Machine Learning - HW7

Tim Kmecl, 2.6.2024

Implementation

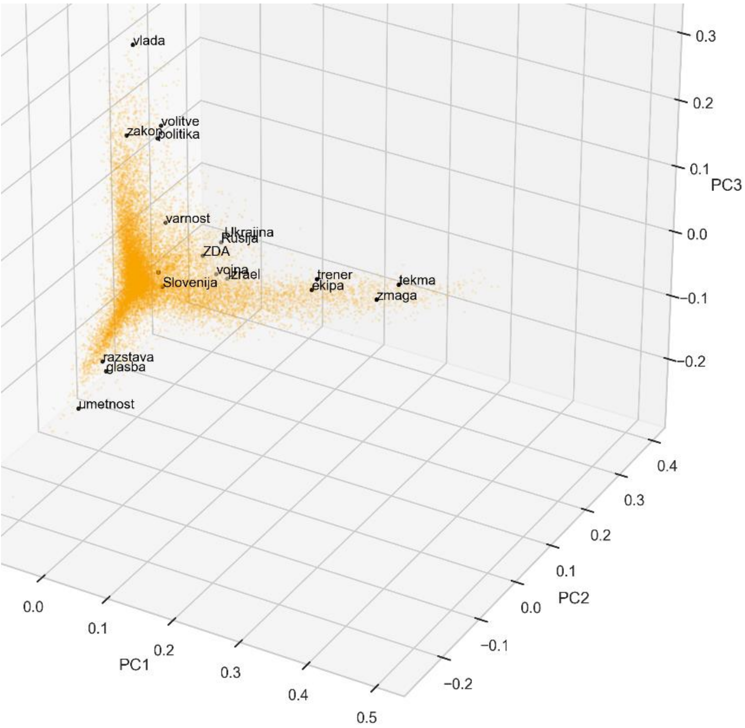
I followed the same proedure as described in instructions: use TF-IDF on keywords and only keep keywords that appear at least 20 times. I used jupyter notebook, and pyplot for all the plots since vispy didn't work on my machine. Plots are not SVG since the resulting file was too large for učilnica due to too many points in plots.

PCA

Data was transformed using 3d PCA and plotted in orange. For each component, in jupyter notebook I printed out the words that contribute the most to high and to low score for that component using the coeeficients from the corresponding eigenvectors (most positive and most negative coefficients). For the plot, 3 words that contribute the most to high and 3 that most to low value were taken, together with 9 most common words. Since there were repetitions, this yielded 18 keywords to be displayed. For each, a fictive sample was created with all features 0 except the one corresponding to that word equal to 1. These were then transformed via PCA – the resulting points are black, with the keyword displayed next to each.

Based on the displayed keywords we can reconstruct the categories. We can see that PC1 separates *sports* (with keywords trener, ekipa, tekma) to the right of everything else. High PC1 therefore means sports. PC2 mostly separates *culture* from other articles, with negative PC2 for culture-related keywords (razstava, glasba, umetnost). On the other hand, high PC2 corresponds to article with war and international keywords. This category becomes more pronounced when we also consider PC3, which distinguishes internal politics from everthing else. We can see that high PC3 should be labeled *politics* (keywords vlada, zakon, politika), and combination of low PC3 and high PC2 *international* (ZDA, Rusija, Ukrajina) – note that this one is less pronounced compared to the other three.

We therefore get 4 specific categories, with a lot of articles falling in the middle, which we could categorise as other. It is interesting that keyword Slovenija, the most common in the dataset, appears there, probably because of multiple contexts where it can be used – both for sports (national teams and our sportsmen), international relations with other countries, internal affairs etc. The plot is displayed below.



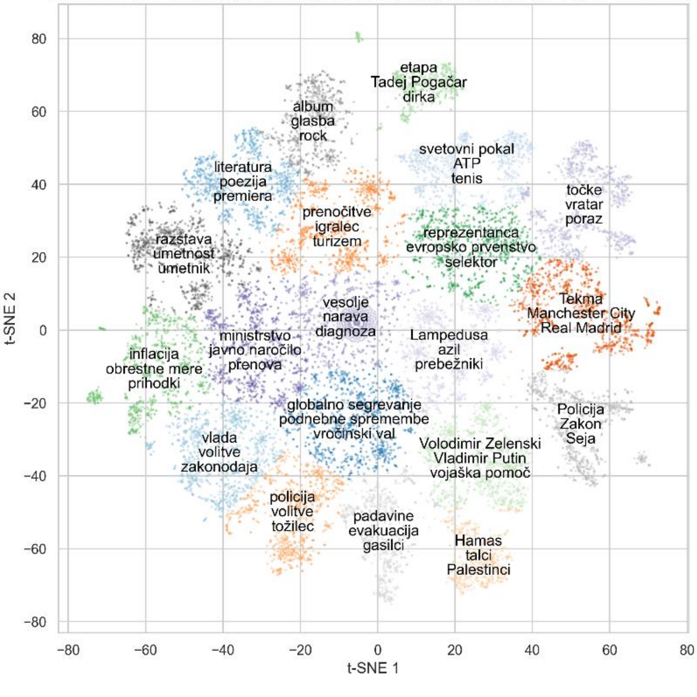
t-SNE

First, I preprocessed the data by transforming it via PCA with 100 principal components, which were then used for t-SNE. I used sklearn implementation with perplexity set to default value of 30 and number of iterations to 1000.

