Big Data - Visualization Tool

Daniel Mínguez Camacho, Yoselin Garcia Salinas

December 9th, 2018

1 Problem characterization

Since the beginning of the master, we have talked about the European commission program 'Horizon 2020' in several courses (Innovation & Entrepreneurship or Big Data for example).

This is the reason why we wanted to take advantage of this assignment in order to to dive a little bit in the different projects, organizations and topics that are part of it and know how much money it is used in this kind of projects.

We have approach this assignment by conceiving our potential users as citizens from the European union looking for information regarding H2020. We have tried to answer the potential questions that they could have regarding the different projects, budget, organizations and researchers that are part of this objective 2020.

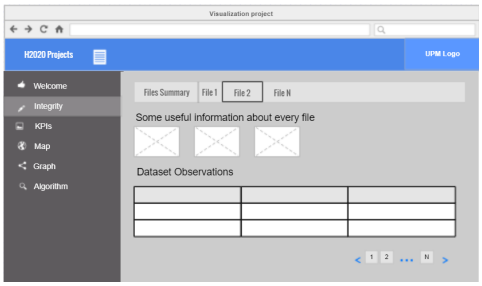
- **Regarding the budget:**
  - Which is the total budget dedicated to this projects?
  - How is the budget divided between countries?
  - Has the budget evolve during years or has it maintain constant?
  - Would be possible to start these research projects without European funding?
- **Regarding organizations**:
  - Which organizations work together in some projects?
  - Which organizations have more researchers?
  - Where are these organizations located?
  - Which are the coordinators of these projects?
- **Regarding projects:**
  - How many projects are being develop in each country?
  - How are the projects distributed along a specific country?
  - Which are the countries with more projects?

  - How many projects started in a year in each country?
  - How many projects started in certain period in each country?

- **Regarding topics:**
  - Which topics are more recurrent?
  - Which topics receive more funding?
  - How topics are related with one another?

After this first approach to the problem, we tried to look for this information in the CORDIS web page here in order to know more about the elements present in our dataset. We have realized that a world map that shows the organizations part of the H2020 is available here and that anyone can look for projects in this link.

We have also analyzed the different datasets that we have chosen (explained in second point) as detailed in the "Input files" tab, in order to have a first approach, we decided to include this analysis in the visualization as in this design:

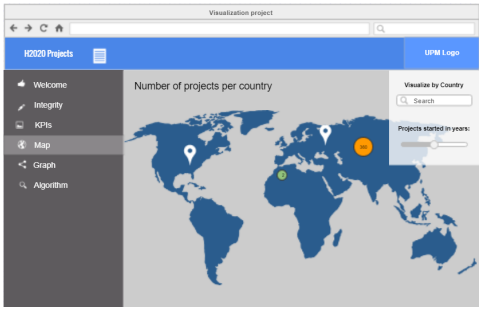


After analyzing the information that we have available and compare this information with the problems that we want to address, we have decided to do the following:

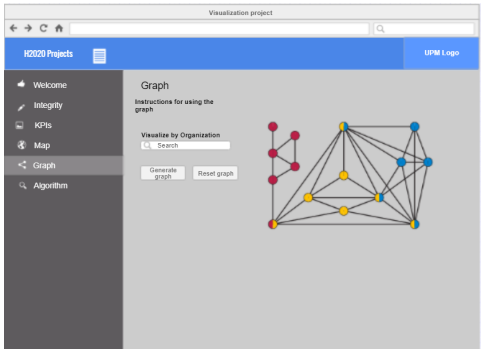
- Perform a **first approach analysis** for each file as detailed in the "Input Files" tab, here it is possible to find some first approach visualization as:
  - **Organizations:** Pie charts representing the contribution per country and the roles of the organizations in projects.
  - **Projects:** A barplot representing the European contribution per year and a scatterplot representing the logs of the European contribution to each project versus the total cost, that represents that almost any project would have been possible without European funding.
  - **General:** It is possible to find tables with the details of each file.

