DS4D Assignment 3

Link to website: https://s1879569.wixsite.com/encyclopaedia-group1
Link to video: https://youtu.be/8OCITifn-FA

Introduction & audience

Our dataset is the eight OCR'ed editions of the Encyclopaedia Britannica, which were released between 1768-1860. The Encyclopaedia is an aggregation of general knowledge and therefore is an interesting dataset as it showcases how knowledge and its categorisation evolved throughout time. Our audience is the general public, anyone who is interested in the Encyclopaedia but is overwhelmed by the amount of content available. We aim to provide an entry point for an initial overview of the content and structure of the Encyclopaedia, in an accessible and engaging way.

Data processing

Due to the data being unstructured text, we had to generate our own structured data from it. We decided to focus on very specific aspects of the data: simple entries of the form "TERM, definition", as well as references to topics of the form "See x", and used reference counts as a proxy for the popularity of a topic.

Topic popularity across editions is interesting to investigate as it showcases how knowledge evolved throughout time, with different topics being popular and more extensively researched at different points in time. Also, by showing example entries for a few topics, and the term with most references in its definition, we are able to give the public an idea of the language used in the Encyclopaedia and its structure and interconnectedness.

We generated dataframes for terms and their definitions, as well as reference counts in each definition, and all reference counts across editions. We performed some cleanup, such as removing very short or long terms and definitions, and removing non-topic references ("See fig", "See Vol").

As our data is the original, unclean OCR, it was impossible to create RegEx queries that wouldn't pick up noise, or leave some data out. However, this isn't much of an issue for us, as we can assume that the same kinds of omissions will be made across editions and terms, so the processed data will still be indicative of the proportional differences between terms across editions. Having noisy data did mean that extracting the longest and most reference-heavy term required manual inspection of the top contenders, as it couldn't easily be done computationally.

Visualisation

In terms of visualisations, we decided to build a website as a platform to show our investigation of the dataset. On the home page, there's a video showing the real physical books as a welcoming and when

you click into the data subpage, there's more to expect. First, there is a "What's in the Encyclopaedia?" slider and you can see the words and value amounts of each edition as well as the longest term and their references when you drag it. This is a quick and brief introduction to let people have a basic understanding of all the eight editions statistics data.

Scrolling down the page, we selected five commonly referenced topics, they are Anatomy, Architecture, Agriculture, Botany and Chemistry. We present the image that is related to these fields and if you move the cursor to those pictures, the number of the referenced topics in each edition pop up and you can see the changes by that. In the lower-left corner, there is also an image showing the popularity of these topics, this was adjusted after our data holder's suggestion that the concept of popularity and number of counts might be misleading to people, since they were initially in different pages.

Then, if you take a look at "LOOKING FOR MORE?" section and click the “more” button, there is a page showing more topics' popularity proportions. When you click the box of each topic, the value changes in each edition will show. In this case, we select the topics that have significant changes to intrigue people and raise their curiosity. Moreover, we also demonstrate the top 20 referenced topics across all editions on the same page to deliver a summary statistic.

Finally, you can find more detailed information of each edition if you scroll to the bottom of the page and click the "MORE SPECIFIC?" section, you can see the top 3 referenced topics in each edition, we use different colour to categorize each of them so people could understand the main changes more easily. By giving the information from the broad overview to a more specific description, we arrange the structure of the website in the same way, therefore, leading people to understand our intricate dataset step by step.

Game

Before designing the game, we know it was hard to show so much text information in such a micro-game, so we chose to filter some data that we want players to know, which should be the most general but little-known things. After considering, we chose the total number of references in each edition because that could indirectly reveal the richness of content in each edition.

Then why do we want to use a game? We also have asked the question to ourselves, and the answer is not just a game, but a 2D game. The game is of Super Mario type so that it can show data much clearer than 3D games, and meanwhile, it only has a single line just like time. Fortunately, Encyclopaedia Britannica was published by volume over time.

The game is a guided tour, also a timeline. When playing, you can find the publication of edition 1 was followed by the first industrial revolution and Scottish enlightenment. After meeting the Jump Graph level, the player can jump on each pad where there are secret bounce pads so that they can reach a fixed height. For showing the data, we give the character a colourful trail after him and we can get a bar chart after jumping. In addition, we chose the top five most referenced topics across all editions that we think most people may be not familiar with. You can jump from one pad to another, and explore in the

new world. One secret thing is, how much time it takes you to jump down from top shows the relative reference number. You can feel the data yourself.

Future work

Future work would involve extracting the long-form articles for topics (which have a more complex structure of parts & sections, as well as annotations of illustrated figures and formulae), and investigating the change in their lengths, to understand more deeply how the structure of the Encyclopaedia changes and complement our existing work: for example, is there a periodic “cleaning” of smaller entries, whereby they get absorbed into the longer-form article of a topic? Another improvement would be to contextualise topic popularity changes by relating them to historical events and discoveries.