UNIOR NLP @ COFE Datathon 2022
Task 2 - Idea clustering

Gennaro Nolano, Maria Pia di Buono, Johanna Monti
UNIOR NLP Research Group
Unversity of Naples ”L’Orientale”, Italy
{gnolano,mpdibuono,jmonti}@unior.it

Abstract

This paper presents the methodology proposed by the UNIOR NLP Research Group for Task 2 - Idea Clustering of the COFE Datathon, a data mining competition organised by the European Commission in the context of the Conference on the Future of Europe (COFE).
We apply a clustering approach relying on a keyword extraction and a graph representation of each idea proposal which includes a category assignment derived from EuroVoc classification. Subsequently we build a graph network and apply a similarity measure between nodes to cluster more similar idea proposals.

1 Introduction

Within the COFE Datathon[1], the second challenge aims at clustering idea proposals from conference open data, namely data provided by users and collected through the COFE digital platform.
The task can be framed as a topic detection process, by whom prominent topics within the idea proposals collection are found and then used to measure the similarity among idea proposals to build clusters of proposals. We perform the topic identification by extracting keywords considered representative of each proposal and then developing a graph-based representation used to detect communities of proposals.

2 Dataset

The dataset collected through the COFE digital platform includes: (i) about 17k idea proposals in several languages, (ii) an EN translation when needed, (iii) a user-generated category to classify the proposal, (iv) scope, (v) endorsement, (vi) anonymized users’ information, (vii) comments to the proposal. The provided dataset is characterized by several aspects that influence the workflow, those are:

- Multilingual data;
- Users’ generated data which may contain typos and a lack or a misuse of domain terminology;
- Machine Translations in English (thus, susceptible to translation errors, mainly for under- represented languages)
- Lack of alignment between the source content and the English translation.
- Language classification of post content can be inaccurate, as it is not based on automatic language identification but on other users’ information. Thus, despite the provided language classification, some of the posts are actually written in English.
- Comments to the proposal without EN translation.

3 Methodology

The proposed workflow is based on four steps:

1. Keyword Extraction
2. Graph-Based Proposals Representation
3. Proposal Network Construction
4. Communities Detection

Keyword Extraction This is step is performed by means of the Python Keyphrase Extraction (PKE) module[2], which returns weighted results for the extracted keyphrases. In particular, we

[1] https://futureu.europa.eu/pages/datathon?locale=en

[2] https://github.com/boudinfl/pke

make use of the MultiPartite Ranking algorithm (Boudin, 2018).
The keywords thus extracted are then re-ranked according to their closeness in a conceptual graph. In particular, such a graph is created by fetching hierarchical information from the EuroVoc vocabularies[3], by looking for its top-3 closest matches through the EuroVoc API[4]. For languages not supported by the API, we refer to the keywords extracted from the English translation.

Graph-Based Proposal Representation Each proposal is then represented as a directed graph, starting from a root node for the proposal, connected to a node representing its COFE category. The latter is then connected to a node representing its mapping to an EuroVoc concept categorization. This graph is then connected to the graph-based keywords representation built in the previous step.

Proposal Network Construction Once we extract such a graph for each of the proposals, we construct the union of these to create a single, unique graph which comprises of the information for all the proposals, their categorization according to COFE and EuroVoc, and the conceptual information of their keywords.
We extract on the nodes referring to proposals, and the shortest paths connecting them (when present), and construct a graph of proposals as exemplified in Figure 1.

Communities Detection Using the graph generated in the previous step, we create communities by employing Clauset-Newman-Moore greedy modularity maximization (Clauset et al., 2005) through the Python package NetworkX[5].
Using this model, we detect the amount of communities needed, and then extract from these communities all the nodes representing proposals.

4 Results and Conclusion

We present the workflow used to cluster idea proposals, based on a graph representation which allows communities detection.
The results of our workflow have been submitted to the COFE datathon in the form of three separate tsv files, for 10, 20 and 50 communities respectively. In each file, each column contains the id of a proposal and the id of the community it belongs to. The results, in this form, have been evaluated by the organizers.
In future works, we aim at integrating further external knowledge, in particular with the objective of making the workflow usable for more languages than those available through the EuroVoc API used in this work. Furthermore, we aim at exploring other combinations of features (e.g., different algorithms for community detection and different algorithms for the network costruction) and evaluate how they would affect the final results. We also plan to evaluate our results against a baseline.

Acknowledgments

Maria Pia di Buono has been supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione e Mobilità Internazionale dei Ricercatori” Avviso D.D. n 407 del 27/02/2018.

References

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs.

Aaron Clauset, M Newman, and Cristopher Moore. 2005. Finding community structure in very large networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 70:066111, 01.

[3] https://op.europa.eu/en/web/eu-vocabularies for the most similar matches for each of the authomatically extracted keywords.
EuroVoc data, classified according to a taxonomy and available through an API, are a multilingual and multidisciplinary thesaurus covering the activities of the EU and containing terms in 24 EU languages.

[4] https://www.vocabularyserver.com/eurovoc/
As of the time of writing this report, the API supports only 12 of these languages (Bulgarian, Spanish, Czech, German, Greek, English, French, Italian, Dutch, Polish, Portugues and Slovenian).

[5] https://networkx.org/documentation/stable/index.html

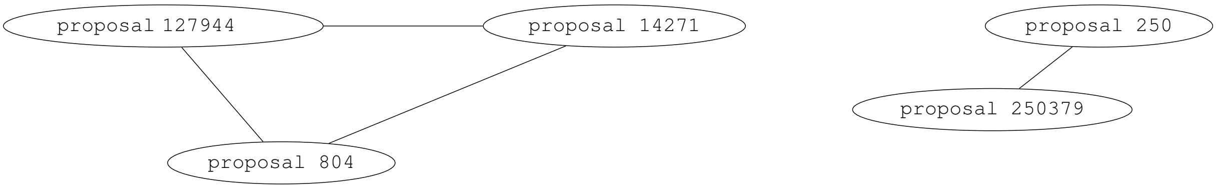

Figure 1: Example of a graph used for keyword extraction. Dashed edges indicate the presence of a path connecting the two nodes.