- **Represent the funding that each country receives**, since little information is available about this topic in the web at a first stage. In order to do this we have decided to create a treemap representing the number of projects and the funding received per country, including more details about which regions/organizations receive this funding. This is the first design we did:

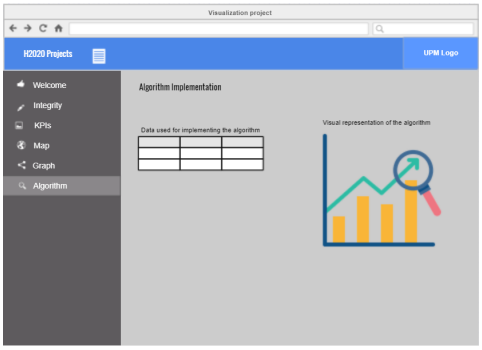
- The map available in the web represents all organizations across the world, but we considered more interesting to know **where are located the organizations that are coordinators** of some of the projects, in order to know if entities outside Europe can coordinate and request projects, in order to do that, we decided to elaborate a map locating the different coordinators:



- In the map available in the web it is possible to know **which organization works with another**, but in some cases it becomes unfeasible to follow the connections (for example if you go to Madrid you can see around 300 dots in Puerta del Sol). This is the reason why we chose to create a Graph, including also researchers, in order to make this analysis easier.



- Finally, in the web page we have several classifications regarding topics (by country, by general topic), but unless you know what you are looking for, it is difficult to have a **global overview of the different topics**, this is why we decided to make a TF-IDF scatterplot, showing the 10 more relevant words for each topic visually.



2 Dataset chosen

We found the following files in the European Data Portal[1]:

- **cordis-h2020projects.xlsx**: Projects part of Horizon 2020.

Field	Type	Field	Type
rcn	Categ/Num	objective	Text
id	Categ/Num	totalCost	Categ/Text
acronym	Categ/Text	ecMaxContribution	Categ/Num
status	Categ/Text	call	Text
programme	Categ/Text	fundingScheme	Numeric
topics	Categ/Text	coordinator	Text
frameworkProgramme	Text	coordinatorCountry	Text
title	Categ/Text	participants	Text
startDate	Date	participantCountries	Text
endDate	Date	subjects	Text
projectUrl	Text	-	-

- **cordis-h2020organizations.xlsx**: Organizations that participate in at least one of the projects.

Field	Type	Field	Type
projectRcn	Categ/Num	city	Categ/Text
projectID	Categ/Num	postCode	Categ/Num
projectAcronym	Categ/Text	organizationUrl	Text
role	Categ/Text	vatNumber	Numeric
id	Categ/Num	contactForm	Text
name	Text	contactType	Text
shortName	Text	contactTitle	Text
activityType	Categ/Text	contactFirstNames	Text
endOfParticipation	Date	contactLastNames	Text
ecContribution	Numeric	contactFunction	Text
country	Categ/Text	contactTelephoneNumber	Numeric
street	Text	contactFaxNumber	Numeric

[1]The data was obtained from EU Open Data Portal. Also available here

- **Researchers.rds**, this file has been created by concatenating the followings:
  - **cordis-h2020-msca-fellows.xls**: researchers working on projects funded by the Marie Skłodowska-Curie actions under the Horizon 2020 framework program.
  - **cordis-h2020-erc-pi.xlsx:** This dataset lists the Principal Investigators (PIs) working on projects funded by the European Research Council (ERC) under the Horizon 2020 framework program (H2020).

Field	Type
projectId	Categ/Num
projectAcronym	Text
fundingScheme	Categ/Text
title	Text
firstName	Text
lastName	Text
organisationId	Categ/Num

- **cordis-h2020reports.xlsx:** Reports of some of the projects.

Field	**Type**	**Field**	**Type**
rcn	Categ/Num	country	Categ/Text
language	Categ/Text	projectID	Categ/Num
title	Text	projectAcronym	Categ/Text
teaser	Text	programme	Categ/Text
summary	Text	topics	Categ/Text
workPerformed	Text	relatedFile	Text
finalResults	Text	url	Text
lastUpdateDate	Date	article	Text

We also have used external data in order to retrieve the geographical location of the different organizations.

