Wikipedia Editor Demographics

Farrukh Tanveer

[1] RWTH Aachen University, Germany

Abstract: Sociodemographic involves the quantifiable characteristics of a human population. Mostly these characteristics are gender, age, location and education etc. Wikipedia is a free encyclopedia with the concept of openly editable. Since it is openly editable anyone can edit the articles, therefore it is necessary to confirm the demographical characteristics of the editor as the source of information not trust worthy or credible. Different approaches can be used to determine different sociodemographic characteristics of editors of Wikipedia. This paper discusses three methods. First method is based on key words that indicate the demographical information about the editor. Second method identifies the gender of the editor based on the name of the editor whereas third method is based on predicting the gender of the editor on basis of the provided text.

Keywords: NLTK(Natural Language Tool Kit), NLP(Natural language Processing).

1 Introduction

Wikipedia is a multilingual, web-based, free encyclopedia which uses the concept of openly editable and viewable content. It has the largest collection of articles among all the encyclopedia websites and is one of the most popular websites on the world wide web. It is biggest freely available source of information. If a person searches any general information on internet, then most probably the first link that a search engine provides is of Wikipedia. As Wikipedia is openly editable, anyone can write new articles or can edit previously written articles therefore no one knows from where the editor or the writer is getting the information. Whether the person who is editing the article have knowledge of that domain or not, what is the knowledge domain of that person or what is the level of education of that person. A person with just a computer and Wikipedia account may gather some related information about the topic and can write an article therefore, the source of information is not credible. This makes the source of information doubtful therefore even being the largest source of information on internet one cannot cite Wikipedia as a source of information in any research paper or scientific writings. Therefore, it becomes important to gather information about the editor to determine the trustworthiness of the information provided by the editor. Demographics means the study of dynamics and dimensions of in a certain context. Sociodemographic means the study of quantifiable characteristics of human population such as gender, age, education, interests etc. Sociodemographic have great importance as they tell about the information or characteristics of a person. In context of Wikipedia it gives us the information about the editor or the author of an article. These demographics help in determining different facts about the author. The problem faced in editor demographics is that some editors provide complete information about them in a very organized way in the summary box that tells summarizes each and everything about the editor whereas there are editors who don’t provide the information in the summary box and instead they write a whole paragraph about them. Then there are some mysterious editors who do not like to provide any information about themselves. In order to process and extract the demographical information NLP (Natural language Processing) in combination with different methods is used. NLP is a sub-field of artificial intelligence that focuses on enabling computers to understand and process human languages, to get computers closer to a human-level understanding of language. For the processing the natural language a library named as NLTK(Natural Language Tool Kit) is used. The main purpose of NLTK is to tokenize the paragraphs into sentences or words and to assign the parts of speeches to the words. This paper provides various methods to determine different demographical features of the Wikipedia editors. 1st method named as direct method provides all the important features if they are provided either in the summary box or in the biography text that user has written about himself. This method uses a bag of words containing all important words that are used in biography to describe various characteristics such as age, gender, education, location and interests. Those words are searched in the text and all the sentences with those words are provided as an output. 2nd method is the name-based gender prediction method, this method creates a list of proper nouns using NLTK and compares proper nouns against a predefined, gender categorized list of (male, female, mostly used for male and mostly used for female) names to determine the gender of the editor. When user has just provided information about himself without mentioning any name or gender then 3rd method is used to predict the gender of the person based upon the text that the person has provided.

2 Collection of Data

For the collection of data, a web base crawler was built in python that crawled the Wikipedia articles for their editors. Only English articles with English language were crawled. Crawled data was stored in a MongoDB database. The Wikipedia page ids along with the list of the editors of the pages were stored in one collection of MongoDB and the raw text of editor profile along editor id was stored in a separate collection of MongoDB.

3 Methods

3.1 Direct Method

Direct method is the simplest method. This method can be used when the editor has provided information about him either in the summary box or in the biography text that the editor has provided. In this method the raw text from the editors collection in MongoDB is fetched. A bag of words containing all the words that could determine or describe the editor is created. The table below shows the words used to collect the demographical features of the editor from the raw text.

