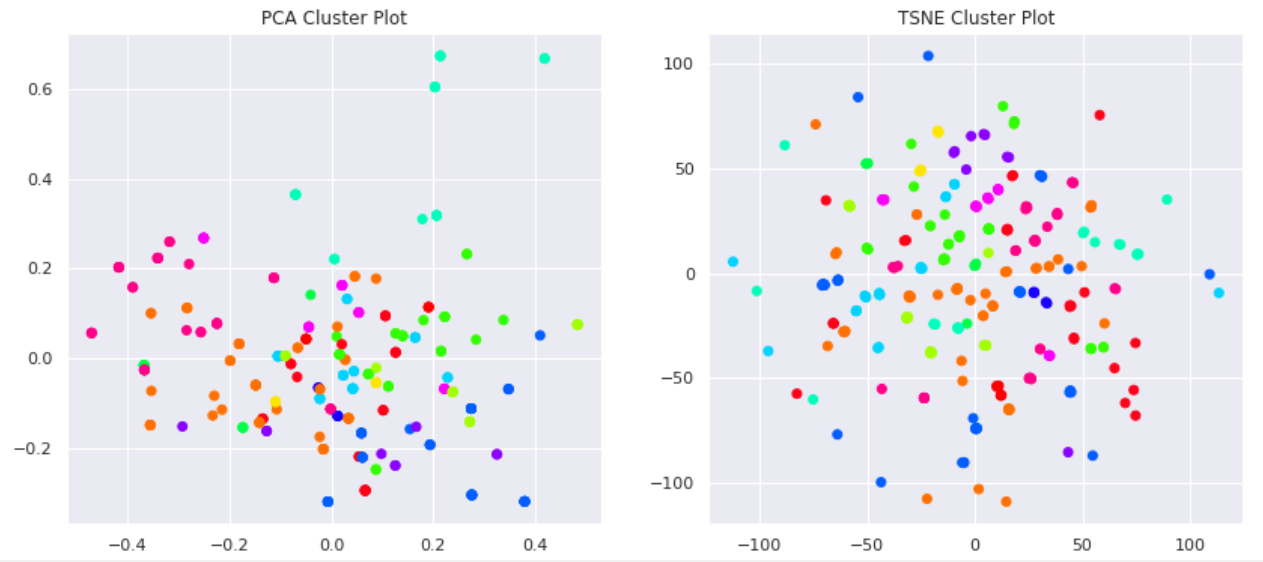
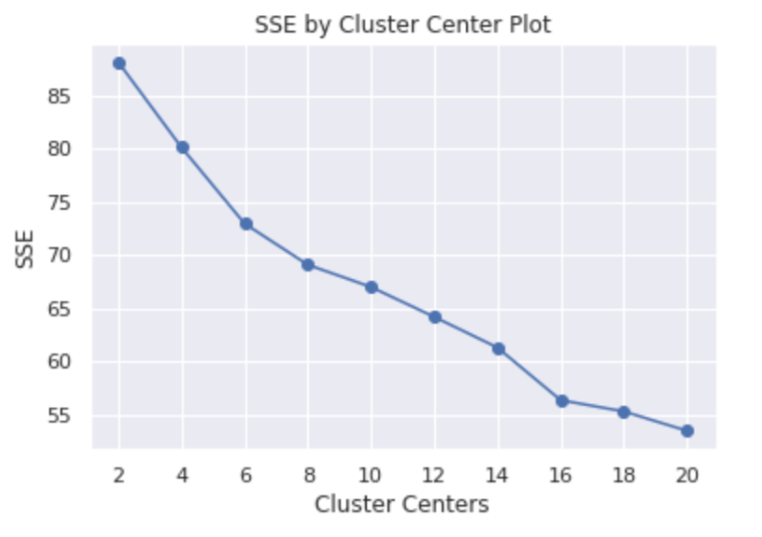
question kmeans & relation trees SBERT sandbox.ipynb

Rregrouping top keywords by clusters’ size with *tf-idf* weights, with the questions 