2.1 Integrity

We have checked the integrity of the data by joining the different data sets, obtaining the following (We have removed incongruous records from our analysis).

- **Regarding Organizations:**
  - There are 7 Organizations id in the researchers files that does not appear in the Organization file, so we could not retrieve their names.
  - There are some organizations with different Organization ID but the same description, we treat them as different organizations.

- **Regarding Projects:**
  - There are 59 projects in the researchers files that does not appear in the Projects file, so we could not identify them.
  - There are some projects without date, we have include them as NA value for the date.
- **Regarding Researchers:**
  - There are 158 projects in the researchers files that does not have any researcher associated, we assume this is not a problem since only some relevant researchers are included.
  - There are duplicate researchers that vary only in the title field ('PROF' and 'DR' for example).
  - There are 1.859 researchers without organization, again we assume this is not a problem.
- **External data regarding geolocalization:**
  - We had some problems locating some organizations and we set up their location in the capital of their country.

2.2 Assumptions:

Based on the information that we have available, we have made the following assumptions in order to work with it:

- All organizations that appear in cordis-h2020organizations.xlsx are part of some project.
- Reports are not mandatory for all projects or most of the projects have not sent yet theirs (Or they have not been published).
- There are organizations with different id and the same description, we suppose this is because they belong to different projects and/or they have different people in charge.
- There are projects with different id and the same Acronym/Description, again we suppose this is because they have different stages or branches and/or different persons in charge (or even different projects with the same acronym).

3 Development

3.1 Map

In order to answer the questions regarding the project's development along different countries, we decided to use an interactive Map as visual representation. In the following sections, we are going to describe the decisions made for the design and development of this visual representation.

3.1.1 Data tasks abstractions

- **Why?** Our main goal is to provide to the users an **easy way to explore and summarize** where the projects are being developed along the world in order to incentive them to create new hypotheses based on the information displayed.

  Our users want to make a locate search because they already know in which countries, they are interested in, but they don't know where to get the summarized information.

  Actions: On one hand, the user should be able to identify the number of projects coordinated in every country, as well as perform comparisons over the quantity of projects developed among different countries.

  On the other hand, the visual representation should show to the user how many projects are coordinated in a specific country and how those projects are distributed in the different regions in the country. The user also can have information about the location and the universities involved in the development of the projects.

  Additionally, the user should be able to explore the how many projects started in in every country in certain period. Finally, we want to provide to the users the possibility of stating new hypotheses based on the information discovered in the visual representation.

  Targets: We want to focus on finding an individual value to be displayed in the visual representation, in this case we want to get the number of projects developed in each country across a period.

- **What?** Creation of an interactive map.

- **How?** We need to work on aggregating the number of projects by area, merging and transforming the data sets in order to get the necessary attributes, such as the name of the country, the latitude and longitude where the projects are developed, to draw the map.

  If a user wants to have information about the number of projects developed in each country across a period, we will also need to perform some filtering over the data set.

3.1.2 Interaction and visual encoding

In order to display how many projects are being developed in a country/region in a more effective way, we decided to use an interactive map instead of other visual representation because this representation is more intuitive and expressive for our users to know and to compare the quantity of projects developed in an area, as well as where are located the universities that are working on those projects.

In this way, the users can read easily relate the areas with the universities where projects are being developed.

Design Criteria

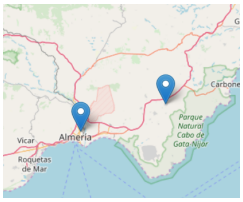
- **Dynamic interaction**

  We decided to make an interactive map in order to generate relevant views for the user and help him to get the information that he/she requires. Additionally, we want to provide the possibility to select one or multiple countries to be display, make groups based on the spatial position of the marks when zoom is done, and select one point that represents the research center where a project started.

- **Marks and Channels**

  We used two type of marks to represent the number of projects and the location of them on them depending on the zoom of the map.

