Climate Maze

Data Visualization

Team ESNAware

I. Introduction

Overview

The world is currently going through one of its biggest crises in recent history. The coronavirus hit us by surprise, an unknown enemy that rendered the strongest of countries vulnerable and helpless. However, this is only a taste of what's to come. A bigger enemy is looming at large and this time, it is known. Climate change is threatening to end humanity as we know it, but we are still far from being prepared to fight it.

Climate change impacts all of us and the world's population has a central role to play in the fight against it. For this reason, it is important to inform people about the different issues related to climate change and to improve engagement on the battlefront. We wish to implement a platform that allows users to have a global understanding of issues relating to climate change. We want to display information in a way that makes navigation through the issues easy and interactive. The idea would be to give them the ability to view basic data stories about each of them expliciting their causes, their actors, their consequences and the solutions being brought up. One of the aims of this project is therefore to raise awareness in a different and novel way.

Throughout this project, we are going to implement a platform that allows easy navigation through the interconnected, underlying issues of climate change and easy access to interactive information about the different issues. This can essentially be represented by a knowledge graph.

Motivation

The inability to have a broad understanding of interconnected global issues hinders our collective ability to tackle them. This typically applies to climate change. One of our team members had this notable experience with their family. Indeed, their family did not know about the relationship between the food on our plates and global warming. This instigated a whole conversation about how to get information about climate change and the lack of broad view of the issue. This particular episode led to discussions between our team members and this spawned our project idea.

Target audience

This project is intended for any person who is curious to know more about the topic and for people who have difficulties linking the dots and grasping the whole picture. However, we must all be mobilized in the fight against climate change. Therefore, the main objective of the project is to make the visualization as simple and as attractive as possible to any kind of audience. It is then necessary to have a very user-friendly and appealing platform to attain our objective.

II. Dataset & Preprocessing

Dataset

The dataset we use in this project is extracted from Wikipedia. We started by scraping all the articles present under the category ***Global Warming*** and all its children categories. To limit the number of articles we scrape, we decided to limit the depth of the tree we are traversing to 12. We finally collected 25593 articles, scattered over 4680 different categories. This dataset was still too large, so we decided to pick a subset of categories and only keep the articles that fall in it. The subset contained the following categories : ***Methane***, ***Extinction Rebellion***, ***Fuel taxes***, ***Hydraulic fracturing***, ***Exxonmobil***, ***Gazprom***, ***Self-sustainability***, ***Industrial ecology***, ***Ecovillages***, ***Eco-towns***, ***Wildlife smuggling***, ***Urban forestry***, ***Biofuels***, ***Sustainable gardening***, ***Animal waste products***, ***Oil platform disasters***, ***Coal phase-out***, ***Climate change denial***, ***Building energy rating***, ***Active fire protection***, ***Industrial minerals***, ***Composting***, ***Reforestation***. This reduced the dataset to 937 articles.

Exploratory Data Analysis

After shrinking down the size of the original dataset, we went on to explore the subset we obtained. In order to build a meaningful knowledge graph, we had to find links between articles. Therefore, we looked at several aspects of a Wikipedia article:

- Its text
- The category it belongs to and the path from ***Global Warming*** to it
- The outgoing links to other Wikipedia articles
- Causal relationships

First of all, the text of a Wikipedia article encloses essential information about the topics it explores, independently of the category the article belongs to. This lead us to use *Latent Dirichlet Allocation (LDA)* on the text of the articles. Given a number of topics, LDA determines the characteristics of each topic and a probability distribution of those topics in each article. We then proceeded to compute a cosine similarity between all pairs of articles.

Second of all, the category an article belongs to and the path of the category in the tree yield interesting comparisons. We therefore extracted for each article the set of categories present in the path to it, and applied a Jaccard similarity between all pairs of those sets.

Third of all, the outgoing links to other Wikipedia articles are another metric that allows to distinguish topics in articles. We therefore apply a Jaccard similarity between all pairts of sets of outgoing links.

