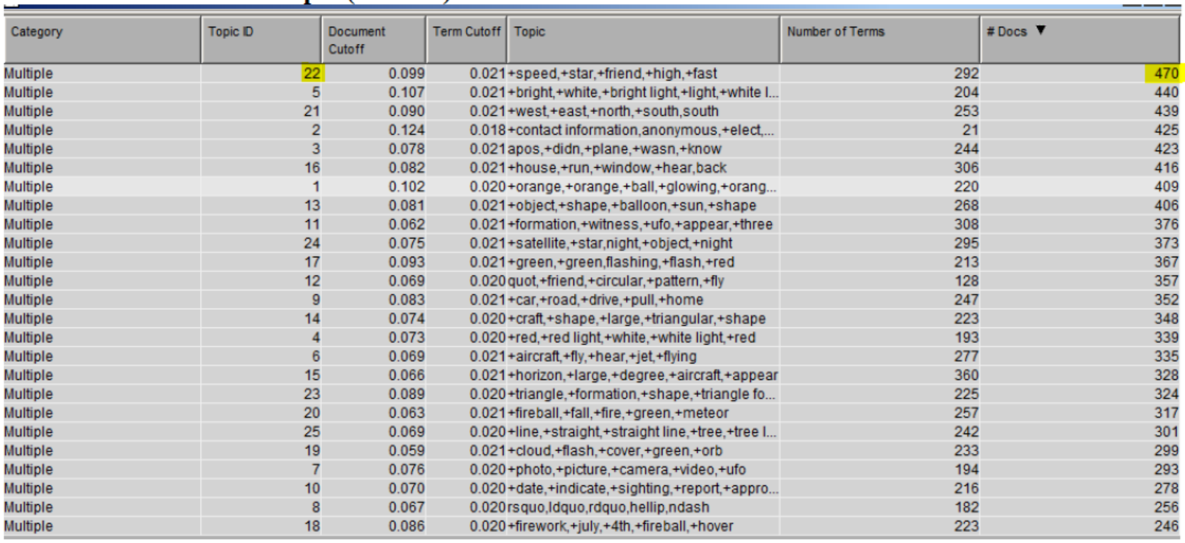
✓ Modify the properties of the Text Filter node. Change the Check Spelling property to **Yes.** Use the dictionary **UFO_SYN** from your library by clicking the ellipsis button next to Dictionary in the properties panel of each Text Filter node.

✓ Use the default setting for **Text Topic (Default)** Node and run the flow.

1. Examine the results **Text Topic (Default)** Node. **(1 point)**
   a. Are there any terms that repeat between different topics?
   b. If yes, point out some terms.

c. Which topic occurs in the highest number of documents and what terms does it contain?

The results of **Text Topic (Default)** are as follows:



Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs ▼
Multiple	22	0.099	0.021	+speed,+star,+friend,+high,+fast	292	470
Multiple	5	0.107	0.021	+bright,+white,+bright light,+light,+white l...	204	440
Multiple	21	0.090	0.021	+west,+east,+north,+south,south	253	439
Multiple	2	0.124	0.018	+contact information,anonymous,+elect,...	21	425
Multiple	3	0.078	0.021	apos,+didn,+plane,+wasn,+know	244	423
Multiple	16	0.082	0.021	+house,+run,+window,+hear,back	306	416
Multiple	1	0.102	0.020	+orange,+orange,+ball,+glowing,+orang...	220	409
Multiple	13	0.081	0.021	+object,+shape,+balloon,+sun,+shape	268	406
Multiple	11	0.062	0.021	+formation,+witness,+ufo,+appear,+three	308	376
Multiple	24	0.075	0.021	+satellite,+star,night,+object,+night	295	373
Multiple	17	0.093	0.021	+green,+green,flashing,+flash,+red	213	367
Multiple	12	0.069	0.020	quot,+friend,+circular,+pattern,+fly	128	357
Multiple	9	0.083	0.021	+car,+road,+drive,+pull,+home	247	352
Multiple	14	0.074	0.020	+craft,+shape,+large,+triangular,+shape	223	348
Multiple	4	0.073	0.020	+red,+red light,+white,+white light,+red	193	339
Multiple	6	0.069	0.021	+aircraft,+fly,+hear,+jet,+flying	277	335
Multiple	15	0.066	0.021	+horizon,+large,+degree,+aircraft,+appear	360	328
Multiple	23	0.089	0.020	+triangle,+formation,+shape,+triangle fo...	225	324
Multiple	20	0.063	0.021	+fireball,+fall,+fire,+green,+meteor	257	317
Multiple	25	0.069	0.020	+line,+straight,+straight line,+tree,+tree l...	242	301
Multiple	19	0.059	0.021	+cloud,+flash,+cover,+green,+orb	233	299
Multiple	7	0.076	0.020	+photo,+picture,+camera,+video,+ufo	194	293
Multiple	10	0.070	0.020	+date,+indicate,+sighting,+report,+appro...	216	278
Multiple	8	0.067	0.020	rsquo,ldquo,rdquo,hellip,ndash	182	256
Multiple	18	0.086	0.020	+firework,+july,+4th,+fireball,+hover	223	246

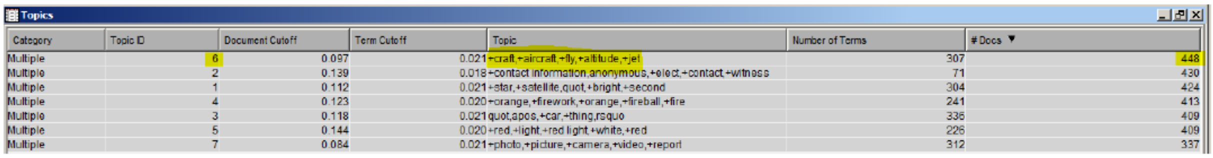
Many terms repeat between the topics. Some of them are "shape", "star" , "white", "white light" etc.,

The topic ID 22 occurs in 470 documents which is the highest. It contains the words "speed, star, friend, high, fast."

✓ Modify the properties of the **Text Topic (Custom)** node. Change the number of multi-term topics to **7**.

2. Run the **Text Topic (Custom)** node and examine the results. **(1 point)**
   a. Are there any terms that repeat between different topics?
   b. If yes, point out some terms.
   c. Which topic occurs in the highest number of documents and what terms does it contain?

The results of **Text Topic (Default)** are as follows:



Topics

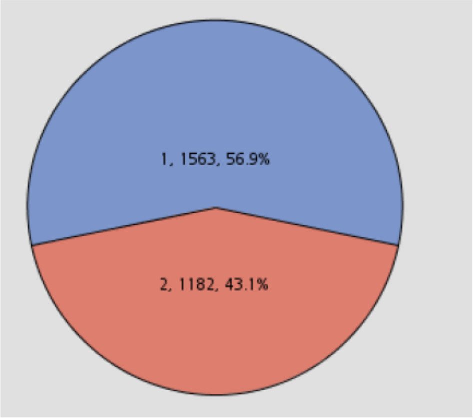
Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs ▼
Multiple	6	0.097	0.021	+craft,+aircraft,+fly,+altitude,+jet	307	448
Multiple	2	0.139	0.018	+contact information,anonymous,+elect,+contact,+witness	71	430
Multiple	1	0.112	0.021	+star,+satellite,quot,+bright,+second	304	424
Multiple	4	0.123	0.020	+orange,+firework,+orange,+fireball,+fire	241	413
Multiple	3	0.118	0.021	quot,apos,+car,+thing,rsquo	336	409
Multiple	5	0.144	0.020	+red,+light,+red light,+white,+red	226	409
Multiple	7	0.084	0.021	+photo,+picture,+camera,+video,+report	312	337

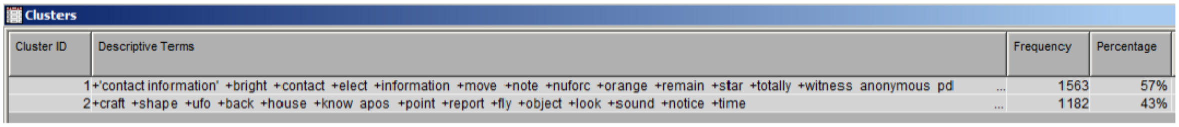
There is only one topic "quot" which repeats between the topics.