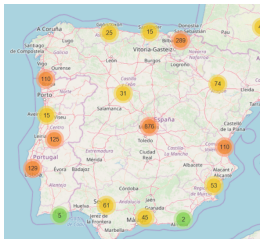
  When we zoom in over a specific area of the map, we are using the identity visual channel based on the spatial position on a common scale. In order to display where the projects are specifically located, we considered to use points marks as in the following image is represented:



When we zoom out the map, we are using a magnitude channel to represent how much projects are being developed in an area. In order to specify what the visual encoding should express, we decided to use an area encoding representation, in which we use containment marks by grouping the individual marks (specific points in the map) into clusters using the spatial proximity where the projects are located. In other words, containment marks enclosures projects based on the geographical area and summarize the number of projects in the same geographical area.

We considered to used 3 colors to represent the quantity of projects grouped by their spatial proximity. A containment mark colored in orange represents those areas which have more projects. Yellow represents that in that area there are considerably less projects that in the orange areas. Finally, green represents an area where there are less than 10 projects.

We can see an example of this approach in the following image:



3.1.3 Algorithmic implementation

For creating the map, initially we used two datasets: cordisref-countries.xls and cordis-h2020projects.xlsx.

- **Computing geolocation**

  From the projects dataset the API R library ggmap has been used to assign a default location to each of the countries present in the dataset. The problem with this when representing the projects in the map is that the amount of information that is provided is limited, as we can't differentiate projects from different cities in the same country.

  To solve the problem stated above was solved by using two Google Maps APIs and implementing them using Python. The first was the Google Places API[1] where we extracted the latitude and longitude of the universities where the project was developed. However, some universities that weren't found using this API were retrieved using the Google Geocode API [2]. This way a big amount of the universities locations was obtained. However, in some cases this location was not correct, as to obtain the coordinates a text search was being used. To correct mistakes when retrieving the location, an additional check to discard the coordinates that were from incorrect countries was implemented. This way if we retrieved for a university from Spain coordinates from Italy, we discarded them and use the default location computed in the first place.

  - [1] Google Maps API - Places search.
  - [2] Google Maps API - Geoencoding.

  After getting the location where each project is located, we created a new dataset in which these attributes are added to the corresponding row.

- **Drawing the map**

  For creating the map, we used leaflet which is a JavaScript library which has integration with Shiny. In the UI we invoke the leafletOutput, and on the server side you assign a renderLeaflet call to the output. Additionally, in the server side we provide to Leaflet the dataset with the projects to be drawn as well as the location (latitude and longitude) of each of

projects. The information about the location is found in the processed dataset explained in the previous session.

3.2 Treemap

3.2.1 Data tasks abstractions

- **Why?**
  - **Actions:** Provide to the user the possibility to compare the quantity of capital invested for the development of the projects in each country, as well as helping them to identify the relation between the number of projects develop in a country and the money invested on those projects.

    We also want to give to the user the chance to explore how the and the money are distributed along the different regions of a specific country.
  - **Targets:** Two quantitative attributes at leaf nodes. We want to focus on finding an individual value to be displayed in the visual representation, in this case our target is to display the number of projects of the different geographical areas along the world. However, we also want to show to the user how is the money investment distributed along these are. Therefore, each rectangle (leaf node) in the treemap it will show the values for those two values.
- **What?** Creation of an interactive treemap.
- **How?** We need to work on aggregating the number of projects by geography. The map it will contain different levels of granularity because it is intended to represent a hierarchy of three levels, therefore the treemap will be able to display the number of projects in different levels. First, it will show the countries with more projects and bigger project investment. If the user wants to know more about the number of projects and the investment in a particular country, he/she will be able to surf into a region or a City.

3.2.2 Interaction and visual encoding

The map is a good tool to discover where projects are located and to detect areas with big amount of projects. However, we also wanted to allow the user to easily get an idea of the number of projects per geographical area as well as its investment amount.

A treemap was a great fit, as we had a hierarchical structure of countries, regions and cities of the projects. This way using the area to represent the number of projects we could represent all the main countries together and easily grasp how the project are distributed. Then if we zoom in one country, we could

see the details of its regions. And if we zoom in a region the most granular detail per city could be seen.

