FROM TEXT TO POINTS

Which fields were used. Labels were not included, but used as some reference, which does not necesarrily hold weight. Perhaps informatively EXTRACTED KNOWLEDGE

EMBEDDING

TFIDF and SBERT perform best. Comparisons later down

Distances ... can normalize to achieve cosine similarity ... DBSCAN does not support cosine out of the box but no real difference ...

Reduce to 25D Why?? TFIDF really important because of speed Sparse, high-dimensional vectors but also because of noise and important features extraction Curse of dimensionality ... distances less meaningful

UMAP VS TSNE DIMENSIONALITY REDUCTION

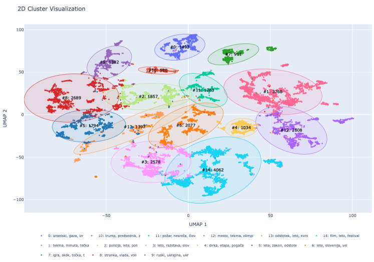
I was choosing between two popular and promising techniques UMAP and TSNE. Umap. TSNE also doesnt really make sense to use for clustering.

Also better parameters on UMAP. TSNE only perplexity ... which emphasizes everything

Speaking of determinism i would like to say that even umap i used is not necesarrily deterministic. svd or pca on the other hand for dimensionality reduction are deterministic, but did not produce better results than umap, since umap preserves nonlinear relationships. THIS FOR 25DIM

Visualization TSNE was not clear enough, increasing preplexity just emphesizes these aggressiveness ... not the goal. TO SAY TSNE might be good for inter-cluster analysis TO SAY Might be better for analysing inside clusters ... not interested ... i saw someone have success here

Noise is scattered randomly ... a lot of overlap UMAP more smoothly



sbert dbscan

Picture of tsne Here we see that tsne can just overemphasize the clustering and does not necesarily care about global distances. When perplexity is lower and clusters are not as tight, clusters start to overlap ... SI LAHKO PREDSTAVLJAS SAY Je pa lahko tud cis okej na kksnem tfidf

TSNE Distances between clusters are not neccesarly preserved, which is good with umap (but umap also pushes them appart sometimes?)

Cluster explanations in explanations.txt for this example. Primer na prosojnice samo to

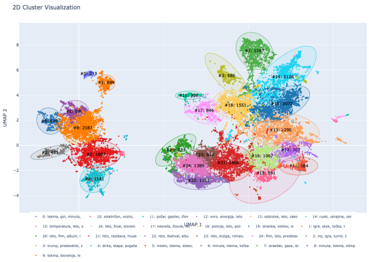
CLUSTERING

DBSCAN VS KMEANS

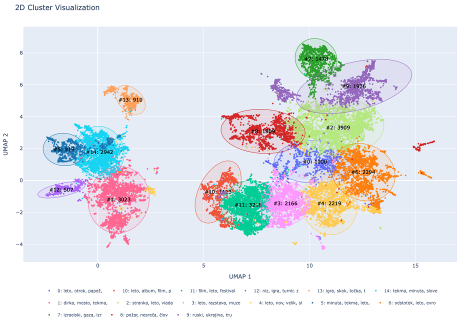
dbscan i wanted to preserve outliers

DBSCAN produced semantically similar clusters as kmeans, just that with dbscan i had to use knn to classify outliers (my choice). With kmeans, which is also much faster clustering was semantically similar and coherent, getting also the highest scores..., but the main choice was the ability to set number of clusters explicitly. DBSCAN identified 26 clusters, out of which most are easy to interpret, but some are really not easy and too similar. For that kmeans was used. The downside of kmeans is that it is naturally not deterministic, which means that some clusters will fall apart on some restarts and viceversa. However, with testing, all decompositions were semantically correct, so here maybe even some knowledge can be gained like, ... movies, music, literature ... sometimes generalizes into culture.

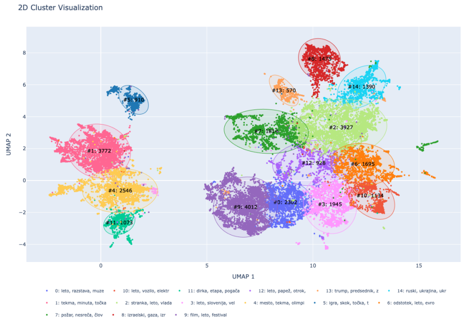
Dolocit kateri clustri so mal ambiguous in jih je morda kasneje smiselno združiti na roke