Words	Description
This User	The information provided by the user in the summary box contains this phrase
Birthday	To determine the birthday of the user if mentioned in the text
Birth day	User can also mention it in this form although a typing mistake
Born	Another word used to determine the birth date of user
Country	The country in summary box indicates the country of the user
Lives	Used for the location of the user
Live	Used for location of user
Location	Mostly used by Wikipedia in the summary box to describe the location of editor
Current location	Mostly used in summary box to describe the current location of user
male	Describes the gender of the user
female	Describes the gender of user
age	Describes the age of the user
education	Mostly used in the summary box to describe all the educational content about the user

bachelor	Describes the educational level of the person
master	Describes the educational level of the editor
B.Sc	Describes the educational level
M.Sc	Describes the educational level
Diploma	Describes the educational level

Table 1: showing the words used to extract the demographical features

The paragraph of raw text fetched from the editor’s profile is tokenized into sentences using NLTK and then each sentence is searched for the following words in Table 1. If the tokenized sentence contains any word from these words then the sentence is displayed in the output as this sentence provides information about a demographical feature of the editor. Following flow diagram in Figure 1 shows the important processing steps to extract the demographical features. The flow diagram shows how the direct method works and extracts the necessary information provided by the editor.

The advantage of this method is that no prediction of any demographical feature is involved. Every feature extracted is mention by the editor himself which makes the information accurate and authentic. Using this method, a summary about the user can also be generated. It can also serve as an algorithm to filter out user related information for further processing. The disadvantage of this method is that if editor has not provided the information then it would not be extracted or predicted from the text. It totally depends on that the editor should have mention it by himself means if we want to detect the gender of editor and editor has not mentioned it in his profile neither in the summary box nor in the biography text then this algorithm cannot tell the gender of the editor. To counter this problem the second method is introduced to determine the gender from the text.

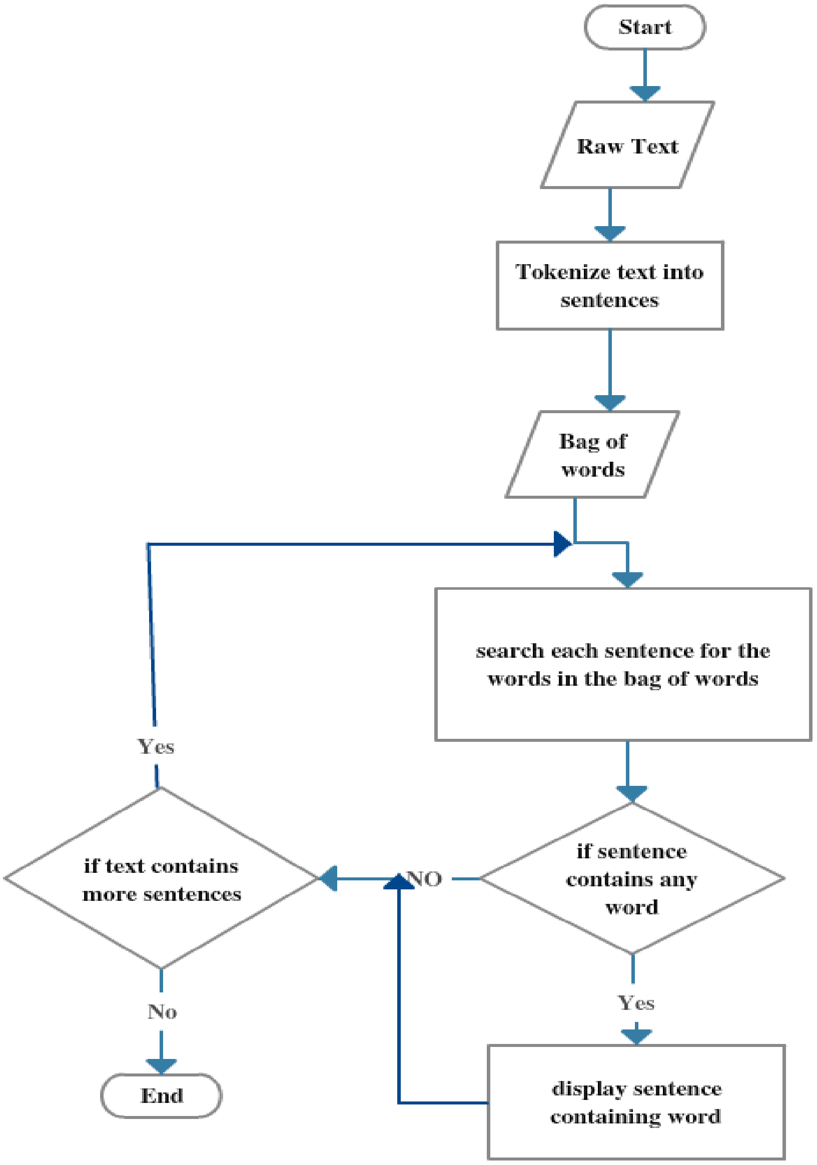
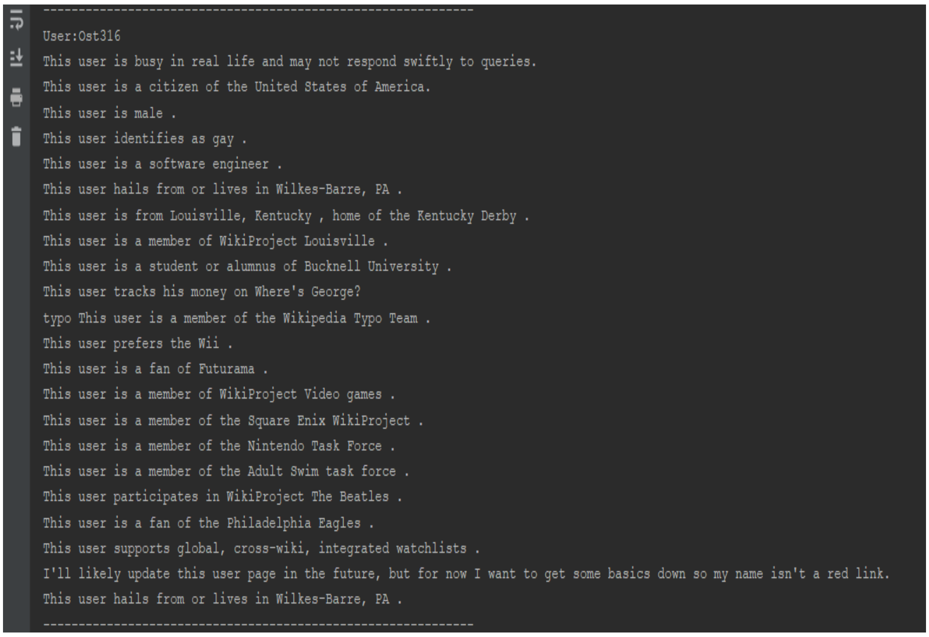