Last of all, causal relationships improve the understanding of the interconnectedness of the different issues. The causal relations are extracted semi-automatically from Wikipedia articles. We apply pattern matching using a dictionary of causal terms and verbs, and then we manually select the most relevant results.

III. Visualizations

Design

The visualization comes in two different components : a node-link graph representing issues and their relationships (Figure 1), and a panel used to display detailed information about an issue (Figure 2). Initially, we are faced with just the graph spread out across the page, but upon clicking a node its information page will pop up and cover about 75% of the screen.

The graph will be squeezed in the remaining upper 25% and take the following determined shape :

- Middle : the current node
- Left : the nodes representing the immediate causes of the issue
- Right : the nodes representing the immediate consequences and solutions of the issue
- Background, position uncontrolled : the rest of the nodes

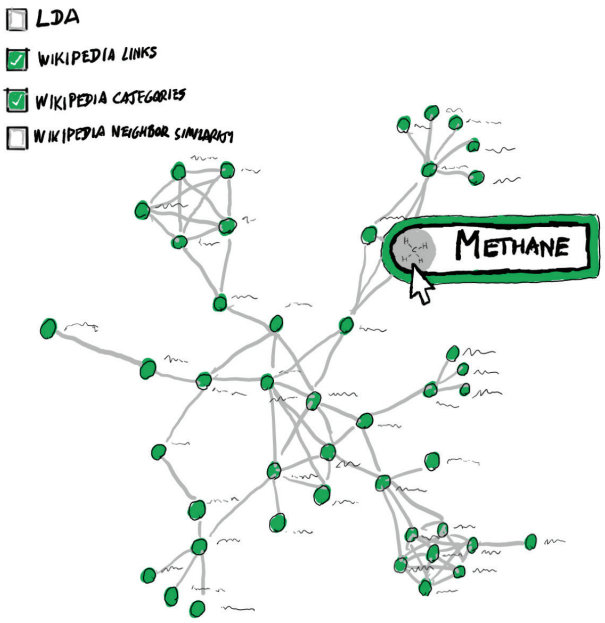


Figure 1

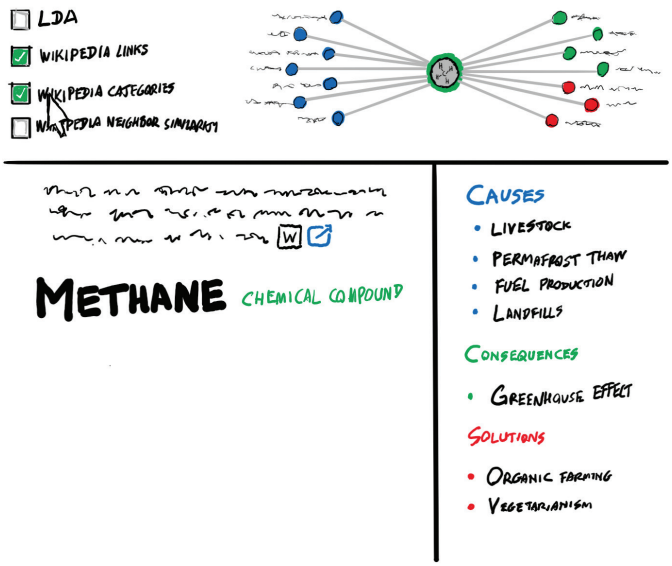


Figure 2

The user can navigate to causes, consequences and solutions by clicking them in the information page or directly on the graph, which will rearrange itself to display the chosen node in the same fashion. It is possible using checkboxes to toggle the different linkage methods listed higher (LDA and others).