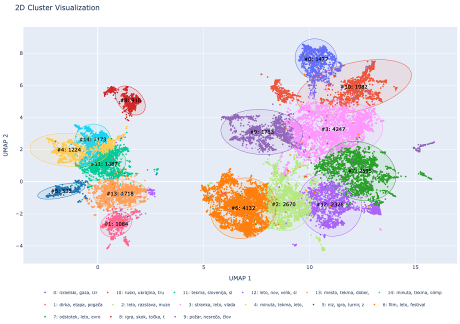
sbert dbscan



sbert kmeans clustering 0



sbert kmeans clustering 1



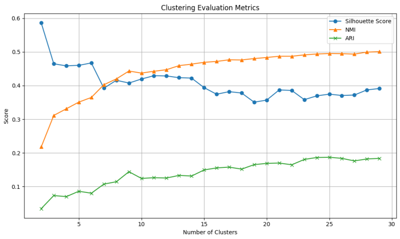
sbert kmeans clustering 2

DBSCAN: Forms clusters based on local density. UMAP might not "pull apart" similar dense areas because they're connected.

🔵 KMeans: Assigns every point to a cluster (no noise). Enforces globular (spherical) clusters in feature space. Might - Clusters are clean and non-overlapping by design. KMeans uses global centroid distance —

UMAP might exaggerate this.

Določitev clusterov:



sbert kmeans number of clusters

EVALUATION

D Interpretability: After clustering, inspect top terms (via TF-IDF within clusters) to qualitatively evaluate if clusters make sense thematically or semantically. D Human-in-the-loop check: Show random samples from clusters and verify if similar articles are indeed clustered together. Evaluate the quality of cluster labels or keywords extracted from each embedding.

Gensim Coherence model ... can be used for keyword extraction shown later NPMI Output Coherence scores per cluster (e.g. NPMI) Similarity stats per cluster (intra/inter) Tko kt semantic similarity je pac precej visok povsod, ker clustering dela dobro? Zato uporabljene tiste tri metrike, predvsem pa silhueta in analiza ter vizualnost + !!Semantic/interpretability evaluation: Check how easy it is to interpret clusters.

Explain metrics Maximum scores

Metoda vložitve	**Silhueta(25D)**	**Davies-Bouldin**	**ARI**	**NMI**	**Poravnanost vložitve**	**Povprečni NPMI**
SLOBERTA-DBSCAN	0.310	1.026	0.294	0.564	0.999	0.091
SBERT-DBSCAN	0.319	1.009	0.232	0.495	0.999	0.087
TFIDF-DBSCAN	0.325	1.019	0.178	0.424	0.999	0.095
SBERT-KMEANS	0.366	0.995	0.232	0.495	0.999	0.090

Semantic/interpretability evaluation: Check how easy it is to interpret clusters.

TODO: move to README: koda za ta eval

EXPLANATIONS

KEYWORD EXTRACTION

a plus of TFIDF - getting this while using sbert metode

✅ Example Use: Gensim's CoherenceModel You can do something like:

✅ Compared to TF-IDF Feature TF-IDF PMI/NPMI Focus Individual word frequency Word pairs and co-occurrence Pros Simple, fast Captures topical coherence Common artifacts Words like “dejala”, “leto”, etc. Fewer generic terms Interpretability Medium High (shows word relationships) Language dependent?

Somewhat Less so — works on token co-occurrence DO TU -------- basically poves da se isto obnasata, ker so semanticno kr tight te clustri in je jasno kere besede occurajo najbl TODO: check ... ce sploh hoces govort o tem

D Stanza + Lemmatizaion D To achieve explainability comparable to TF-IDF: Use SBERT embeddings to cluster semantically meaningful groups. Then apply TF-IDF or KeyBERT keyword extraction to each cluster independently, gaining TF-IDF-like interpretability at the end.

D On preprocessed tokens extract keywords

SUMMARY

it was decided that the use interpret from given collection of keywords, since LLMs are not deterministi enough. On keywords they hallucinate, on leads they overfixate on specifics.

VIZUALIZACIJA

importance podat velikosti gruč. Tukaj bi lahko tudi združil katere grupe v postprocesiranju, vendar je currently avtomatsko Ključne besede... podaj na hover top tfidf, top keybert ... each in its line

Tukaj govoris potem o tem explainabilityju ... dejanskem importancu pokazat vse skupaj, da lahko uporabnik on the spot vidi in si predstavlja kaj je v tej gruči. NPMI se izkaze za zelo podobnega TFIDF zato ga v vizualizaciji ni

Alphashape konture niso najbolj smiselne. Zaradi robov, ki itak niso sharp: Zato elipse izracun elips, da je hover clean

Tukaj tudi pomemben pol noise kar je še en razlog zakaj sem se odločil za kmeans in umap, ker ga je blo še najmanj, pa da je vseeno vse semantično blizu in da se da na vizualizaciji gledat tudi relativno bližino (umap to respecta načeloma).

Potem lahko pogleda v dejanski izpis ... text file