HackerCamp Summer 2018

Division - Analytics

Sparsh Dutta [1]

[1]Department of Electronics and Communication
The LNM Institute Of Information Technology

March 2, 2018

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Problem Statement

- Variation in names leads to difficulty in identifying a unique person and hence de-duplication of records is an unsolved challenge.
- The problem becomes more complicated in cases where data is coming from multiple sources.
- The problem to solve is to find a simple solution to find all the variations in the name and replace that with a single name.

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Tools Used

- Python 3.5.4 and Jupyter Notebook
- OS Module
  - For Operating System related operations.
- Pandas Module
  - For Data Management and Representation.
- Dedupe Module
  - For Machine Learning related operations.

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Input

	A	B	C	D
1	ln	dob	gn	fn
2	Frometa	24/11/34	F	Vladimir
3	Frometa Garo	24/11/34	F	Vladimir Antonio
4	Frometa Garo	24/11/34	F	Vladimir A
5	Frometa	24/11/34	F	Vladimir
6	Frometa G	24/11/34	F	Vladimir
7	Frometa	24/11/34	F	Vladimir A
8	Frometa G	24/11/34	F	Vladimir A
9	Dutta	24/11/34	M	Sparsh
10	Dutta K	24/11/34	M	Sparsh
11				
12				
13				

Figure: Input Dataframe

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Output

	A	B	C	D
1	ln	dob	gn	fn
2	Frometa	24/11/34	F	Vladimir
3	Dutta	24/11/34	M	Sparsh
4				
5				
6				
7				
8				
9				

Figure: Output Dataframe

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Approach-1

- In Approach-1, I tried to create word embedding of the names using google pre-trained model and was successful in that.
- Then after that I tried to find the similarity between the duplicate names and other names in the dataset, distance metric used was euclidean distance.
- I was successful in finding the duplicate names but their were anomaly in the result and the process was taking time to finish. So, I started thinking for a new approach.

Content

1. Introduction
   - Problem Statement
   - Tools Used
2. Inputs and Outputs
   - Inputs
   - Outputs
3. Approach Used
   - Approach-1
   - Approach-2

Approach-2

- In Approach-2, I found a python module named "dedupe" which solves all the problems specified above.
- The "dedupe" module uses the same approach as specified in approach-1 but on top of that it does one thing very special that it creates cluster of similar string and it does it very fast which makes it scalable for heavy loads.
- This "dedupe" module is easy to implement and you need to first train it according to your need and it will easily do the job specified.