Assignment 1: Part 2

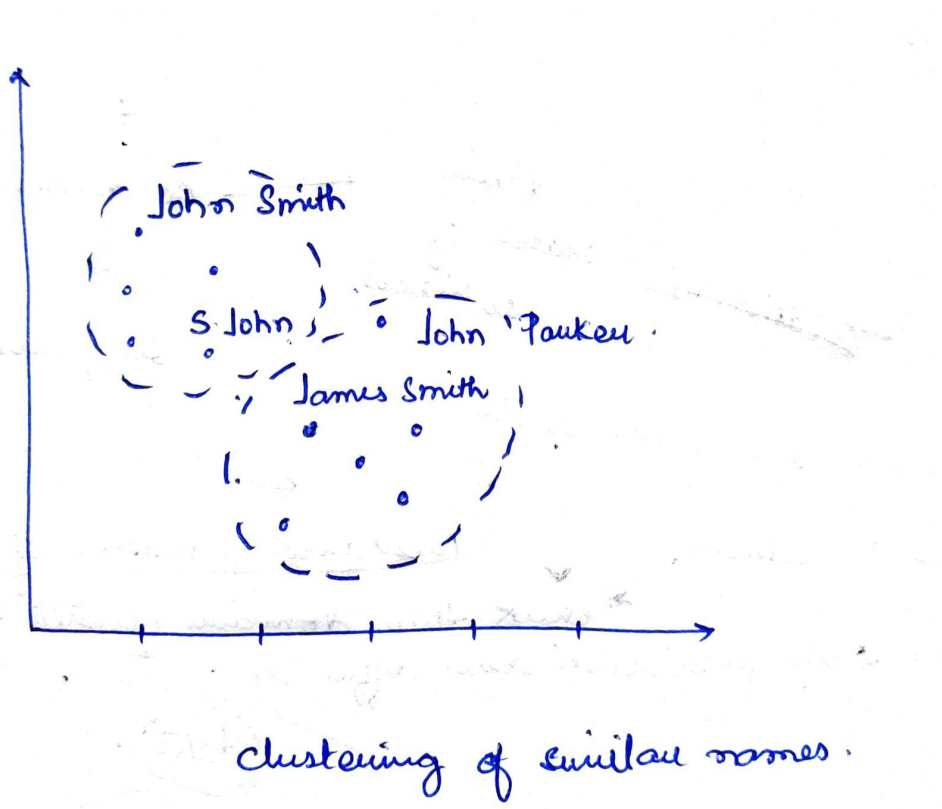
Advanced Programming in Python(DSE 309)

Author Name Disambiguation

Name: Ajay Choudhury **Roll no.:** 18018 **Date:** 20th Sep 2021

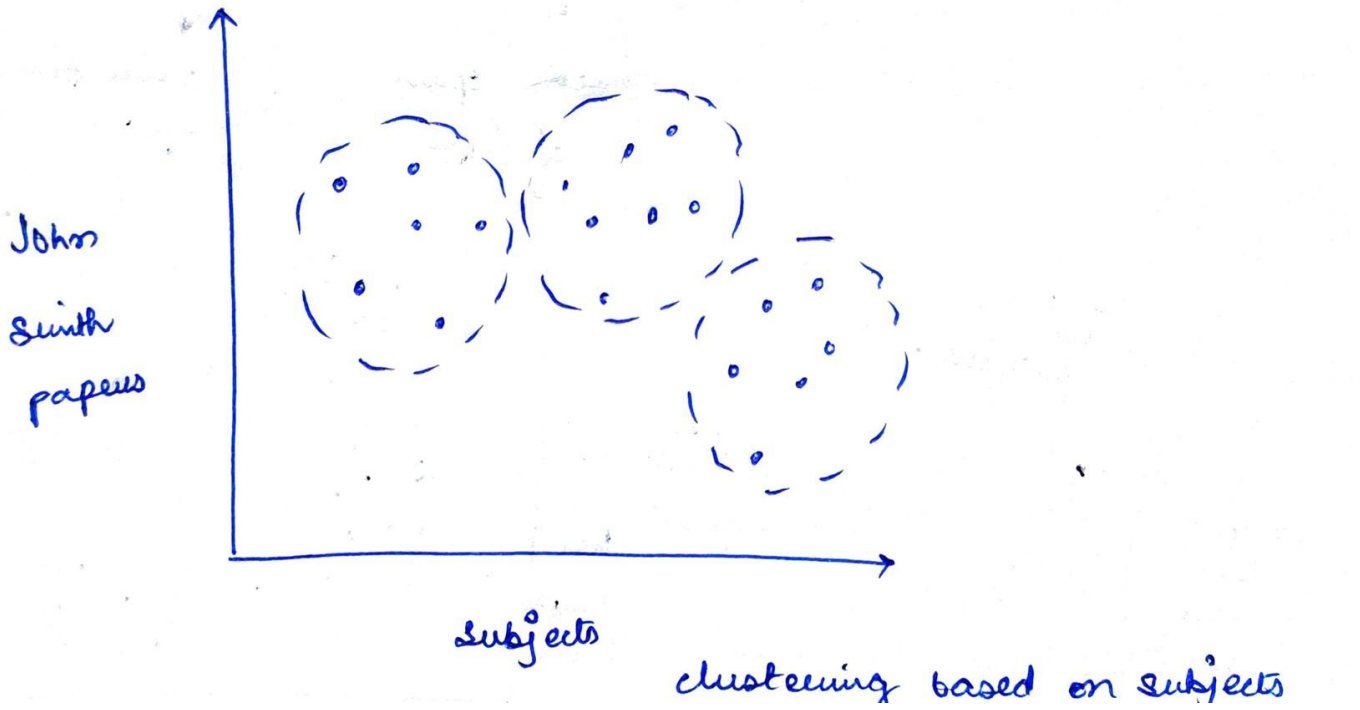
A short summary of how we can proceed with author name disambiguation. First of all, a dataset of various authors, their published material, its topic and relevant subject, date of publishing, and co-authors are required. We can obtain these data from sources like Kaggle or Github.