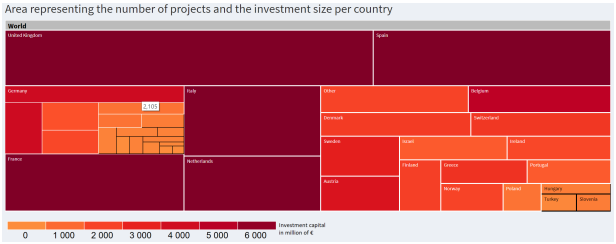
Design Criteria

- **Dynamic interaction**

  We decided to make an interactive treemap in order to generate relevant views for the user and help him to get the information that he/she requires. Additionally, we want to provide to the users the possibility of making specific queries using different levels of granularity.

- **Marks and Channels**

  As in the case of the Map, we are using containment marks order to display hierarchical data using nested rectangles. At the same time, we are using a rectilinear orientation of our elements. We are considering that the size of each node is represented by the number of projects in a geographical and the color of the node represent the total money invested for the development of those projects.



3.2.3 Algorithmic implementation

For creating the map, initially we used a modified file cordis-h2020projects.xlsx with the new attributes about the localization of the projects.

For the creation of the map idiom, we had to extract the cities, regions and coordinates where each project is located. That information was useful for the development of the treemap idiom.

- **Reducing the number of categories**

  One problem with a treemap visualization is that we had a large number of categories that contain very few projects, so the area was very small, and the text identifiers were overlapping. To avoid this, if the number of projects was less than a certain threshold then those countries, regions or cities were classified as Other. This threshold varied depending on the granularity: 100 for cities,10 for regions, and 5 for cities. For instance, if a country had less than 100 projects then that country appears in the treemap contained in the category Other. This solution made the treemap

more appealing visually, as well as more useful, as labels were not overlapping and rectangles with very small areas that couldn't be appreciated were not present in the plot.

To generate this plot, a tree structure needs to be computed and used to store all the information about countries, regions and cities.

- **Color dimension**

  The number of projects per location and with different levels of granularity was something key to identify areas with more projects. However, the amount of the investment is also very important as some areas might have more projects but with less capital invested. As we also wanted to represent this information, we decided to use a color gradient from red to yellow to represent the size of this investment, dark red representing high investments and pale yellow representing low ones. The scale chosen was thousands of millions of euros, as the original values had too many zeroes, reducing the readability of the legend.

3.3 Graph

3.3.1 Data tasks abstractions

- **Why?** The main goal of this visualization is to help to discover in an easy way which organization work with another, letting the user discover the projects and which researchers are involved in them.

  Actions: Firstly, the user should be able, given a node she is interested, to generate the graph representing the nodes related with it. Also it should be able to investigate the connections among these nodes, generating new levels of connections. At the same time, the user have to be able to identify which organizations receives more funding from which project.

  Targets: Network of projects, organizations and researchers.

- **What?** Creation of a graph visualization

- **How?** We decided to merge the files cordis-h2020projects.xlsx, Researchers.rds and cordis-h2020organizations.xlsx together in order to be able to generate the nodes and edges that are necessary to create a graph, weighting the edges by ecContribution that represents the European contribution in the project to an specific organization.

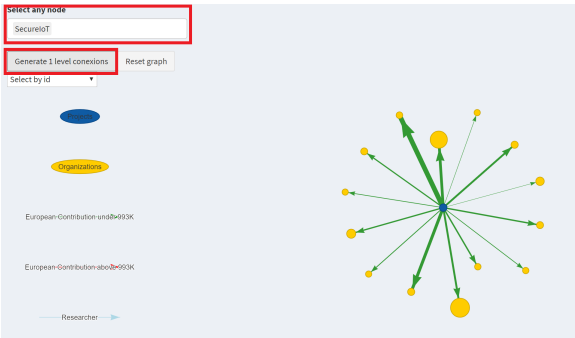
3.3.2 Interaction and visual encoding

As our purpose with the graph is to answer how the organizations, projects and researchers are linked together, we decided to do the following.

