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

Scientific Advisor - Olga V. Khotskina

- 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.

- 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.

1. BULL-101	1. NO. 10/10/100	1011.01	Co. Allowing
a ladare Meloci	and, Dest J. 181-21, her in black		100
instant 40-1 (burdled a more to 100).	Robard Address World, Md-1, MD, Tran To-	And in case of the local division in the loc	
	Induite Taples (alleges) (H)	A NAME AND ADDRESS OF TAXABLE	
And the state of the second seco			14. K.B.
a man has I the to Find hand in	anders, Segmen, - Names Annual II.	Name and Address of the Owner of Street, or other owner, the owner of the Owner of Street, or other owner,	
Conception of the local division of the loca			
a new property line of the line house of			proph. The rate of a loss of party sharing that is 10.1
		which we will be taken.	
\$ more, finding of the light of the states of	Resolution Smith of A such (The sum) (20)	Aprile, Manar pt concepts, spatia, from the basics William Applica, Sell, St. Ster. on	
Annual States, a later & inglish etc.			
	In case of the local division of the local d		
Startin will	the same of a local division of the same	1143 (0.10) (0.10) or (0.10) to (0.10) (0.10) (0.40) (0.10) (0.10) (0.10)	
		The last of the Lot, C. Anna is the	Think takes
		Real Spite Spaller die Heller Ander Soller Steer auf Spite Stadie staat is Solar Spite	Index of the lot of the lot of the
	Support of the little	Control of States in part (178), Aug.	
Address of the local division of the local d	graphent () THE OF ME	"". She bight of \$ 1000, fea to take 100	
Index in a holive per FLA factor by			Dis. March Vol., and Diving.
	have of Wardson's 1973.	** Next table table do Net o faile with all (1996-10), Franks balance (1995-10), 60.	
salar per Plat Same in 198 taux hits	Tagen, Jone, a John Tagentin.		
the support of the local data			Name in column 1 distants in formation
	And the Post of the local division in the	Apples, Real + New Apple Chapter size State	And in case of a line of the
on First Association in the second state.	spatia di farta di	Table without Without Applicate Applies, Advant as Applicate Applicate	
Adus, the suffice, or Adus in Water will write a Will-			and their leafer of it will be be been
	And a line of Anna 128-18, 428, 428, 428,		Married Annual of Street of Street and Street and
Manuel, 1880 (2) 1212-102, 1882/001, 188, 188, 149	per glind, tong Kin, Silk, Silk, Sulla business or of spaces, the cold, time implement of business	Non-Robert of Street, 51 (197), 881	
40.	Sparts, Same in 1981, the support of private same		
Man, Salari (i. 177). An Di, Fuelto etter, a Toron Mare,			super little. And any start fact any proper little.
	Ansated 167 in 1972 and 168 in 1988.		
		Read of Concession, or Specific of Specific	
Address of			Non. Soluti (1 143). 6 1847, for SK Speed
			the 1400, June Law marked by 1407-1800, and have the 1440, June Personal stated a William
	Manuel all, day loss \$10, Names in 1975	the billion of the local	
sciences are 3 (100-80, 100-90, 98, 80.4	Base & Portage.	- then have of figures of its local on , the part	- Set
			"", has all of it hills and still.
ph. 204, 1864, 854, 854, 864, 86, 700 start a fatast tabelity as praity of the	And in Contra of the lot of the line of th	With Party 2 (Street, Area, but 10)	
		Contractor of the local division of the loca	
Annalog, Robust (J. 177-1877), Proc. in Column Inter Address, Miller, Mill, and Address			And American Street Street
	Appendix State (1997), for 187, feasible object of	THE NAME	Report Ann.
Speciales, Andreas, A. Sadaras, Sadardicito.	Supervision of the OCL DWD Advant. No.	Manual Rad Store addresses loss 100-5. Manual Dr. (1998), & M. (1999), & Collin, No.	a cristing loop a line change
			NAME AND POST OF A POST OF ADDRESS OF
	with frees		
4	17 m		

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 Edward 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 ( tt. 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 50-8G 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 sumame Richard ( fl 1327-32 Year ) 994-5 part 1100				
Possibly related Henry Abbotestone Inhabited 994-5 Draper Subsidy 4s (1327) 6s (1332) Abbot John (11 1582 Year ) Part 351 Abraham Sumame				
Pinch Name (fl 1230 d. 1236 ) 281-8D Jew Accused theft florins 1230 ( Justi/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 281-88 C. Jew Abraham son Ellas Name (fl c.				
1222-73 ) Part 27/30-2 Jew Claimed debts Wilton 1244-5 Year John Name de Pratellis sumame owed money occupation 1248 Year John				
Name Holebury sumame owed money 1253 Year (Cal Plea REJ 97 Cal Pat R 1247-50 pp 20 119) Butcher Name Abboteston sumame				
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 sumame Reynham sumame part 66 Brother Robert Name Abraham sumame Inhabited 994-5 Subsidy 6s (				
1340) Isabel Name daughter (fl 1347 1357) From father 994-5 Joan Name de Abboteston Sumame widow (fl 1347 Year ) From X husband				
994-5 Mother Isabel Name Inhabited 994-5 father Abraham de Reynham sumame part 66 Brother John Name Abraham Sumame A Abraham				

## 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.





- 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 provide a custom based named entity recognition for custom entity recognition for computational sociolinguistics doing research in this field to extract entities.
- We can use these approach to get information from other old archieves which are still not in digital form and accessible by researchers.

- 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.

- 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.

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

## THANK YOU

Scientific Advisor - Olga V. Khotskina