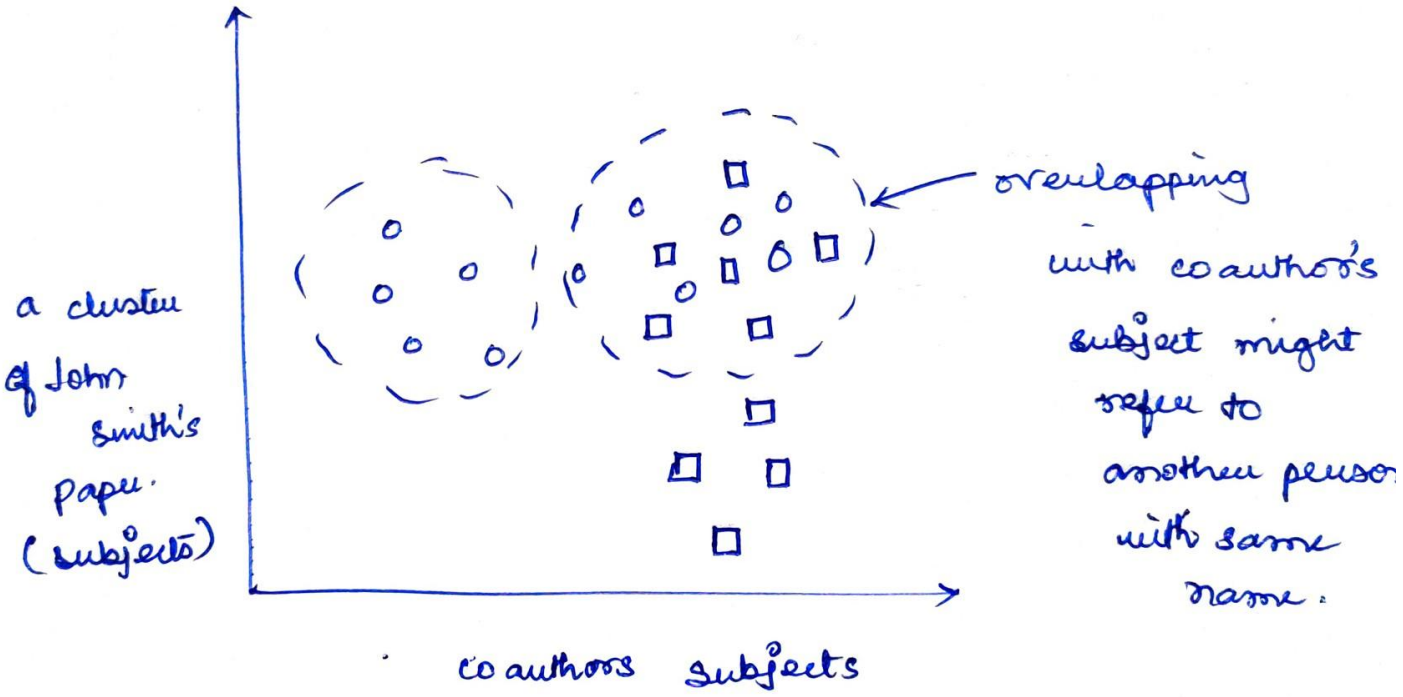
Now, we can create a pipeline to move ahead with the task. As we detected the authors' names, we can now prepare a clustering model to collect all the names that spell similar or are abbreviated like John Smith, J. Smith, Smith J, J Smith, Johnny Smith, etc.



