Online tool for linguistic and sociolinguistic studies accessing open online resources (based on "Survey of Medieval Winchester" name and occupational material).

Scientific Advisor : Prof. Olga Khotskina

Rishabh Tiwari

Novosibirsk State university

December 29, 2020

Introduction

- This study shows development of social categorisation based on the social positions and relationships.
- It is considered as combination of linguistics and sociology.
- It takes language samples from sets of random population subjects and looks at variable that include such things as pronunciation, word choice, and colloquialisms.
- It takes language samples from sets of random populationsubjects and looks at variable that include such things aspronunciation, word choice, and colloquialisms

Aim Of Research

- To provide an online tool incorporating data from the Index of property holders from the Survey of Medieval Winchester and linking this data to related open online resources for specialists conducting their research in the field of linguistics and sociolinguistics.
- To test Python and NLP on digitized Winchester material and its ability to extract chunks of data on names and occupations.
- The format and output should be modified to guarantee more varied resource helping scientists and researchers to combine various resources in one too.

Data

- The source material for this thesis is drawn from the Survey of Medieval Winchester.
- It is the logical continuation of the volume Winchester in the Early Middle Ages.
- It includes various manuscripts of legal, religious, personal and other origins.
- It is the reviews of the legal structure, property distribution, town organization, town development, population size and development, the trading system.
- I selected a large sample from the whole material of the Survey of Medieval Winchester collection and described development of English society based on the first name material

Architecture

- Data Extraction From medieval Urban Data (Winchester material)to convert all data into text format from images and pdf.
- Data Pre-processing(Data Cleaning) Word Tokenization, Stemming , Lemmatization ,Part of speech tagging (POS),Named entity recognition and Chunking.
- Data Labelling using supervised learning to extract names and occupational information from Winchester material.
- Implement model Using word2vec and BERT to gather all the useful information from the data .

Data Assemble

- To read the data from multiple sources, PDF Documents Scanned Documents.
- I used optical Character Recognition Technique.
- I used Python for analysing the data but the data need not be in the required format always. In such cases, we convert that format (like PDF or JPG etc.) to the text format, in order to analyze the data in better way.

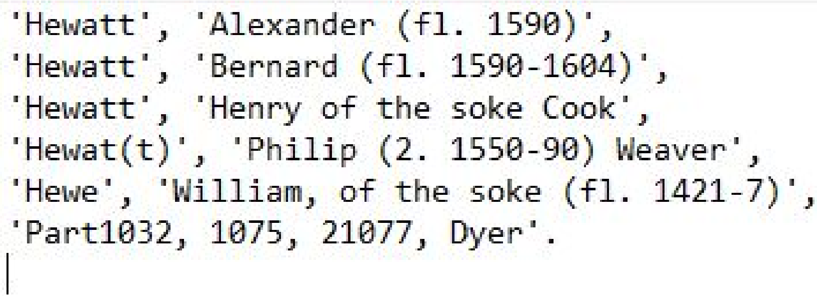
Continued

```
me Adderly, Edward (ff. 1604-29). From his ?relative
Richard Adderl(e)y: 614-15, 640-1, 643. From his
?relative Stephen Adderly(e): 183.

Adderly, John, v. John oo

Nicholas (fl. 1590). :
cee ha ee oa Richard, - Richard Adderl(e)y.
See 27/30-2. Married (by Adderly, Stephen, v. Stephen Adderly(e).

: Adderlye, John, v. John Adderl(e)y.

Adderly(e), Stephen (fl. 1590-1604). 4, 183, 184, 185,
589-90. Possibly related to John Adderl(e)y and Ed-
ward Adderly.
|
Haberdasher.
~~, family of (fl. early 17th cent.) 184.
Adrian, Willi . 1502-3). ?Part 905.

le Ac(h)atour, John (cont.)
, Inhabited 622-3. Described as master by 1320,

citizen 1325.
```

Data Pre Processing

- Regular Expressions
- Data Pre-processing(Data Cleaning) Word Tokenization, Stemming , Lemmatization ,Part of speech tagging (POS),Named entity recognition and Chunking.

Continued



```
'Hewatt', 'Alexander (fl. 1590)',
'Hewatt', 'Bernard (fl. 1590-1604)',
'Hewatt', 'Henry of the soke Cook',
'Hewat(t)', 'Philip (2. 1550-90) Weaver',
'Hewe', 'William, of the soke (fl. 1421-7)',
'Part1032, 1075, 21077, Dyer'.
```

Futiure Work

- Data Labelling using supervised learning to extract names and occupational information from Winchester material.
- Implement model Using word2vec and BERT to gather all the useful information from the data .

THANK YOU