Start
Raw Text
Tokenize text into sentences
Bag of words
search each sentence for the words in the bag of words
if sentence contains any word
Yes
display sentence containing word
NO
if text contains more sentences
Yes
No
End

Figure 1: Flow diagram for direct method

Example output obtained from direct method is shown below



```
-------------------------------------------------------------
User:Ost316
This user is busy in real life and may not respond swiftly to queries.
This user is a citizen of the United States of America.
This user is male .
This user identifies as gay .
This user is a software engineer .
This user hails from or lives in Wilkes-Barre, PA .
This user is from Louisville, Kentucky , home of the Kentucky Derby .
This user is a member of WikiProject Louisville .
This user is a student or alumnus of Bucknell University .
This user tracks his money on Where's George?
typo This user is a member of the Wikipedia Typo Team .
This user prefers the Wii .
This user is a fan of Futurama .
This user is a member of WikiProject Video games .
This user is a member of the Square Enix WikiProject .
This user is a member of the Nintendo Task Force .
This user is a member of the Adult Swim task force .
This user participates in WikiProject The Beatles .
This user is a fan of the Philadelphia Eagles .
This user supports global, cross-wiki, integrated watchlists .
I'll likely update this user page in the future, but for now I want to get some basics down so my name isn't a red link.
This user hails from or lives in Wilkes-Barre, PA .
-------------------------------------------------------------
```

Figure 2: Detailed output from direct method

3.2 Gender Prediction

author profiling is of growing importance in a variety of areas. Mostly used in marketing, security and forensics. In marketing, for example, companies are interested in knowing the gender-based interests of their target group in order to achieve a better market segmentation. In Wikipedia the interest could be that what type of articles are written by female editors and what type of articles are written by male editors. What topics interest the male writers and what topics interest the female editors. We can know which gender like or dislike which topics. And, using this information, they will have a sort of first feedback from their clients and, so, they can adapt their strategy concerning the articles. This paper provides two techniques for predicting the gender of the editor. The first technique is basic technique that uses a categorized list of names to be compared against the proper nouns that occur in the raw text. If the name exists in the list then the gender of the editor is displayed. The second technique describes for gender prediction is based on the probability measure of different features that identify that the editor is male or female. Different features from the text are measured and probability of being male or female is calculated. For this purpose the bayes theorem is used.