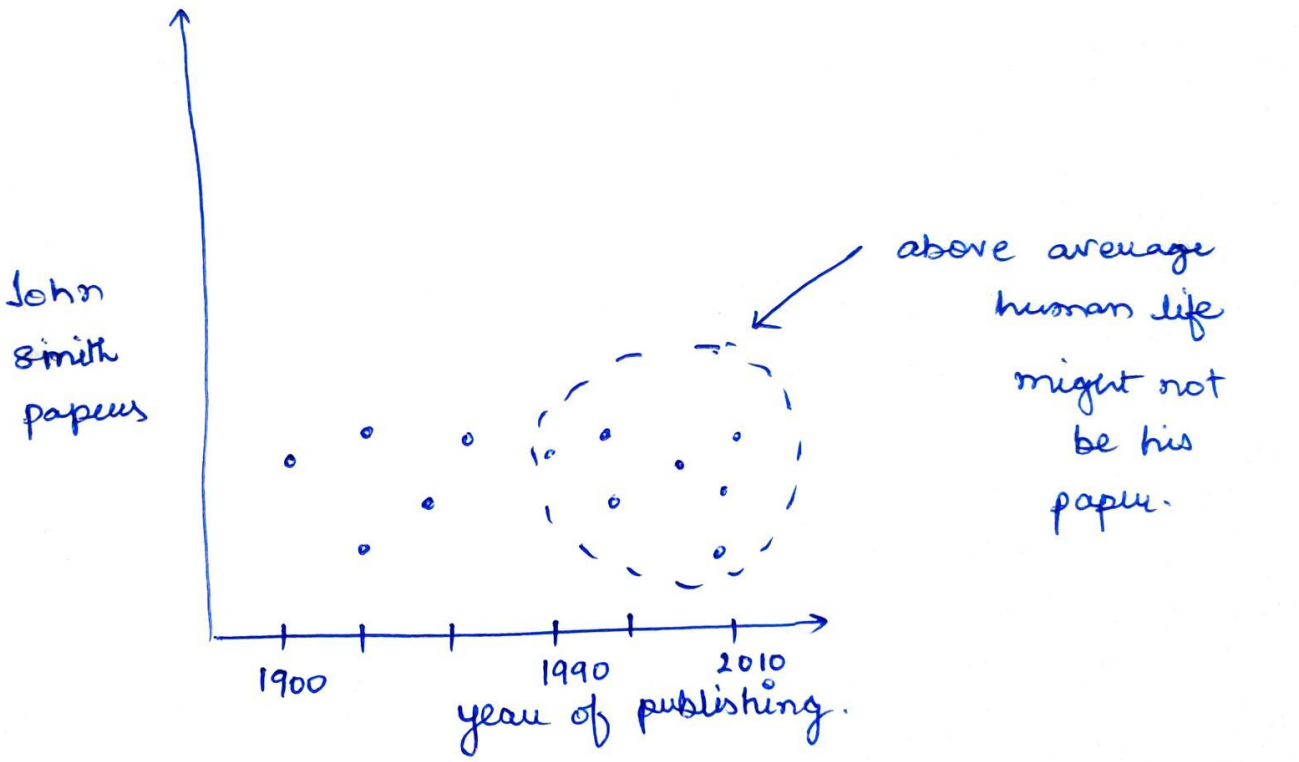
After getting all the names of the author we can cluster the similar names accordingly and check how much the names are related to each other, this will give us a naive cluster of different authors but still a cluster can contain more than one author with the same name.

Now, we can analyse the subjects of each author and their co-authors(if other papers of them are available in the dataset) and can correlate with them if they match or not, again clustering the data can differentiate between two or more authors with the same name but very different domains of interest.





Dates can also play a vital role in determining if two papers of almost related subjects are written by the same person or not if the difference between their dates of publishing is very close or greater than an average human lifespan.



After we train our multilabel classifier model on the above data and clusters, we can try our model on the test dataset to first differentiate between authors with the same or similar names and then assign their papers under their names.

We can also take the institute name (or location and origin) of authors and co-authors for tuning our model as well, but I think that won't be very helpful all the time (for most of the recent datasets), it might work well for older datasets.

Resources I am following and studying:

1. **Github:** diging/author-disambiguation
2. **Github:** Makobyte/Author_Name_Disambiguation