The topic ID 6 occurs in 448 documents which is the highest. It contains the words "craft, aircraft, fly, altitude, jet."

3. Run the **EM Cluster** node (this is text cluster with default settings and renamed) and examine the results. **(1 point)**
   a. How many clusters are made?

EM Clustering with default settings gave 2 clusters.





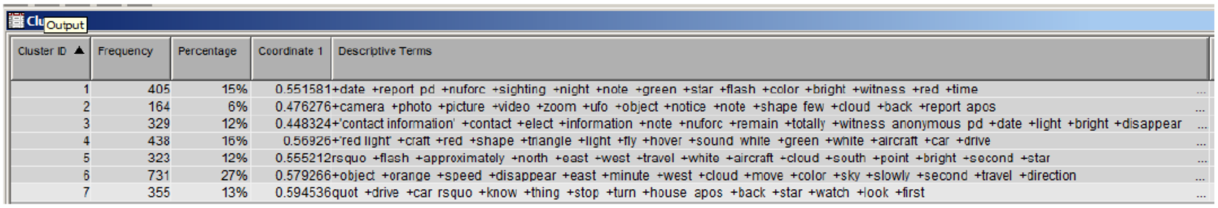
Clusters

Cluster ID	Descriptive Terms		Frequency	Percentage
1	+'contact information' +bright +contact +elect +information +move +note +nuforc +orange +remain +star +totally +witness anonymous pd	...	1563	57%
2	+craft +shape +ufo +back +house +know apos +point +report +fly +object +look +sound +notice +time	...	1182	43%

- ✓ Modify the properties of the **EM Cluster** node. Change **Exact or Maximum Number** of clusters in the properties panel to **Exact** and **Number of clusters** to 7.

4. Run the nodes. **(1 point)**
   a. Comment on the results.

EM Clustering output with 7 exact number of clusters.



Clu Output

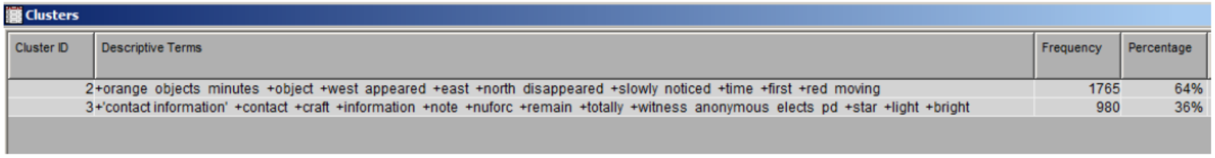
Cluster ID ▲	Frequency	Percentage	Coordinate 1	Descriptive Terms	
1	405	15%	0.551581	+date +report pd +nuforc +sighting +night +note +green +star +flash +color +bright +witness +red +time	...
2	164	6%	0.476276	+camera +photo +picture +video +zoom +ufo +object +notice +note +shape few +cloud +back +report apos	...
3	329	12%	0.448324	+'contact information' +contact +elect +information +note +nuforc +remain +totally +witness anonymous pd +date +light +bright +disappear	...
4	438	16%	0.56926	+'red light' +craft +red +shape +triangle +light +fly +hover +sound white +green +white +aircraft +car +drive	...
5	323	12%	0.555212	rsquo +flash +approximately +north +east +west +travel +white +aircraft +cloud +south +point +bright +second +star	...
6	731	27%	0.579266	+object +orange +speed +disappear +east +minute +west +cloud +move +color +sky +slowly +second +travel +direction	...
7	355	13%	0.594536	quot +drive +car rsquo +know +thing +stop +turn +house apos +back +star +watch +look +first	...