As mentioned previously, the goal of the project is not only to allow easy navigation through different subjects related to climate change but also to allow easy understanding of those subjects. In order to implement that, we opted for a certain layout for key information corresponding to a particular issue. It would incorporate among others yet to be determined:

- A summary of the article

- A title
- Different descriptors including the link to the corresponding Wikipedia page, the category of the article, the dominant LDA topic and its associated keywords, and finally the thumbnail of the Wikipedia article (if available).
- Causes, consequences, solutions

Implementation

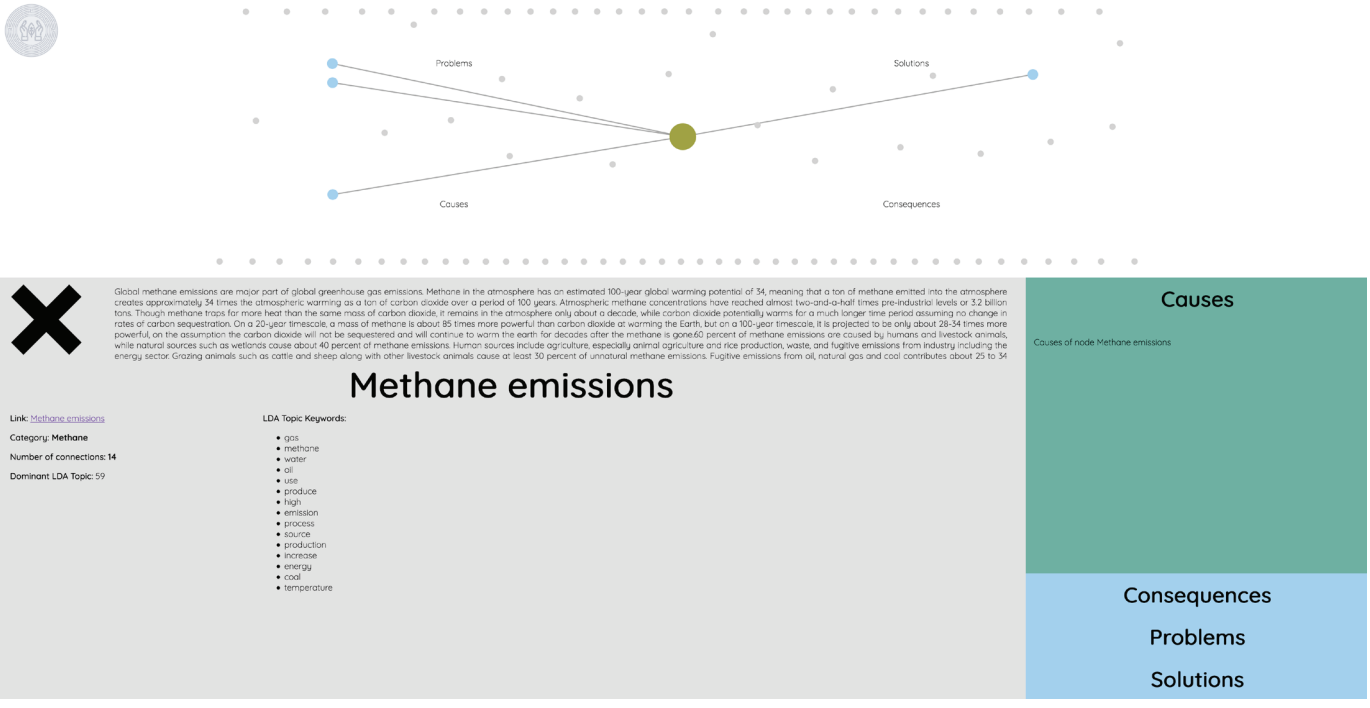
After Milestone 2, we started implementing the different components of our visualization. We used `D3.js` and in particular `D3.js`'s force layout graph to represent our data structure. Our system comprises three main components :

- The side menu
- The node graph
- The informative panel

Upon startup, the informative panel is absent and the graph nodes are grouped together. When the user hovers over a node, the name of the article corresponding to the node is show on the side and its lnks are highlighted. The user can click a node, and this rearranges the graph around the chosen node, showing it at the center with the related nodes (problems, solutions, causes and consequences) structured around. The rest of the nodes are spread over the canvas and can still be selected.

Clicking a node also brings up the informative panel which comprises the information related to this node, including the title, a short description, the Wikipedia link, the category of the article, the number of articles it is connected to, the dominant LDA topic, the LDA topic's keywords, a Wikipedia thumbnail (if available) and the related nodes. This panel retracts when clicking the cross on the side.

