May 18, 2017 Report

Extra preprocessing step added: remove semicolons at the end of the word.

Cluster Centers

Intuitively doesn't make any sense. But trying to observe what they mean.

The intuition is that clustering groups words in a way that one word is the primary center, and the cloud around it has the words that belong to the primary center word. Therefore, the assumption is the cluster center represents a group of words as a whole, which makes it a keyword.

When number of clusters is set to some high number, in this case, 15, and if we apply a threshold of 0.1 to limit the number of keywords generated, we are getting the words that can summarize the document.

This method was tested on a few documents, and observed marginal improvements/gains in our **score** metric compared to clustering and PCA.

There is an issue with this method, which we observed while testing some documents. The factor we are taking into consideration is that when similar words are grouped together, the cluster center will be a representative of those words. This is good, but in some rare cases, this falters. For example, in the Mahabharata example, Kauravas and Pandavas are members of a similar cluster. This is good, but the issue is, we are taking on the center. Therefore, we are going to miss one of those. The big task is to figure out a way of separating important clusters that might contain these kinds of attributes. The same cluster was shared between Arjuna, Krishna, and Rama. But the center was closest only to Arjuna, therefore, we missed Krishna and Rama.

So, our task here should be to find an important cluster. The question is, what makes a cluster important?

The next task would be to write a clustering algorithm that can group the most similar words together based on the user input.