Computational Modelling in the Humanities and Social Sciences

mbkb74

1 Introduction

The task I have chosen is to model the features of castles in England by geographic location. There are many ways in which a castle can be built, such as if it has a moat, a portcullis and so on.

2 Sources of data and used modules

The main source of data will be the National Heritage List for England by Historic England [1]. Data will be obtained through this using Beautiful Soup, a web scraping tool for Python [2]. The data will then be processed by Stanza, a Natural Language Processing Python library [3]. Data is then visualized using plotly in order to make the patterns clearer to see [4].

3 Models and implementation

The analysis of text uses multiple of the features from Stanza, first in using Part of Speech Tagging in order to find nouns and plural nouns in the sentences, as these are most likely to be the features I am looking for.

I am also using the provided lemmas feature in order to just extract the singular versions of the nouns, so that if in some cases a feature is discussed in plural, and in some cases singular, they will be treated the same. This has both benefits and drawbacks, providing a benefit in increasing the matching, however the use of the plural may convey information about the features, such as if a castle has multiple towers, compared to one tower.

The dependencies provided are also used to match together compound nouns, such as "Curtain wall", as this provides more information than the separate nouns "curtain" and "wall".

In order to categorize the castles into sections, I am using NUTS1 regions of England [5], this is an EU specification which divides up countries, in this case, dividing England into 9 regions. There is the option to get more granular with NUTS2 and NUTS3 regions being smaller than NUTS1, however given the limited dataset of around 230 castles, I wanted to ensure there were many castles per region so that the trends would be visible.

The location of each castle is extracted from the webpage, using an Ordnance Survey Grid Reference. This is then turned into latitude and longitude coordinates, which is finally converted into a NUTS region. One area in which this caused some trouble is in castles on small islands, as the converter must have been using a map that excluded these for performance. However the only case in which I discovered this was Lindisfarne castle, so not a significant proportion of the data, and so shouldn't affect the trends seen.

In order to simplify the categorization process, first I ran the code to generate an ordered list of features by how many times they occur. This list was then filtered to just the top 50 that I wanted to observe. This was done to ensure they were significant findings, as if it was just from one castle, it doesn't imply any trend. For these 50 features, a dictionary was created under each NUTS1 region with each features, and as the program went through the castles, the number was incremented if the feature was found in that region. Also kept track of was the total number of castles in each region so that the number of features could be divided by the number of castles to get an average value, ensuring that regions that just had more castles didn't look like they had more of a certain feature.

4 Evaluation of models

On evaluating the heatmaps, features did come out as showing as more prominent from certain areas, they can be shown below. You will see that London is white on this map as none of the castles in the dataset were in the London area.

The simplest evaluation of these results is to look at the description of building materials. Due to the difficulty of transportation at the time castles were being built, they use materials from the local area, and so they reflect geological surveys.

The most obvious of these are flint and sandstone, as shown below

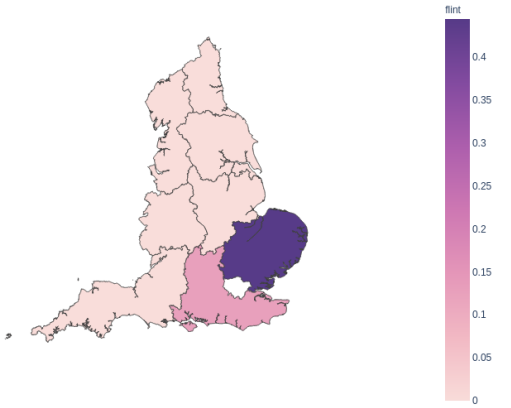

Figure 1: Flint

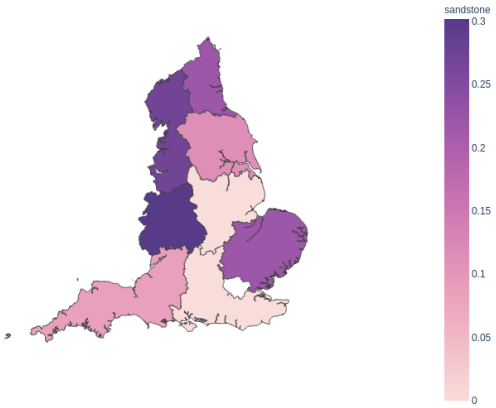

Figure 2: Sandstone

Data from the British Geological Survey shows a large flint deposit in the south east area where flint is featured so heavily [6]. Sandstone covers the west of England, curving round to the North East, with the rest featuring limestone instead. The image to view this needs to be seen at a large scale, and so it provided as an appendix at the end of this paper.

In addition to these two natural resources, the derivatives of natural resources can also be seen, with the South West using a lot more brick than other areas.

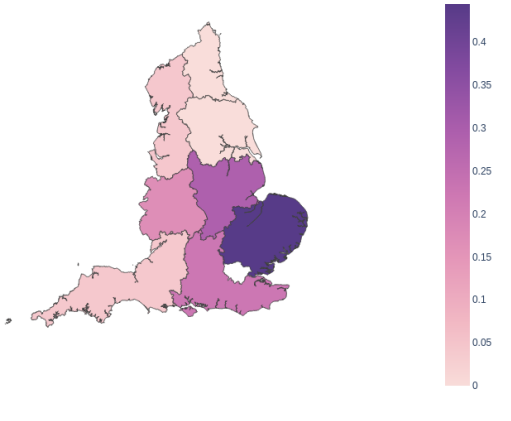

Figure 3: Brick

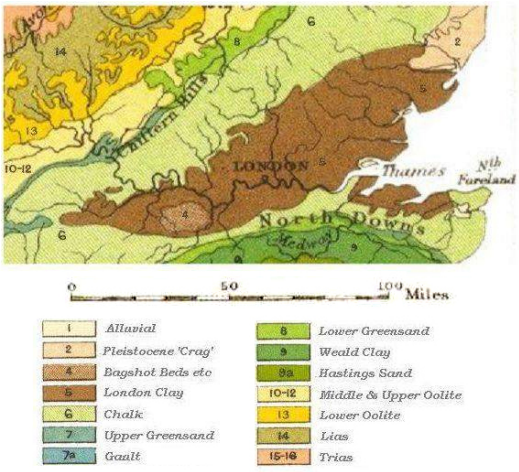

Figure 4: London Clay

This is likely due to the large amount of Clay in the Thames valley, named the London Clay Formation [7]. This can be seen in Figure 4.

5 Conclusion

References

[1] Historic England. *National Heritage List for England*. 2021. URL: https://historicengland.org.uk/listing/the-list (visited on 04/19/2021).

[2] Leonard Richardson. *Beautiful Soup*. 2004. URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (visited on 04/19/2021).

[3] Peng Qi et al. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[4] Plotly Technologies Inc. *Collaborative data science*. 2015. URL: https://plot.ly.

[5] Eurostat. *Background - NUTS - Nomenclature of territorial units for statistics - Eurostat*. 2021. URL: https://ec.europa.eu/eurostat/web/nuts/background.

[6] British Geological Survey. *Geology of Britain viewer*. 2021. URL: https://mapapps.bgs.ac.uk/geologyofbritain/home.html.

[7] MG Sumbler. *London and the Thames Valley*. Vol. 13. British Regional Geology S., 1996.