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

RISHABH TIWARI

Scientific Advisor - Olga V. Khotskina

Novosibirsk State university

19 May 2021

Introduction

- My approach is to provide analysis of winchester data using image recognition and natural language processing techniques to provide a search tool for linguistic and social studies.
- To provide an information retreival model based on winchester data.

Data

- The Survey of Medieval Winchester, a typical example of a secondary source that was compiled for the City of Winchester on the basis of very rich and varied material.
- Each individual entry includes the following information: family name, given name, dates, properties, occupation (the first the most important one), relatives, other properties, and law cases.

Data Extraction

- To extract data from multiple sources PDF and JPEG scanned documents.
- Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and "read" the text embedded in images.

```
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 Prepossessing

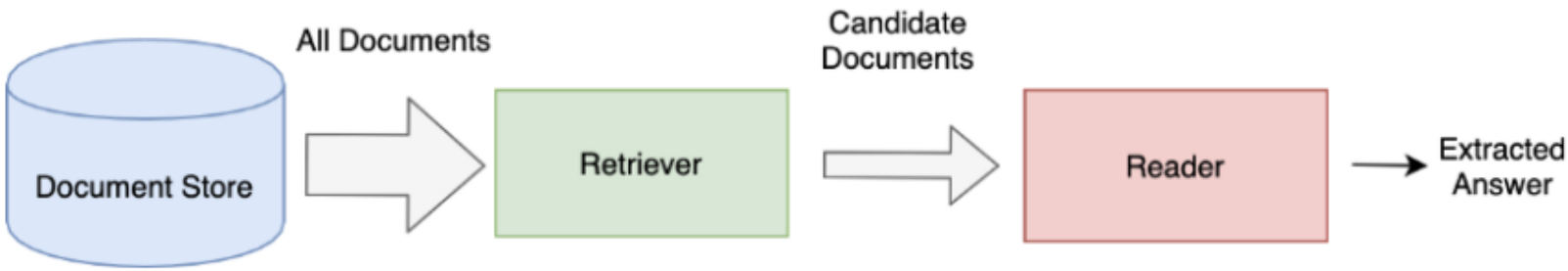
- Tokenization.
- Removing Stopwords.
- Named Entity Recognition from retraining Spacy(NER) model for custom entity(names,surnames,occupation and Year) recognition on own custom labeled data.

Abbat William (A 1464-71) Abbott 372 Inhabited 372 Baker entered liberty city 1469-70 (CA) Subsidy 8d (1464) Abbelyn Agnes (ft. 1440-54 Year) From relative John Name Ablyn Surname 1005 Abbod Surname Roger (A. 1309) Part 734 Abbodeston William Name (fl 1350-91) Abboteston de Abbotestone .. 50-6G H 61 62 305 Married (1366) Eve Name Butcher Occupation 1350-86 Year inspector carcasses Occupation 1364 (CR) Assessor fines Occupation 1371 (CR) citizen 1374 Eve wife (fl 1366 1391) de Abbodeston Surname Richard (fl 1327-32 Year) 994-5 part 1100 Possibly related Henry Abbotestone Inhabited 994-5 Draper Subsidy 4s (1327) 6s (1332) Abbot John (fl 1582 Year) Part 351 Abraham Surname Pinch Name (fl 1230 d. 1236) 281-8D Jew Accused theft florins 1230 (Justl/775) Roger Alis Thomas de Bromden Reginald de Moyun owed money 1230-1 (Cal Close R 1227- 31 pp 428 539 550) Hugh de Godeshull owed money 1235 Year (Cal Close R 1234-7 p. 81) debtors included Adam de Rutherefeud Bartholomew de Ellested William de Boleyn Easton (Exc Fine R 8 234 Cal Plea REJ 196) Hanged felony 1236 (allegedly child-murder) (Cal Close R 1234-7 pp 239 271 341 Just1/775) Abraham (son) Cokerel Name (fl 1290) Year Part 281-8B C. Jew Abraham son Elias Name (fl c. 1222-73) Part 27/30-2 Jew Claimed debts Wilton 1244-5 Year John Name de Pratellis Surname owed money Occupation 1248 Year John Name Holebury Surname owed money 1253 Year (Cal Plea REJ 97 Cal Pat R 1247-50 pp 20 119) Butcher Name Abboteston Surname William Name v. William Name Abbodeston Surname de Abboteston Surname Joan v. widow Henry Name Abbotestone Name AAbraham John Name (fl 1335) From Year father Abraham de V Abbotestone Henry Name (fl 1340 d. 1347) Abodeston From relative Richard Name de Abbodeston Surname 994-5 Abraham Robert Name (fl 1335 1362) From V Married (1347 Year) Joan Name de Abboteston Surname Father Isabel Name * Abraham Surname Reynham Surname part 66 Brother Robert Name Abraham Surname Inhabited 994-5 Subsidy 6s (1340) Isabel Name daughter (fl 1347 1357) From father 994-5 Joan Name de Abboteston Surname widow (fl 1347 Year) From X husband 994-5 Mother Isabel Name Inhabited 994-5 father Abraham de Reynham Surname part 66 Brother John Name Abraham Surname A Abraham Surname Thomas (fl 1350) 1002 V Abyndone John v. John Abyn (g) done Abyn (g) done John (fl 1408-17 Year) Ablyndon 195 501 Clerk

Architecture of Information Retrieval

- A corpus of data.
- Transformers library to build information retrieval model.
- Haystack library to scale information retrieval model to thousands of documents and build a search engine.
- I used tow Bert model trained on squad dataset deepset roberta-base-squad2 and distilbert-base-uncased-distilled-squad.

Architecture



Results