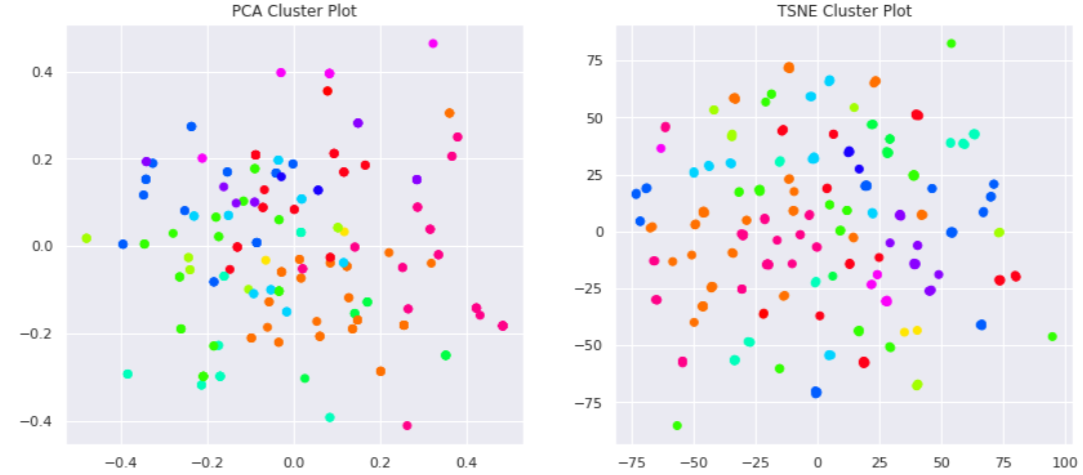
Principal component analysis t-Distributed Stochastic Neighbor Embedding



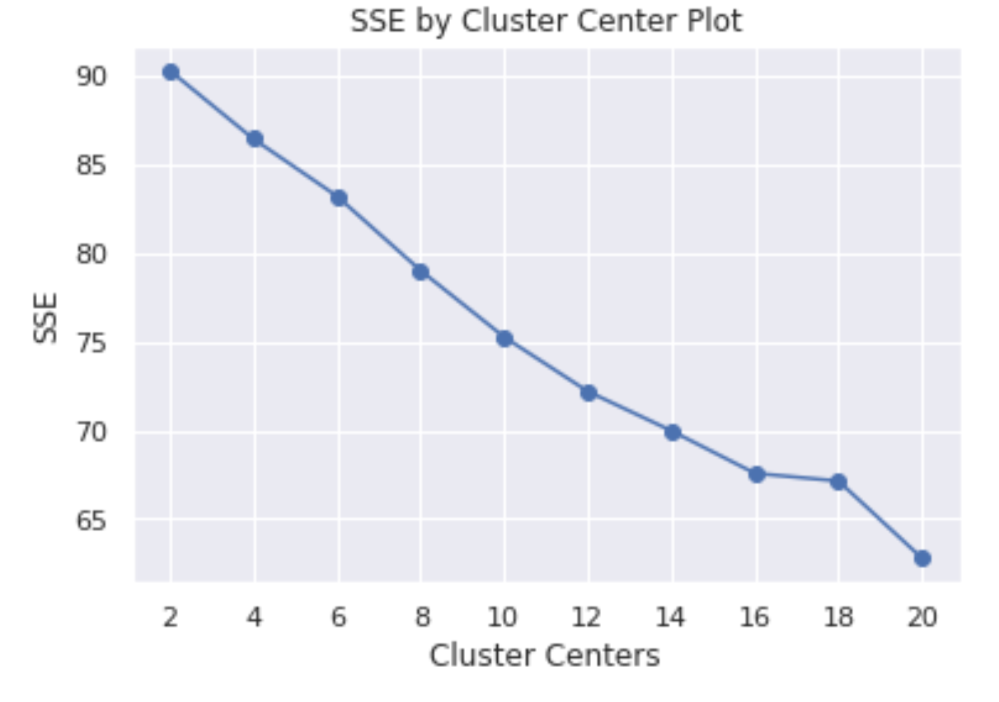
Square of Sum Error for each clusters’ centers’ size, with the questions

*entire\_phrase kmeans & relation trees SBERT sandbox.ipynb*

Rregrouping top keywords by clusters’ size using *tf-idf* weights, with the *entire phrase*

**

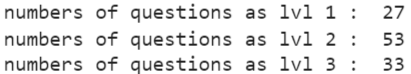
Principal component analysis t-Distributed Stochastic Neighbor Embedding

Square of Sum Error for each clusters’ centers’ size, with the *entire phrase*

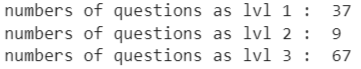
*Adding further data to our model (entire\_phrase) increases the number of clusters needed in order to decrease the SSE.*

*\**

*Questions-only Kmeans*

**

Kmeans classification according to the *tf-idf weights*

**

Kmeans classification with *Wikipedia4gb* as the embedder

*The sentenceTransformer Wikipedia4gb* tends to classify the level 2 question : as level 3, or as level 1.

This underlines how it does not take full advantage of the words’ tf-idf weights in the level 2.

*entire\_phrase Kmeans*



Kmeans classification according to the *tf-idf weights*

Same results with the tf-idf weights



Kmeans classification with *Wikipedia4gb* as the embedder

*Wikipedia4gb* with the *entire\_phrase* tends to over-classify the level 1

It takes more advantage of the tf-idf weights than the *question-only* model, with its classification of level 2