We can see better clusters as we have forced the clustering algorithm to make 7 clusters. The clusters now have many related terms unlike the clusters we obtained with default settings.

- ✓ In the **Hierarchical clustering** node, change **Cluster Algorithm** to **Hierarchical** in the properties panel and rename the node to **Hierarchical Clustering**.

5. Then run the **Hierarchical Clustering** node and examine the results. **(1 Point)**
   a. How many clusters are there?

Hierarchical Clustering with default settings gave 2 clusters.



Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
2	+orange objects minutes +object +west appeared +east +north disappeared +slowly noticed +time +first +red moving	1765	64%
3	+'contact information' +contact +craft +information +note +nuforc +remain +totally +witness anonymous elects pd +star +light +bright	980	36%

- ✓ Attach a **SAS Code** node to the flow containing the **Hierarchical Clustering** node.

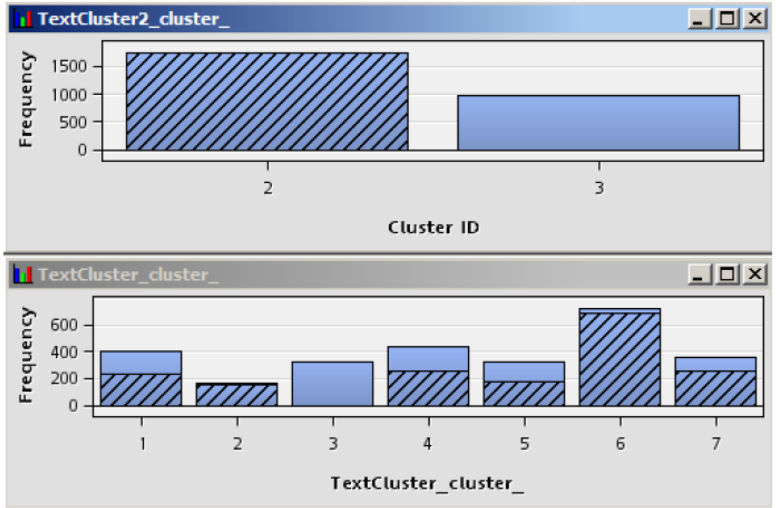
- ✓ Right-click the **SAS Code** node and select **Edit Variables**.

- ✓ Then explore the two cluster variables to get a sense of overlap between cluster memberships from both methods.

6. Examine how the terms in cluster from Hierarchical Cluster node are split across other clusters from the EM cluster node. **(1 Point)**
   a. Comment and explain about what you see.

(*Hint:* In the pop-up box, sort the column Role to see the two variables with the Segment role. While holding down the Ctrl key, click the two variables with Segment roles and click Explore.)

Hierarchical Clustering node vs EM Clustering node

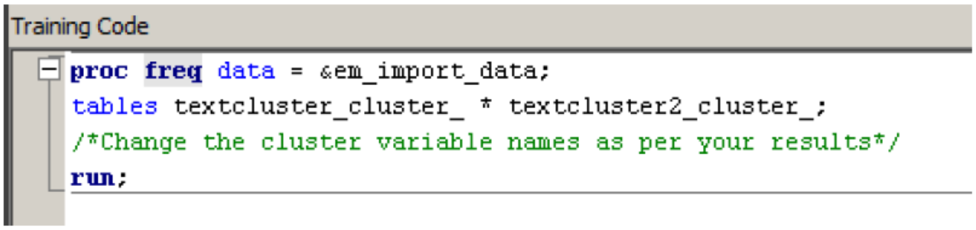


From the above output, we can see that the terms in cluster 2 of the hierarchical method are split across clusters 1, 2, 4, 5, 6 and 7.

- ✓ Attach a **SAS Code** node and write a code to do a crosstab between clusters from the two different algorithms.