Name Based Gender Prediction

This method is used when the editor has not mentioned his gender either in the biography that he has written about himself or in the summary box that Wikipedia displays about the editor. It is based on the technique of identifying parts of speech from the text. For this purpose NLTK is used. To achieve this, the raw text is broken down into words and each word is assigned a part of speech. This is done by using NLTK. NLTK tokenizes the whole raw text into words and a list of the words is created. Then in order to assign the parts of speech to each word in the created list NLTK parts of speech tagger is used and each word in the list is given a part of speech depeing on wether the word is a noun, pronoun, proper noun, adjective, adverb etc [1].

NLTK uses following abbreviations for tagging some common parts of speech.

Abbreviation	Full Form
CC	Conjunction
IN	Preposition
JJ	Adjective
NN	Noun
NNS	Plural noun
NNP	Proper noun
NNPS	Proper noun plural
PPR	Personal pronoun
RB	Adverb
VB	Verb
VBD	Verb past tense

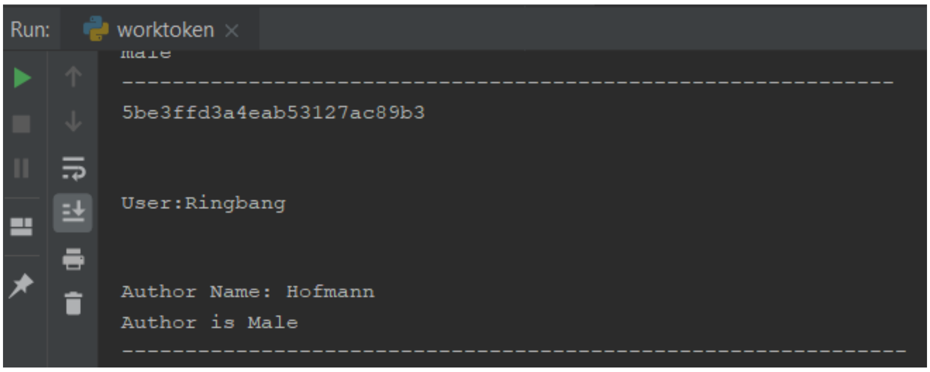
Table 2: Shows the partsof speech tags used by NLTK

The tagged list of words is then filtered for Proper nouns indicated by NNP in NLTK. It is filtered just for proper nouns not for plural proper nouns as one cannot write his name in plural. Proper nouns are the nouns that describe the name of the noun and is always written with first capitalized letter. Whenever a person writes his/her name they always use the first letter as capital. The list is filtered just for proper nouns to reduce the search space and search complexity of the algorithm. This list of proper nouns is then compared with the categorized list of names. The list of names is categorized as male, female, mostly male and mostly female. The categorization mostly male indicates that the name is mostly used for males and mostly female indicates that the name is mostly used for females. The categorized file of name is a CSV file. Which uses the following mapping for the indication of the categories.

Gender	**Indication in CSV file**
Male	M
Female	F
Mostly male	?M
Mostly female	?F
Unknown	?

Table 3: Gender indication in CSV file [2]

The name list used in this paper is taken from US security reports taken from the module genderizer [2]. If the proper noun from the text matches the name in the CSV file then the next column to the name column is check for the categorization of the name. The categorization is indicated by the following keys indicated in Table 3. Then the name and gender of the person is displayed in the output. The flow diagram in Figure 4 shows the working of the name based gender detection method, It shows all the necessary steps followed in this method. An Output from this method is shown below



```
male
---------------------------------------------------------------
5be3ffd3a4eab53127ac89b3


User:Ringbang


Author Name: Hofmann
Author is Male
----------------------------------------------------------------
```

