DS4D - Assignment #3

Group 6: Anastasia Athanatou, Jingyi Chu, Yidan Yuan, Yongchang Zhu

Data holder: Sarah Ames

2 Minute Video: https://drive.google.com/drive/folders/1vMdyqz7mDUw3XwPYwd4wPjZE8psHOYIM?usp=sharing

Interactive Website Link: https://relic-dyebt.github.io/scottish-exam/

Introduction

For our project we built an interactive webpage that offers an insight into the Scottish School Exam papers, from 1888 to 1963, which have yet to be explored or presented to the public. Our design choices were fueled by our target audience: the general public; we wanted people of all ages and backgrounds to understand and appreciate our findings.

Initial Challenges

As the dataset was unexplored, our data holder gave us freedom to explore the dataset in any direction we wanted to. She suggested for us to begin by exploring areas such the key themes and topics, references of authors, representations of Scottish places, and gender dynamics within the exam papers.

Data Background

Description of data Our dataset came in two forms: plain text files and images of the exam papers. We primarily focused on using the OCR text files that were uncleaned and required preprocessing.

Preprocessing steps We used Python as the main programming language. With the help from Python libraries such as *Spacy* (for natural language processing), *jieba* (for word tokenization) and *re* (for regular expression matching), we implemented preprocessing modules to clean the text files, e.g., tokenization, lowercasing and punctuation removal. After data cleaning, we conducted basic syntactical/semantic/topical analysis, including word frequency calculation, POS tagging, named entity recognition, and dependency parsing. These analysis serves as the basis for further investigation into different aspects of the data. For instance, we use the result of dependency parsing to acquire sentential contexts for gender words, which are then associated with of timestamps of text files, to gain an understanding of the temporal dynamics of gender context.

We used *Pandas* and *Numpy* for data storage and manipulation. *Seaborn* and *Matplotlib* were used for the final visualization.

Four Perspectives

- **Gender**

The aim of this analysis was to identify potential gender bias in the exam papers and its patterns throughout the time. We conducted the analysis from three aspects: general gender analysis, temporal analysis and sentential context analysis. In general analysis, we manually collected two gender lexicons from the web, and calculated the frequency of male words V.S. female words in all exam papers. The result shows that there is a significant difference in the total number of male and female words, with a ratio of approximately 5:1. To visualise this information, we sketched a school clock (representing a pie chart) and used its dials to indicate a 17.4% ratio – the ratio of female words in the exam papers. In temporal analysis, we studied the dynamics of gender word frequency over time and observed a negative correlation between years and the ratios of male to female word frequency. Lastly, we acquire sentential context via dependency parsing, and identified gender bias clues by quantifying gender dependency (i.e., male->female and female->male).

- **Exam Topics**

When we extracted the number of subjects from the exam dateline and drew a plot of the number of subjects with respect to years, we discovered a continuous increase of subjects throughout the 75 years of exams. The increase of tested fields was slow in the first 40 years but around 1950 the rate rapidly increased.

We investigated this further by manually extracting and analyzing 4 single years of exam topics. In 1888, the exams had only 12 topics, in categories of language, math and science. By 1921 (33 years later), the topic number slightly increased to 16. Twenty-nine years later, the topic number doubled from 16 to 32 and two new areas appeared: music and liberal arts.

Besides this, we noticed that the science area was enrichened by more topics: zoology, chemistry, etc. This trend continued till 1936. The findings from these 4 years were presented in the website as a drawing of school bags and labelled books so as to fit our findings with the Exam papers' theme.

- **Location & Authors**

The most frequent writer mentioned in the exam papers is Shakespeare. His plays include many kings, so it inspired us to represent the data in a form of a king with a crown. We also calculated the frequency of the word "Scotland", to find out how Scotland was represented in the dataset. The frequency, which features the degree of localization, increased and peaked in 1953.

We also made an animated bubble map showing all the Scottish cities mentioned in the exam papers from 1888 to 1963. The sizes of the bubble denote the frequency of each city mentioned in a particular year. Most major cities in Scotland were referred in the exam papers, e.g., Edinburgh and Glasgow.

- **Circulars**

Compared to the Scottish exam papers, circulars are shorter and more readable. We found that some circulars were connected by a single theme, such as food (milk and vegetables), while others are a completely unique. We also found that there have been many debates throughout the years. One good example of this is extension of school age, and the policy was firstly proposed in 1919, but finally implemented from the 1st September 1939.

Another observation was that some policies has taken a long time from being proposed to being implemented, whilst some propositions were not implemented at all. Further analysis is needed to better understand circulars' propositions and the discussions that led to its complete rejection.

Design: Processes & Challenges

Process To communicate our findings, the group agreed on developing an interactive website which was manually developed with HTML, CSS and JavaScript, and deployed on GitHub, and the content of the website were sketched by paper and pencil.

During the brainstorming stage, our first challenge was deciding between a digitally drawn interactive comic or a pencil-sketched data visualisation. Since our target audience was the general public, we decided that a simple data visualisation would be more suitable over a comic, as it is often associated with and appreciated by a youthful audience.

We also found writing the code for the interactions challenging and needed to re-arrange our content to enhance the UX.

Future Improvements While it was part of our design concept, contrast between the pencil drawings and the white background highlighted emphasized any minor imperfections. Furthermore, several drawings were scaled up and down so the pencilwork was no longer coherent in its thickness and density. We would work on correcting this in the future by re-drawing several illustrations and editing the particles out with Adobe Illustrator.

Lastly, we want to continue working to improve the website alignments, as our current website has several minor UX issues - particularly in terms of aesthetics and intuition.