Extraction and Description of the Narrative Structure of TV Series

Aman Z. Berhe

Jan 2018

1 Report

1.1 Data Representation

I have played with the data a lot. I have extracted the text of each scene of the nine episodes of GoT. I have done visualization of the data in the following representations

1. TF_IDF: tf_idf representation and clustering using the tfidf representation
2. Count Vector: this is almost similar to Tf representation
3. Word2vec representation
4. Doc2vec representation: each scene(segment) has been assumed as documents

I have done the clustering of each scene(segment) using the above representations. I have used different evaluation techniques for the clusters as can be seen on the table below.

N	data repre	Clustering Algorithm	Evaluation		
			Purity	NMI	Accuracy
1	TF-IDF	kmeans	0.137	0.077	0.023
2	TF-IDF	Affinity Propagation	0.005	0.545	0.042
3	TF-IDF	**Mean Shift**	**0.120**	**0.547**	**0.042**

N	data repre	Clustering Algorithm	Evaluation		
			Purity	NMI	Accuracy
1	Count Vector (cos-sim)	fcluster	0.0	0.094	0.051
2	Count Vector(ecld)	fcluster	0.0	0.083	0.037
3	Count Vector (cos-sim)	Kmeans	0.160		
4	Count Vector (ecld)	Kmeans	0.105		
5	Count Vector (cos-sim)	Affinity Propagation	0.005		
5	Count Vector (ecld)	Affinity Propagation	0.005		
6	Count Vector (cos-sim)	Agglomerative	0.202		
7	Count Vector (ecld)	Agglomerative	0.349		

N	data repre	Clustering Algorithm	Evaluation		
			Purity	NMI	Accuracy
1	doc2vec	Agglomerative	0.224	0.127	0.059
2	doc2vec	kmeans	0.306	0.132	0.059
3	doc2vec	Affinity Propagation	0.146	0.161	0.050
4	doc2vec	Mean Shift	0.161	0.547	0.050

There are some images that illustrate what is happening and try to indicate the clusters using colors, on the python file Scene_Clustering_Episodes, in the folder codes.

2 Next Task

- Try different parameters of the algorithms (such in Kmeans the good Perfect k and the perfect iteration).
- Use the Book scene segments for generating the doc2vec model
- Fix the error on the length of the training labeled and predicted labels for the count Vector representation
- Try to use TVD to generate more data
- Read more papers