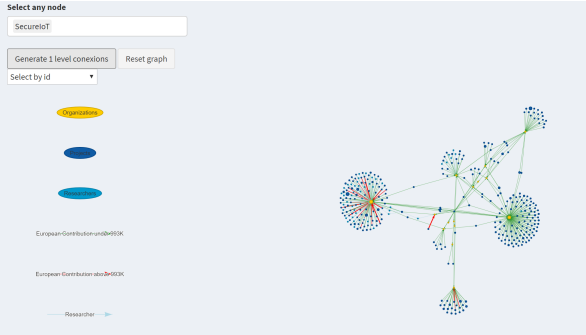
Design Criteria

- **Dynamic interaction**

  The first problem that we faced was performance, since the graph has 135K edges and 64k nodes, it was not feasible to represent the whole graph at once, this is why we decided to let the user decide which node she is interested in and plot the nodes that the original node has connections with.



Afterwards, in order to allow further investigation and details discovering, we decided to include the option of generate one more level of connections, which works recursively.



Furthermore, we include the options of interact with the nodes (in case a representation hide some of them) and zoom.

Finally, and for performance reasons, we do not include the names of the nodes unless the user hover one of them, we also hide the edges when the user move the graph.

- **Marks and Channels**

  We distinguish the nodes regarding researchers, projects and organizations with different colors (yellow and blue as the European colors for projects and organizations, light blue for researchers). The nodes have different sizes depending the number of edges (both going in and out) they have.

  The color of the edges is green, but those edges that represent a contribution considered an outlier (above Q4 plus 1.5 the interquartile range) are represented in red. The size of the edge also depends on the contribution.

3.3.3 Algorithmic implementation

In order to develop the graph, first of all, we have created the nodes and edges files from the original files:

- Nodes: The nodes are conformed by the organizations (by id), the projects (also by id) and the researchers (by the concatenation of firstName plus lastName). The weight of the node is calculated once the edges are calculated by taking the number of edges that each node has (both going in and out)

- Edges: The edges are retrieved from two sources:

  - Projects-Organizations in cordis-h2020organizations.xlsx file, from this file we also take the budget contribution for the weight of the edge. In the case of 0 contribution or the researcher, we apply a value of 1 (minimum).
  - Researchers-Projects and Researchers-Organizations in Researchers.rds file

As the number of nodes and edges is large, we have decided to generate the graph in an interactive way as we have explained in the previous point.

For that, we have use the visNetwork library with a igraph representation in order to improve performance, the main characteristics are:

- Type of graph, after trying several layouts provided by the igraph library, we have decided to work with the layout_nicely representation, which choose an appropriate layout for a given graph.

- Options to improve performance, as detailed before.

- We have observed that due to the fact that some organizations/projects have different ids with the same names, sometimes when a node is selected, the graph will be unconnected, but we have decided to maintained it this way.

3.4 TF-IDF scatterplot

3.4.1 Data tasks abstractions

- **Why?** This tf-idf calculation and representation is used in order to give a graphical representation to the main ideas and concepts developed by the the different projects across the topics.

  Actions: The idea is to have a visual representation of the main words along the different topics. This will allow the user to find which ideas are more recurrent and use them to access the topics in which this words appears. Then the user should be able to filter the projects and reports files (in the tab 'Input Files') to find out the details of the projects associated to these topics.

  Targets: Most relevant words per topic according to tf-idf measure.

- **What?** Creation of a scatterplot visualization, followed by an interactive way of discover different topics.

- **How?** We have decided to calculate the tf-idf based on the tittle of each project, we took the description of the project as the text. After the calculation we extracted the 10 words with higher tf-idf per project and join them to the project topic. The we would display a sctatterplot showing the average tf-idf per word across the different topics and the number of topics in which the word appears. this would be associated to a way of find out which topics have a word, which will allow the user to find the projects associated to this topic in the projects file.

