Article categorisation

Blaž Erzar

Machine Learning For Data Science 1 2023/24
Homework 7 report

1 Introduction

In this homework, we explore the underlying patterns of the RTV SLO articles dataset using unsupervised learning. Each article is described by a set of ChatGPT-generated keywords, from which we want to reconstruct the categories. We use two approaches, visualisation using a 3D PCA biplot and t-SNE.

2 Methods

2.1 Data preparation

First, we preprocess keywords by removing special characters and numbers. We remove any keywords shorter than 3 characters and replace the space character with an underscore. This is done because `scikit-learn` uses whitespace for tokenisation when computing tf-idf. This algorithm is used to generate the final features. We remove keywords with a frequency lower than 20.

2.2 PCA

Before doing PCA, we standardise the data. We use the first 3 PCA components to plot the data. To select the keywords to plot, we manually select the directions in which the data spreads the most. Then, we find the loading vectors which agree with these directions the most.

2.3 t-SNE

To visualise the data in 2 dimensions, we use t-SNE. For better results, we first standardise the data and transform the data into 10-dimensional space using PCA.

To explain the embedding, we use k-means clustering on top of the original dataset. For k, we select the number of clusters which seem well-grouped. To describe each cluster, we use the top 3 keywords of the articles in the clusters.

3 Results

Looking at the results in Figure 1, we can see that the data is mostly spread in 3 directions. Keywords associated with these directions are *Israel*, *NHL* and *race*. There is also some spread in 3 more directions shown by keywords *art*, *victory* and *government*. Based on these findings, we may propose article categories **politics**, **culture** and **sport**.

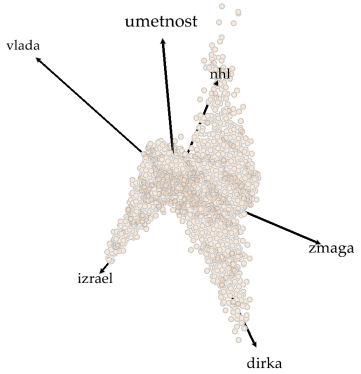

Figure 1: *3D PCA biplot containing vectors of keyword loadings which agree with the spread directions the most. Each vector is scaled to the same length for better visibility.*

With the results in Figure 2, we can see similar results but also some differences. There are two additional categories which can be obtained from keywords *floods*, *accident*, *evacuation* and *government*, *law*, *elections*. Based on these, we propose categories **culture**, **sport**, **local news**, **local politics** and **foreign politics**.

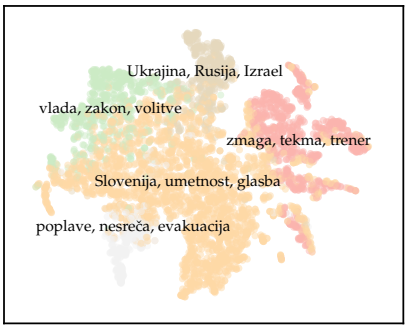

Figure 2: *Scatter plot of 2-dimensional t-SNE embedding with k-means clusters of the original space. The top 3 keywords for each cluster are shown.*

4 Discussion

Using both methods, PCA and t-SNE, we obtained some knowledge about the underlying article categories. PCA already provided some insights, but we were able to construct two additional categories using t-SNE. Even though the clusters are not clearly visible in the visualisation, the ones obtained from k-means seem sound based on their embedding.