Intuitive Semantic Graph Tool for Enhanced Archive Exploration

Maria Teresa Artese, Isabella Gagliardi IMATI – CNR, Via Bassini 15, Italy
email: {teresa, isabella} @mi.imati.cnr.it

The work

To define a pipeline able to create semantic graphs in an Intangible Cultural heritage archive with two aims:

1. To offer a global view of the content of the archives for any users;
2. To enhance traditional ways of searching and browsing data on the web.

Innovative way to visualize the contents of an archive, as a multilevel graph

Characteristics of the approach

- The pipeline is completely unsupervised
- Semantic graphs are layered, in case of very large archives
- Use of pretrained language models makes it suitable also for a few hundred items
- Multilanguage o language specific pretrained language models make it suitable for documents in English, Italian, French, even mixed.

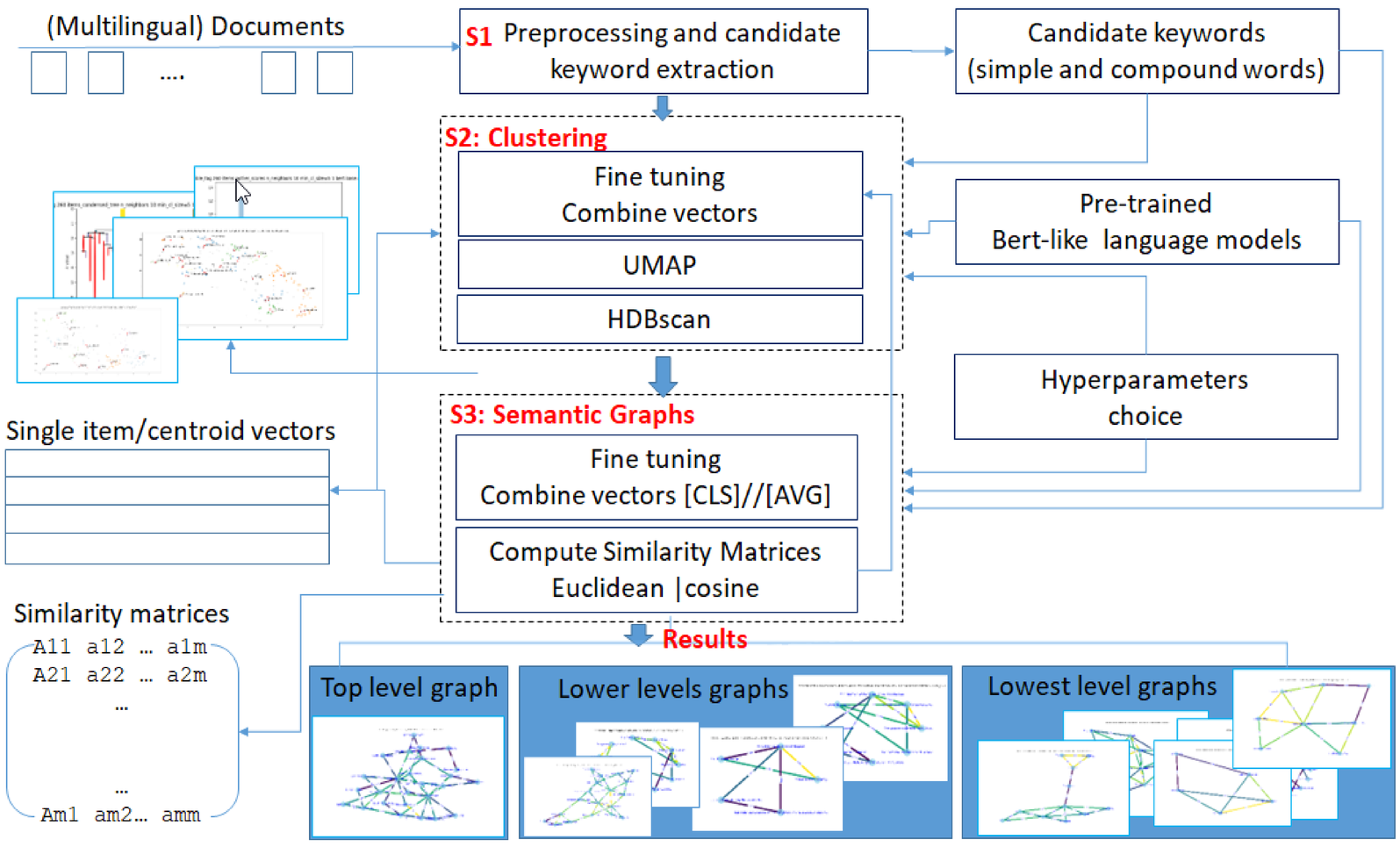
Innovative elements of the pipeline

1. **UMAP** is a dimensionality reduction technique. UMAP uses manifold learning for mapping high-dimensional data to a lower-dimensional space while preserving the local structure of the data
2. **HDBSCAN** is a clustering algorithm that does not require the user to specify the number of clusters or the size of the neighbourhood to be searched. It automatically detects the number of clusters and the shape of clusters
3. **Transformers** are a type of neural network used in natural language processing. Transformers use self-attention mechanisms to focus on different parts of the input and capture long-term dependencies.

The Data used in the experimentation from QueryLab

- A portal designed to manage intangible cultural heritage data.
- Textual metadata from title, tags, description, …

The pipeline for the unsupervised creation of multilayer semantic graphs



 # Task 1: Dataset Preparation
- Preprocessing
- Extract items to be used

Output: items of interest

 # Task 2: Items clustering
- Choice of transformers and pre-trained models
- Fine tuning of pre-trained Bert-like models to obtain the vectors
- Choice of hyperparameters for UMAP and HDBSCAN

Output: centroids of clustered items, and elements of each cluster

 # Task 3 Semantic graph creation
- Choice of transformers and pre-trained models
- Creation of similarity matrix using [AVG] or [CLS] tokens

Output: Semantic graphs con k most similar items, with k=1…4

Hyperparameters for UMAP and HDBSCAN

- n_neighbor (UMAP) : 20,15,10,5.
- min_cluster_size (HDBSCAN) 15,10,5,
- min_samples (HDBSCAN) 15,10,5,1

Choice of transformers and pre-trained languages

- BERT Base
- DistilBERT
- MiniLM-L6-v2
- Bert-base-Wikipedia-sections-mean-token

Single language or multilanguage, from Google, Hugging face, Microsoft, …

Qualitative evaluation

We gathered feedback from heritage experts and web users:

- is the clustering and similarity matrix able to extract the significant elements?, and
- Do users find browsing the archive via graph interesting and useful?

Positive results:

PROs: simplicity and usability of the graph visualization
CONs: low-level clusters contained elements that were not closely related, or that some related elements were spread across multiple clusters

Conclusions

- Effectiveness of the pipeline in generating meaningful semantic graphs
- Positive evaluation from users, but graphs with more than 30 nodes (overly dense graphs) are difficult to understand and navigate

www.imati.cnr.it
arm.mi.imati.cnr.it/querylab
arm.mi.imati.cnr.it/mislab

arm.mi.imati.cnr.it/papers/ht2023 for additional materials

Future works
- tools to traverse the graphs
- fish-eye views to overcome the overly dense graphs
- experiments on other datasets

Prototype developed in Python
using standard packages like Numpy, Matplotlib, Pandas, more specific ones for processing of textual data such as NLTK , Gensim, Scikit-learn, Spacy, and packages for managing pretrained language models as Torch, Tensorflow, Transformers,…