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].

3 Implementation of models

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”.

4 Evaluation of models

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`.