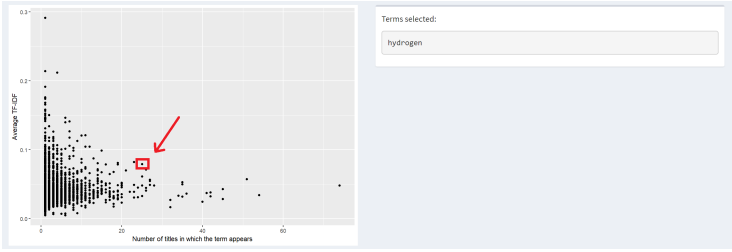
3.4.2 Interaction and visual encoding

Design Criteria

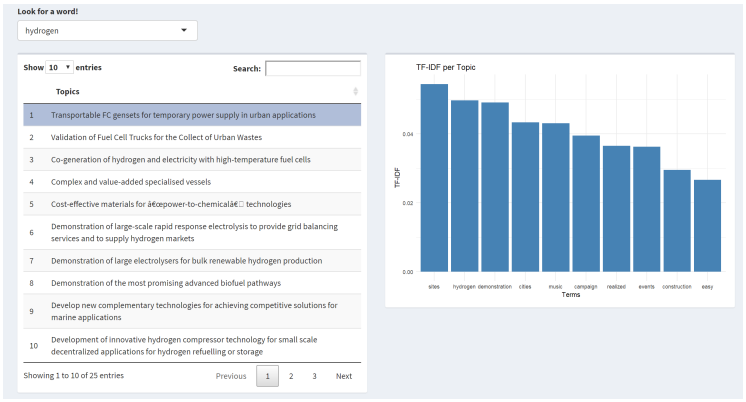
- **Dynamic interaction**

  We split the interaction in two modules, the details can be access in the "Input files" tab:

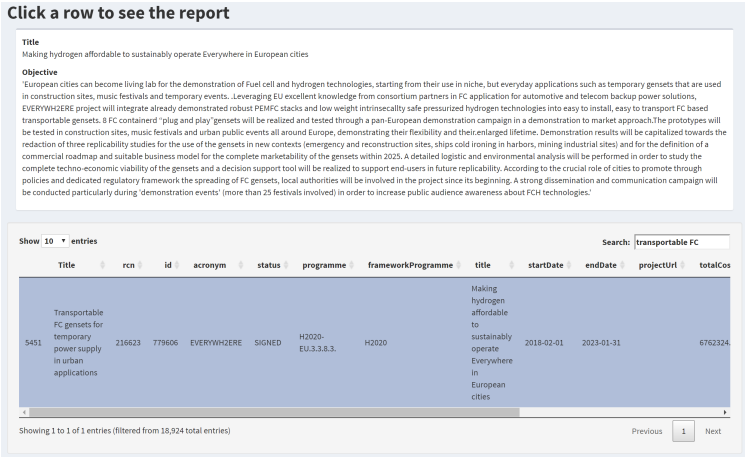
  - **Dynamic scatterplot:** With the dynamic scatterplot the user should be able to find which words have higher tf-idf across multiple topics and it will help the user search:



– **Word searcher:** Here the user will be able to look for a word and find in which topic its relevant, the user then can access to the topic and see which other words are relevant in it too.



– **Details:** The details will be found in the 'Input Files' tab, at the projects or reports sub-tab:



- **Marks and Channels**

  In this case, we just have used a relatively simple scatterplot with dots representing the different words that have been found, allowing the user to select the dots when she click in any part of the scatterplot.

3.4.3 Algorithmic implementation

We have decided to use tm and tidytext libraries for perfoming the tf-idf calculation. We have calculated the tf-idf for each title offline, the text elements

were the concatenation of the title and the description of each project.

After computing the tf-idf we have decided to extract the ten with the higher score and calculate the average tf-idf among the documents, computing the scatter plot plotting the average versus the number of tittles in which it appears.

4 Conclusions

We chose this dataset because we were interested in knowing more about Horizon 2020 and the projects involved on it.

It has been challenging afterwards the process of thinking in which way we could contribute more to the current web CORDIS has available and the process of transforming the dataset available in order to fit the requirements we came up with.

It could be interesting to maintain the code and try to use the next data available for the following projects regarding research and innovation in the EU.