Figure 3: Output from name based gender prediction method

The advantage of this method is that if the editor has just mentioned his name in the summary box or either in the text and that name exists in the categorized list of name s then the gender of the editor can be determined but it solely depends on the fact that the person should have mentioned his name in the description. The disadvantage of this method is that it only works when the editor has mentioned his/her name in the description. Another drawback could be that even if the person has mentioned his/her name in the text and NLTK does not tag it as a proper noun then it would not be compared to the file of categorized name which in return would not detect the gender of the editor. One more negative point could be even if the editor has mentioned his name in the text it is also tagged as proper noun but the name is not mentioned in the list so the gender of the editor would not be detected. The main disadvantage of this method is that if the editor has mentioned any other name in the description text then this method will also detect that name as a proper noun and if that name matches in the categorized list of names then it would also be assigned a gender. Since one editor cannot have two names so this is the major disadvantage of this method.

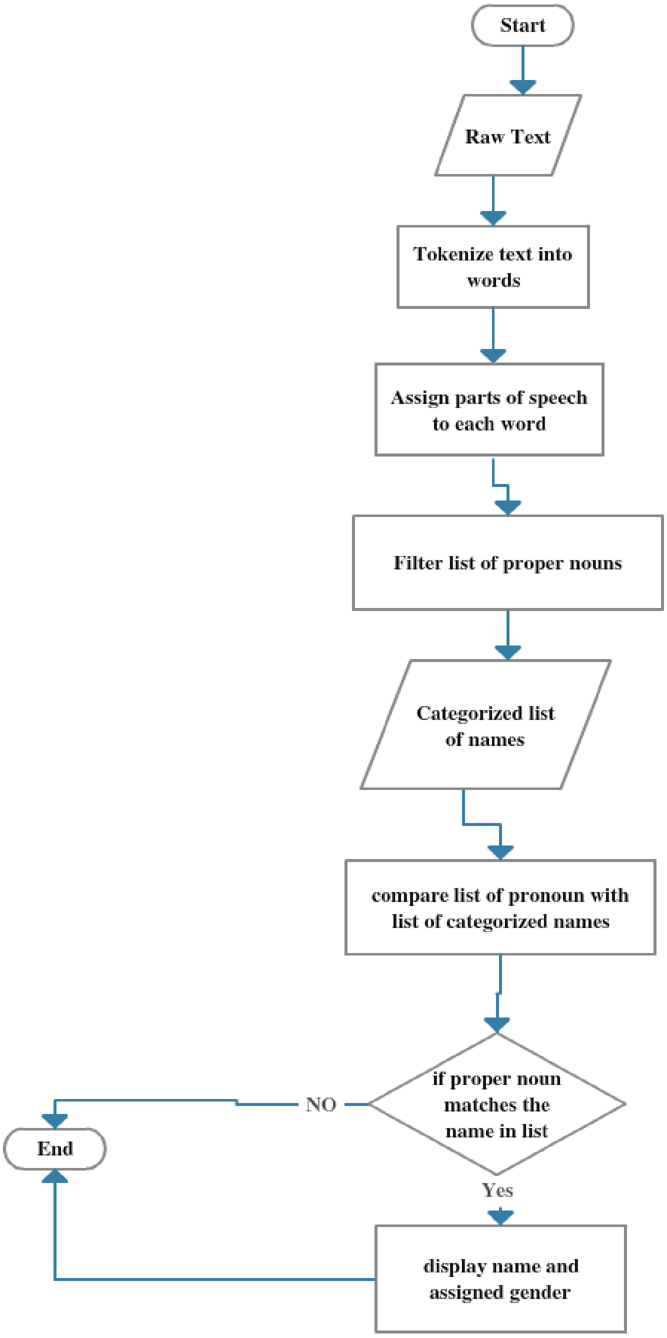

Start
Raw Text
Tokenize text into words
Assign parts of speech to each word
Filter list of proper nouns
Categorized list of names
compare list of pronoun with list of categorized names
if proper noun matches the name in list
NO
End
Yes
display name and assigned gender

Figure 4: Working of name-based gender method

Text-Based Gender Prediction

This method is used when editor has not mentioned his name or gender neither in the biography text nor in the summary box. In this method the gender of the editor is determined on the basis of the text that he/she has provided in the profile. The raw text is given as an input. In this method the Genderizer library is used. This module uses the Naïve Bayes Classifier which is based on Bayes theorem.

