COFE Datathon - Make.org

Overview

Our Submission for the 5th challenge of the COFE datathon provides an interface to answer with a few clicks the question : “What are citizens proposing for the future of Europe” ?

Our interface thus focuses first on providing clues to get a rapid overview of the main topics and ideas supported by the citizens in their propositions. We offer an entry by three complementary axes : the topics of the COFE, top keywords and top ideas. The first is given by the structure of the COFE platform, the latter two are based on semantic algorithms that we will describe later.

The real power of the interface reveals itself when digging into the data. By choosing a first axis entry by topic, keyword or idea, the interface is updated and allows a cross-analysis along the two other axes. The obvious drill is to look at the main keywords and ideas by topics. But one can also then see for instance what are the keywords associated with an idea, and how the proposals of the idea are spread across COFE topics. You can even go deeper and choose a keyword strongly associated with an idea, then look back at the ideas associated with it to reveal ideas close to the original one.

Since a demo is better than a thousand words, our interface is publicly available here :

https://dial.make.org/future_eu

General presentation

Our target is to provide an interface for people that do not speak 24 languages, that is for most of us. For this reason, we have worked on the English translations of the citizens’ propositions and we display English keywords and English names for the ideas.

Our interface is composed of 6 visualization modules that we will describe next.

Top keywords

The first module displays the top keywords associated with the selected propositions. The size represents the number of propositions associated with the keyword.

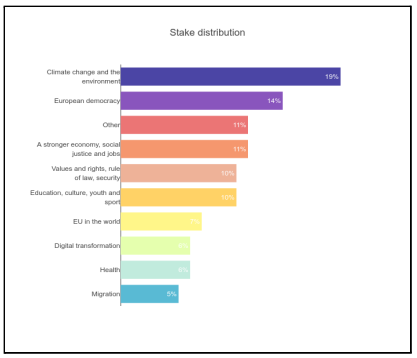
Our methodology to extract the keywords from the citizens’ propositions is as follow:

- For each proposition, we first extract the nominal chunks (bi and tri grams), the named entities and nominal tokens.
- Each extracted gram is scored using multiple groups of features:
    - Consultation based Frequency features
    - External corpus based frequency features

    - POS tag based features
    - Wikipedia based features
- Keywords are then sorted using a score combining these features. The 5 best grams are selected as keywords for each proposition
- The keyword cloud is then built by counting the frequency of the selected keywords

You can see the keywords extracted from each proposition in the table at the bottom of the interf ace.

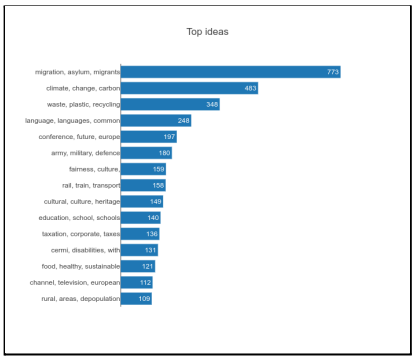
Topics



This module displays the distribution of COFE topics within the selected propositions. The percentage is the share of propositions associated with the topic within the selected propositions.

The topics have been reconstructed by parsing the proposition URLs, since only the sub topics were provided in the data file.

Top Ideas



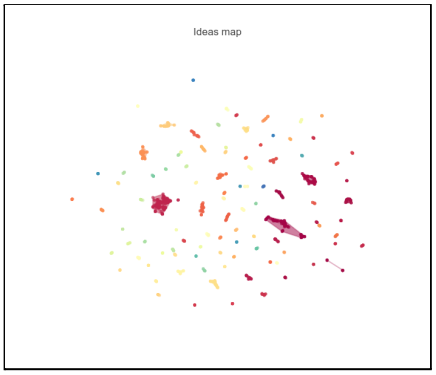
This module displays the weight of the top 15 ideas within the selected proposition. The weight is the number of propositions associated with the idea.

Ideas are clusters of propositions based on the following clustering approach :

- The propositions are vectorised using S-BERT multilingual embeddings
- The vectorised propositions are projected with UMAP into a subspace of 150 dimensions
- The propositions are then clustered using the HDBSCAN algorithm
- The cluster are named by the best 3 words extracted from the title of the propositions and weighted by TF-IDF

You can see the idea associated with each proposition in the table at the bottom . For the sake of clarity, only the 100 best ideas have been kept.

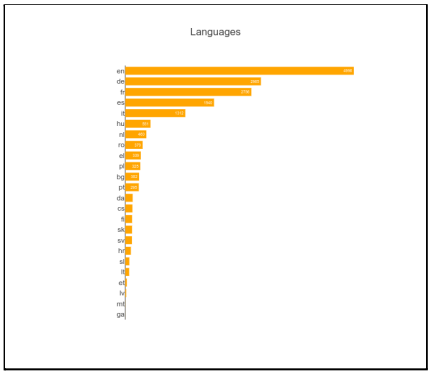
Top ideas Map



The fourth module displays a projection of the 100 top ideas in a 2d plan.

The underlying projection is made with UMAP. The propositions in the same idea share the same color, and a convex hull is shown to make the ideas span more visible. The closer the propositions, the more similar they are. This is not true for the clusters since with UMAP only small distances are meaningful.

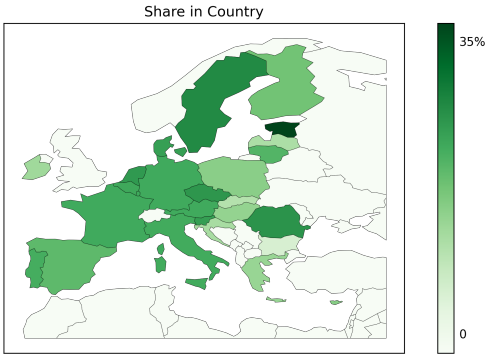
Languages



This module displays the distribution of language of the selected propositions. It provides a reference of volume to make the interpretation of the next module easier.

The graph displays the number of propositions associated with the language within the selected proposals.

Country Map



The last module is meant to provide an insight about how strong a topic, keyword or idea is linked to some specific countries or not.

The map displays the percentage of propositions from the country in the selection compared to the total number of propositions from the country. The intensity is therefore comparable between one country and the other.

Warnings:

- the map emphasizes the contrasts between countries, check the scale to see if the percentages are sizeable
- a small country that displays a high percentage can amount for only a few propositions. Check the language distribution to have an idea of the volume involved. (For instance Estonia on the map above)

The country of the proposition was not provided in the data, we had to infer the data from the closest proxy : the language.

For most countries, there is only one language spoken and this language is used only in this country. But there are exceptions that we have handled as follow:

- English has been used by many citizens for whom it's the natural international language, English is therefore linked to no country (thanks to Brexit, this is not a too big issue)
- Belgium and Luxembourg are multi-languages countries. The percentage of propositions in those countries has been computed as the average of their official languages, French and Dutch for Belgium, French, Dutch and German for Luxembourg.
- German is spoken both in Austria and Germany, they therefore share the same percentage
- Ireland speaks Irish and English but since English is not country specific, only Irish counts for Ireland (alas, there are only 2 propositions in Irish)
- Some countries have very few propositions, to avoid misleadingly high percentage values, we used a bayesian estimate. The prior is based on the English propositions, used as a proxy for the European average.

In the table at the bottom of the interface, we have shown the inferred country for each proposition. The rule here is even simpler : a proposition is associated with the main country speaking its language. There is therefore no proposition associated with Belgium, Luxembourg, Austria and Cyprus.