```
print_answers(prediction, details="minimal")
```

```
[   {   'answer': 'Abbot, John',
        'context': '1364 (CR). Assessor of fines 1371 (CR); citizen by 1374. '
                   'Eve wife of (fl. by 1366; 1391). de Abbodeston, Richard '
                   '(fl. 1327-32). ?994-5, part 1100. Possibly related to '
                   'Henry Abbotestone. Inhabited ?994-5. ?Draper. Subsidy 4s. '
                   '(1327), 6s. (1332). Abbot, John (fl. 1582). Part 351. '
                   'Abraham Pinch (fl. 1230; d. 1236). ?281-8D. Jew. Accused '
                   'of theft of florins 1230 (Justl/775); Roger Alis, Thomas '
                   'de Bromden, and Reginald de Moyun owed him money 1230-1 '
                   '(Cal Close R 1227- 31, pp. 428, 539, 550); Hugh de '},
    {   'answer': 'John de Arundel',
        'context': 'bailiff of the Arundel(1), Peter (fl. 1407-18). 23, 726. '
                   'soke 1405-22 (PRBW); gate-keeper of Wolvesey Palace 1406, '
                   'esquire 1408-17, citizen 1411. ~~, Agnes widow of, v. '
                   'Agnes Bouere widow of Walter Arw, Philip, of Amesbury (fl. '
                   'by 1314). tive John de Arundel: 114. Married Margery widow '
                   'of Philip de Candev(e)re. ~~, Margery wife of, v. widow of '
                   'Philip Arundell, Elias (fl. 1590). Part 645. Married (by '
                   '1411) Elizabeth. Brother of William Archer. Merchant 1402; '
                   'vintner 1407, butcher 1412 (CR); sold '}]
```

Continued

- In this research as per topic we provided a framework of information retrieval system based on custom data and different natural language techniques to extract data from image and preprocess it for machine learning models.
- It shows DistilBERT provide more accurately answer as compared to RoBERTa in information retrieval.
- In this work we providea a custom based named entity recognition for custom entity recognition for computational sociolinguistics doing research in this field to extract enities.
- We can use these approach to get information from other old archieves which are still not in digital form and accessible by researchers.

Scientific Novelty

- This work give approach of nlp and machine learning techniques to work in computational linguistics.
- Customised NER model for name ,surname and occupation material.
- Researchers and students working in this field can use this type of information retrieval to extract information large archives data.
- We can use these approach to get information from other old archieves which are still not in digital form and accessible by researchers.

Future Work

- To get more structured and specific results.
- To add more data available from other sources to make more dataset and good results.
- To make API for this information system for easy user interface.

Publication

- My abstract is submited and approved in V International Scientific and Practical conference "Digital Transformations in Education".

Thank You

THANK YOU