Social and Linguistic Analysis of medieval Urban Data Scientific Advisor : Prof. Olga Khotskina

Rishabh Tiwari

Novosibirsk State university

28 may, 2020

Social and Linguistic Analysis of medieval Urban Data

Introduction

- This study shows development of social categorisation based on the social positions and relationships.
- To measure socio economic indices such as education, income, wealth, occupation, ethnic heritage, and family dynamics to better understand the relationship between language and society.
- ▶ What can forenames tell us about English medieval society.
- The work demonstrates the next step in the development of anthroponymic research.

- 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.
- This field of study deals with anthropological linguistics, dialectology, discourse analysis, ethnography of speaking, geolinguistics, language contact study, secular linguistics, the social psychology of language, and the sociology of language.

- 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

- 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

<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	BOLLATED.		BILLON D	The second
	and the local	salah, Shari J. 189-20, Day in Salar	MR. box MI. bosts strart a March	14
	over all it hashed a more in 1995.		sector free spinsts	The second second second
		Indulte Tagina Address 1981	and a second sec	
	new with of \$1 to 1000 \$ to 1000	tamen, bits a pitte substitute	And in case of the local division of the	
	and the said field of a Norger's length of	and the property of	of West Splitter 5, Trails, state a Line	at Making 48, Correspondingly hole (102) 172-
	LAND OF THE PARTY OF			
<text></text>				P. Name & Content for Content of Content
<text><text><text><text></text></text></text></text>		assailute, Station (J. 118-100), 4, 18, 79, 78,		
	Anna & Col.7 Soc 176-Rolley of	Mild, Audity wheel a job vehicley and he	Applies. Want is Want highlighter also then	
<text></text>	to be here.		Arthur, Some of company, or other	Martin, John & John Marajahn.
<text><text><text><text></text></text></text></text>			Table With Autors 10 Million of	 Among A and a start of the star
<text></text>		Anna Miller J. 192-5, Par 988		
		same finite of the state in case of the same		
	and the second s	the local division of		
	and it is been and it is to see \$			
		Stational States and		
	a day of the second in the second			
		distance of a party hardware		
				states of the state of the local states
	and Plant Survey in 1983, June Arthur	Appen, John, A. John, Applica.		
		Appen, John, A. John: Appelline.	Min. Control part 63, 10, 18,	
<text><text><text><text><text><text></text></text></text></text></text></text>	and of property lines.			
<text><text><text><text></text></text></text></text>				
	a caller, a links in Want of a Viter of State	Applies 40. for tell 40.	repetitor, Rottore, a Manhaer Apalitan	and the local division of the local division
		a spatial later (C. 1986 August 1977 Your Low	Spiller, Killer, a William Spiller,	
	to send more 2 cities rate and		sting Street, of Street, Stree	April 1, 1996, of Strate (2, 1996, 14, Tax 80, 88,
	the h local second and an in-			1987. Pagetto ettere to Thoma taka.
	Robert H. 1978, Saw 20, Faulty street		all federate from of Branchings (8 to cloth, by China	and the part of the later of the later party being but,
An and a set of the form of		of Ladds, Shork, ed Sile.	 An other materials list, spec fights of the 	
	These P. Holl, Francis Topper Science		Statement of the state of the state	days, and () to (see, 40).
 The state of the state of the		The same barrier of it is 10.7. Non Isl		Corean .
And and a second				server, through (1) 1910; A 1987), fast BBL Bartinel
A manufacture of the second s		the last bagter of 2 to 101, 100 million	and the state of the local first from the local	
 Bernard Marker, Samper Marker, Samper	a familier (f. 1976). Mild. Postic read		I wanted and the lateral	
Chi Gao Chang, San Yuang, San				THE REAL PROPERTY AND ADDRESS OF THE PROPERTY ADDRESS
the state of the				be all all the distances
an Adard (2 ref. 2 ref. and 2 ref. 2	at a fished bilitity of posite a low	And	Water, Person 2, 1911-17, Aug. Soc. No. 101	
 a partier 14.0 (A) 4.0 (A) and (the same part on an family of the part		Addition your Bill
(R) HAL (In them cardwords) given, Mark - Ker (Jann), (R) HAL (In the transmission) (R) HAL (In	statute where and and the lots			
Andreas Andrea	OR MAN Half Printle stand to Street			Res Wilson of Long States
ALC: NOT A REAL PROPERTY AND A REAL PROPERTY A		Apple, Mile J. (199-1), for M. Analis - case of	10 type	Report New York, New
ALC: NOT A REAL PROPERTY AND A REAL PROPERTY A	and the second se	successive in the latter water, but	have to part in a stream ing 1814.	A 1974 Date - Date stress
E grade days	And a second second	1100	and the second s	NAME AND POST OFFICE ADDRESS OF TAXABLE PARTY.
Equility Print		Date:		
the for The		A.14		
for Cone	-	- 17 -		
	T+	Cont.		

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 Pre Processing

- Regular Expressions
- Common Python Regex Methods are re.findall(),re.split,re.sub,re.search
- Word Tokenization
- Tokenization is the process of converting a text or corpus into tokens.
- Stemming.
- Lemmatization.
- Part of speech tagging (POS).
- ► Named entity recognition.
- Chunking

Questions

- How many males from the general number had name John in the 14th century?
- How many of these Johns had wives and how many of them were Maries?
- What was the most popular male name among merchants in the 15th century? In 1401-1449 or in 1450-1499?
- Were there any Johns having the surname Smith in the 14th century?
- How many Johns in the 15th century had sons bearing the same first name?
- Which crafts were practiced by women in 12th century? 13th century?
- ► Etc

- Master Thesis in social and Linguistic analysis by Prof. Olga Khotskina.
- https://www.academia.edu/42696597/Ancient text recognition review.
- https://www.academia.edu/42984049/An approach to solve the problem of creating an automatic knowledge control system based on text analysis.

THANK YOU

Social and Linguistic Analysis of medieval Urban Data