Bayes Classifier

The Naive Bayes classifier is a basic classifier [2]. It uses Bayes Theorem to predict the probability that a given name set belongs to a particular gender, P(c | x), from P(c), P(x), and P(x | c) [2]. The original formula for Naïve bayes theorm is :

$$P(c \mid x) = P(c) * P(x \mid c)/P(x).$$

It calculates the probability of occurrence of an event based on likelihood and class prior probability and predictor prior probability [2].

Genderizer

The genderizer modules first requires the data against which the classifier is trained. First it generates a model based on the provided data. In this case the data grabbed from the online profiles of editor is given against which it is trained. Once the machine is trained against the data. Then it takes raw text as an input and based on certain factors classifies the editor as male or female [3]. The Figure 5 shows the flow diagram showing how the genderizer module works. The advantage of this method is that even if the editor has not provided any information about the gender then it can be predicted using this module. The main disadvantage of this method is that prediction is not always accurate. Most of the editor profiles are the bots created by editors on Wikipedia even if the editor is a bot it also classifies it as a male or female. An output using genderizer module is shown below

Input	I was born in 1998, and have studied social science at Nacka gymnasium in Sweden.
Predicted Gender	Female
User name of editor	Josve05a

Table 4: Represents an output of text based gender prediction method

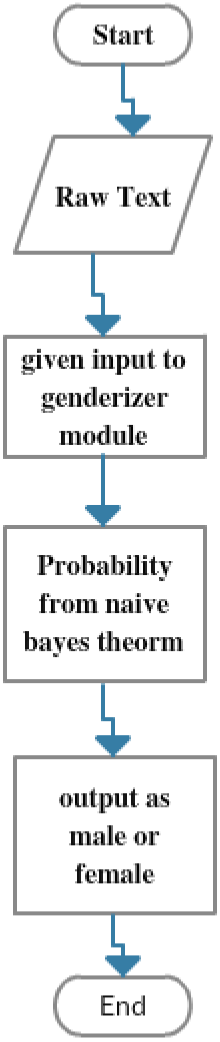

Start
Raw Text
given input to genderizer module
Probability from naive bayes theorm
output as male or female
End

Figure 5: Flow diagram for genderizer module

Discussion and Findings

This paper discusses three methods for determining different sociodemographic characteristics of Wikipedia editors. First method is the simplest method based on a bag of key words that indicate a demographic character of the editor. It is the most accurate method as all the information is provided by the editor himself. It can also serve as a filtered data for making conclusions about other demographic characters of the editor. It can also serve a raw data for prediction of other characteristics that may be based on user interests but the main disadvantage is that it totally depends on the editor whether or not he/she has provided the information or not. 2nd method predicts the gender of the editor on the basis of name mentioned in the biography text. The main advantage of this method is that if name is mentioned then gender can be predicted. The main disadvantage of this method is that if during comparison two names are mentioned by the editor then it assigns gender to both names. The 3rd method is the text based prediction method. The accuracy of this method is least among all methods mentioned in this paper as it predicts just on the characteristics of text. This method works really poor for the editor bots made by different editors because it also categorizes the bot as male or female.

Conclusion

There is no perfect method to predict the sociodemographic characteristics of Wikipedia editors. Every method describes in the paper has its own advantages and disadvantages. In my approach, I tried to formulate three methods for determining different sociodemographic characteristics. The first method is used to extract all the known characteristics if they are provided by author in his profile. Second and third methods are used to predict the gender of the editor.

For future research a new method can be established by combining the direct method and text-based prediction method. First method can serve to extract all the interests of the editor and a machine learning approach can be used to train the data depending on the male and female interest and then the gender of the editor can be predicted.

References

[1] "Natural Language Tool Kit," [Online]. Available: https://www.nltk.org/.

[2] "Naive Bayesian," [Online]. Available: https://www.saedsayad.com/naive_bayesian.htm.

[3] muatik, "github," [Online]. Available: https://github.com/muatik/genderizer.

[4] J. Perkins, Python Text Processing with NLTK 2.0, book. Packt Publishing, 9 november 2010.

[5] H. Zhao and F. Kamareddine, "Advance gender prediction tool of first names and its use in analyzing gender disparity in computer science in the UK, Malaysia and china".