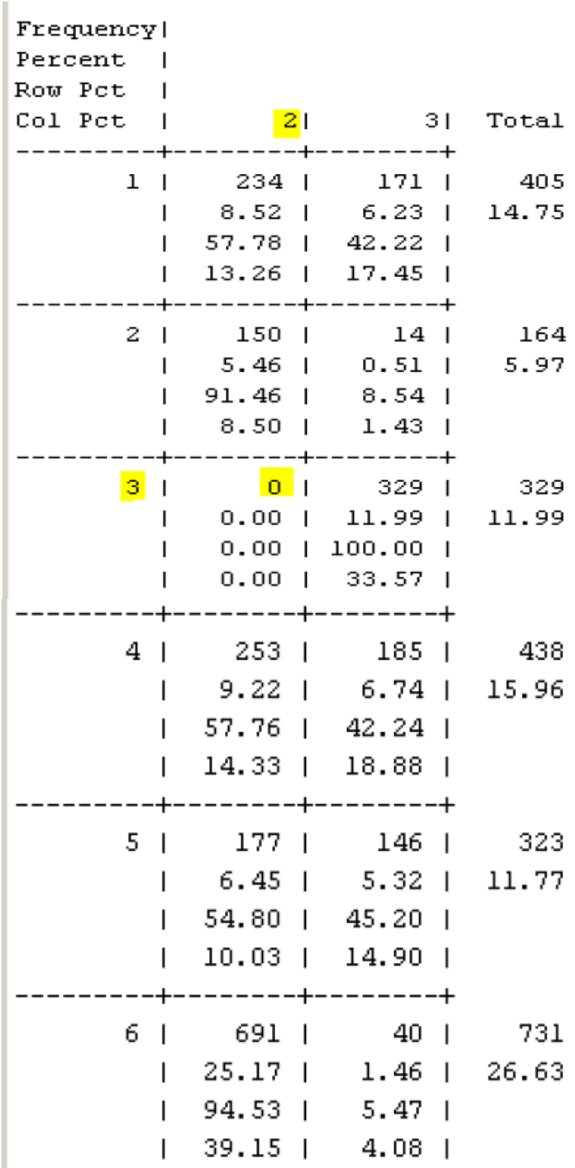
7. Report your results. **(1 Point)**
   a. Are there any clusters that do not overlap between different algorithms? (*Hint* : Use the code provided)

The code used to do cross tab is as follows:



Training Code

```
proc freq data = &em_import_data;
tables textcluster_cluster_ * textcluster2_cluster_;
/*Change the cluster variable names as per your results*/
run;
```



```
Frequency|
Percent  |
Row Pct  |
Col Pct  |       2|       3|  Total
---------+--------+--------+
       1 |    234 |    171 |    405
         |   8.52 |   6.23 |  14.75
         |  57.78 |  42.22 |
         |  13.26 |  17.45 |
---------+--------+--------+
       2 |    150 |     14 |    164
         |   5.46 |   0.51 |   5.97
         |  91.46 |   8.54 |
         |   8.50 |   1.43 |
---------+--------+--------+
       3 |      0 |    329 |    329
         |   0.00 |  11.99 |  11.99
         |   0.00 | 100.00 |
         |   0.00 |  33.57 |
---------+--------+--------+
       4 |    253 |    185 |    438
         |   9.22 |   6.74 |  15.96
         |  57.76 |  42.24 |
         |  14.33 |  18.88 |
---------+--------+--------+
       5 |    177 |    146 |    323
         |   6.45 |   5.32 |  11.77
         |  54.80 |  45.20 |
         |  10.03 |  14.90 |
---------+--------+--------+
       6 |    691 |     40 |    731
         |  25.17 |   1.46 |  26.63
         |  94.53 |   5.47 |
         |  39.15 |   4.08 |
```

From the above Cluster Output we can observe that Cluster 3 of the EM Cluster Node doesn't overlap with cluster 2 of the Hierarchical method.