Dimension reduction, word embeddings and VAE for generating sentences

The project will be sent by email (to `christophe.ambroise@univ-evry.fr`) as a `pdf` file with the corresponding notebook (`python` or `Rmd`). Briefly describe the problem, write the calculations you are programming. The project can be done in pairs or alone.

In order to have reproducible results, you can set the seed of the random number generator to the value 20222023.

Data representation

1. Consider the dataset from `https://www.kaggle.com/datasets/crawford/20-newsgroups/code`, describing a collection of 18,000 documents from 20 different newsgroups.
2. Use a NLP (Natural Language Processing) library to convert your data to TF-IDF format.
3. Compute a matrix of similarities between a stratified random sample of 1000 documents using the corrrelation (cosine similarity).
4. Sample 500 words from your 1000 documents (explaining your sampling strategy) and compute a matrix of word co-occurence (in documents).
5. Use the k-medoids algorithm to cluster the the documents into 20 classes. Comment.
6. Represent words and documents using SVD (Latent Semantic Analysis), t-SNE and UMA. Comment.
7. Use word2vec to create word embeddings, then visualize using t-SNE and UMA. Comment.

Sentence generation

Consider the paper ”Generating Sentences from a Continuous Space” from Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Bengio published in 2016.

1. Use the Variational Auto Encoder from `https://github.com/timbmg/Sentence-VAE` to generate sentences from a given class of documents.
2. Illustrate the performance of your algorithm.