I obtained 20 clusters using k-means clustering. The number was chosen so that the resulting clusters give us sufficient granularity to interpret the plot while not making it too cluttered and since 20 subcategories seem about right amount to me. Those could then be grouped into proper categories. Then, TF-IDF is computed, with clusters being treated as documents – keywords labeling each cluster should be those that appear a lot inside it but rarely outside it, and high TF-IDF score of a word denotes exactly this. I printed the words with highest score in the notebook, and decided to display top 3 for each cluster in the plot centered at cluster mean, with most important one on top. In this way, we can see in which cluster and in which area of the plot specific topics are most common.

On top left, we see 3 clusters for music, literature and art – *culture*. On top right, there are several clusters for different sports: cycling, tennis, football (divided into club and national teams). Going to the bottom from center, there are three clusters dealing with *international* affairs/politics, each for different topic: top one for migrations, one mostly for Ukraine crisis/war, and one for Palestine/terorism. To the left, there are two for *climate*: bottom probably for natural disasters and top for global warming. It is interesting to note that it seems like the entire central bottom area of the plot is related to topics concerning *security* in general, especially considering there is cluster keyworded ''police'' next to the ones described before. Bottom left there are few clusters with keywords clearly related to *politics* or internal affairs. The leftmost cluster corresponds to *economy*. The ''prenočitve'' cluster could be label *lifestyle*, since it seems to also include actors etc.

t-SNE of TF-IDF with K-means clustering and most important keywords (TF-IDF) for each cluster

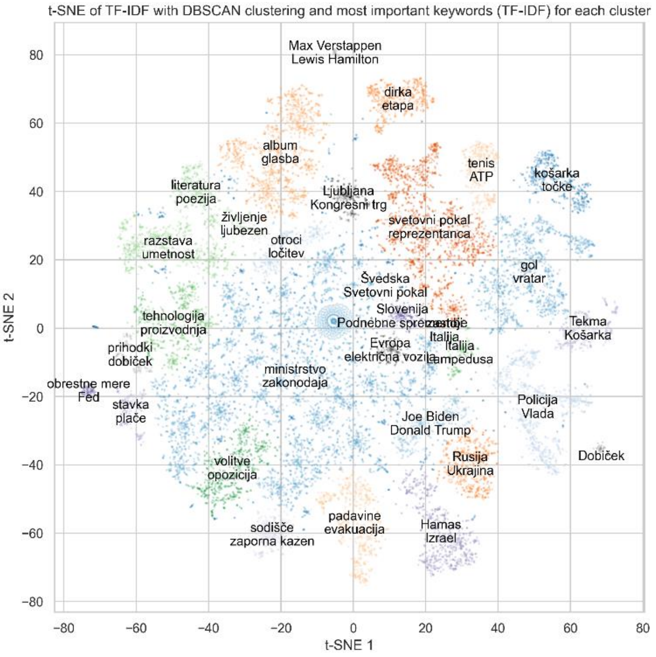


The results obtained from k-means seem sufficient to me to adequatly cluster articles into topics, where the resulting clusters could be joined into more general categories. This visualization allows us to intepret different areas of the t-SNE space, and it is clear that the position inside this space carries a semantic meaning.

We can however see that some borders between obtained clusters seem arbitrary, especially near the middle. I therefore tried DBSCAN clustering with different parameters, and settled on using epsilon 3 with 40 min. samples. This resulted in 30 clusters. In my opinion, borders between clusters better correspong to borders I would personally draw in the scatterplot. However we lose some precision in the central to bottom left area, with different political and climate topics being merged into one big cluster. There are also some really small clusters just right of the center. I like that on the extreme left, some visually evident clusters previously missing are now apperant, adding more differentiation to economy. A bit above the center, to the bottom right of culture clusters, there are now two small clusters with *lifestyle* topics (življenje, otroci, Ljubljana).

To summarize, I would say that the advantage of k-means is that since it divides the space into similarly sized clusters, every area of the plot can be explained in at least some amount of detail. DBSCAN on the other hand gives better or more natural borders between clusters and finds some small but prominent clusters that kmeans misses due to in general cluttered nature of the plot. It however merges most of the bottom left quarter into one cluster, but we saw from kmeans that it makes sense to divide this area into different categories. If there was more separation in t-SNE space, k-means would probably perform better.

t-SNE of TF-IDF with DBSCAN clustering and most important keywords (TF-IDF) for each cluster



Mostly based on the first plot, by merging relevant clusters as described by keywords the top level categories are (approximately looking clockwise from top left) culture, lifestyle, sports, international, climate, politics, economy, most with several subcategories as described above.