Graphical visualization of high dimensional text representations

t-sne

t-sne for t-distributed stochastic neighbor embedding is a method designed for dimensionality reduction, mainly for a visualization purpose. But it can be applied to any vectorized data in order to reduce the space dimension (or number of variables).
Basically, the method try to minize the divergence (kullback leibler) within the input space and the tiniest output space, usually (2 or 3 dimensions). Two neighboors points in input space must be neighbors also in the output space. Each space had a distributional probability, and the goal of t-sne is to reduce the divergence between them.

An example of application of t-sne is given in annexe (DVO) in file visualization.py, you will need the file train.csv :

https://drive.google.com/file/d/1XRCN5k0x97xyGRH8v-1Z0hGT4fryJwIp/view

1) Start by running the program and understand the differents steps.
This python program load the datas, some preatements are done in order to extract the vocabulary, then this vocabulary is vectorized with word2vec and finally t-sne project words a 2 dimensional space the word embedding.

2) Start to design an Information Retrieval system using the dataset CNN. It's composed of pairs of documents and summaries. We will use the summaries as request and the goal will be to find the related document. Your system need to return a rank for all the document.

If your model is perfect (theoretically), the first document of the list is the document paired with the request summary.

To simplify you can start with only the first 100 (Document,Summary) pairs.

You can use Word2vec, TF-IDF or any other pre-treatment.