The side menu allows us to navigate to the about page and this very report, and choose the link weights to use for the visualization. The user can choose between 3 types of link weights: LDA topic similarity, category path similarity and finally outgoing links similarity. The screenshot below gives an idea of what the visualization would look like with full content in the bottom panel.



General Design Choices

Color Palette

One of the most important design choices we had to make was the color palette we were going to use in our visualization and website. We picked colours that were contrastive for the node distinctions and a teal for the presentation. Moreover, we wanted the colours to be color-blind friendly, so we went with the following colors :

Font

We used the `Montserrat` font on our website as well as the process book.

Title & Logo

We wanted to pick a name that would be unique and would have a memorable visual association. Several options were on the table, but we went with Climate Maze. This title resonates with the amount of available information about climate change and the need to disentangle it. Many paths can be taken to reach an information and this is

what the platform intends to facilitate. The icon in the middle of the maze is made by Pixel perfect from Flaticon.

Data used

The initial dataset we intended to use was too big as already mentioned. Therefore we shrunk it down to 937 articles. However, after some trials, we noticed that our

visualization did not scale well to that number. Consequently, we had to reduce the size of the dataset again to 100 articles. The visualization we get at the end is interesting enough to get the essence of the project. The final data can be found on the `data_viz_reduced.json` file. It includes information about the nodes, links and panel data.

Moreover, we kept a simulation with data made of placeholders. The We have built a placeholder dataset with a higher linkage (`data.json`) the likes of which we would expect from a more curated dataset in order to demonstrate the vision we have of a complete product.

IV. Challenges & Future Work

Data

The data we used throughout this project was scraped from Wikipedia and many preprocessing steps had to be applied in order to get meaningful insights. Our initial goal was ambitious and required more data wrangling than what we initially thought.

We ended up not using the causal relationships we extracted in each article because of time constraints and how basic the applied method was. Indeed, the extracted relations were very noisy and contained many false positives, and it was hard to manually filter out the noisy samples. Therefore, we had to leave the Causes, Consequences, Problems, Solutions side of the informative panel empty.

More complex methods are available to extract such causal relations, taking advantage of the part-of-speech tags and the power of artificial neural networks. As this project is intended to be a proof-of-concept that will be further worked on, we will explore such models and fill in the panel with more insights.

Furthermore, we reduced substantially the size of our dataset, that was already a subset of the entire set of articles on Global Warming and Climate Change. We will need to scrape many more articles to create more Cause-Consequence and Problem-Solution relations.

In addition to that, we have to make the visualization more scalable as we used a very reduced dataset, the size of which is only a small fraction of the data that should be used to fulfill the purpose of our visualization. We will therefore need to explore more visualization libraries, and adapt the backend to make a part of the computation offline.

Moreover, we will need to find a way to automate the process of linking articles together on the basis of Cause-Consequence and Problem-Solution. In this project, we picked a few nodes in order to show what it would look like on a larger scale, but this also entails the implementation of an algorithm that would determine those relationships as automatically as possible.

Design

Linking the different requirements of our visualization was not an easy task, and required some workarounds. The visualization did not work out well with the size of our dataset. Additionally, the size of the dataset will certainly grow in the near future, and we will need to make the visualiazation scale accordingly. We will therefore try different visualization tools from the ones we tested during this project.

Additionally, we will need to add more user-friendly links and make it as simple to understand as possible, because the targeted audience can be anyone interested in knowing more about climate change. This also means that we would need to rethink the way we display the current links we propose.

V. Peer Assesment

Yann Yasser Haddad

- Worked on the data preprocessing
- Took care of the data analysis throughout the project
- Worked on improving the visual identity of the website and the project
- Linked the graph visualization to the panel data
- Created the layouts of the process book
- Worked on the content of the process book
- Created the logo

David Resin

- Worked on the graph visualization
- Created the collapsable panels on the website
- Linked together the different components of the website
- Worked on improving the visual identity of the website and the project
- Workde on the content of the process book
- Recorded the screencast