DATA 221
Homework 6 - Embeddings and clsutering

W. Trimble

Due: Friday 2023-05-19 - 11:59pm

For the PCA plots of the epicurious dataset:

1. Embedding

   (a) Using at least 2000 rows from epicurious, calculate the all-against all distance matrix in Euclidean space for the category labels. Produce a clustering.

   (b) Using the same 2000 rows, calculate the all-against all distance matrix with a Minkowski metric for the category labels. Produce a clustering.

   (c) Map the category labels to a word2vec language embedding, calculate bag-of-words vectors for each recipe, and produce a clustering. Visualize the clustering (PCA, TSNE, or clustermap would do)

   (d) Compare the three clusterings visually; make three PCA or TSNE scatterplots and label them suitably.

2. Feature selection / embedding

   The Online News Popularity dataset collected by Fernandes et al. `https://archive.ics.uci.edu/` contains 58 features extracted from 39,000 webpages and one response variable number of shares.

   (a) Extract the text from the 39,000 webpages and check whether you can reproduce the ”text length” field.

   (b) Calculate bag-of-words word2vec feature vectors for each of the articles and perform clustering (kmeans or hierarchical will do)

   (c) Plot the within-cluster-variance vs. cluster number graph and choose a number of clusters.

   (d) See if you can identify the clusters. You can try to use the vectors or you can inspect the cluster members.