Author Name Dissambiguation using Unsupervised Learning Techniques

Dinesh Adhithya

September 17, 2021

Introduction

Author name disambiguation is a type of disambiguation and record linkage applied to the names of individual people. In this work I try to use Agglomerative clustering techniques to cluster authors based on the following set of features describing each paper published by an author: Name , Organization , Journal of published paper , co-author names and title of paper.

Data set Description

The data set Wang et al. (2011) is used for studying name disambiguation in digital library. It contains 110 author names and their disambiguation results (ground truth). Each author name corresponds to a raw file in the "raw-data" folder and an answer file (ground truth) in the "Answer" folder. The data can be downloaded from this link.

Literature Review

The paper Louppe et al. (2015) uses following pipeline for this task :
(i) a linkage function determining whether two publications have been written by the same author; and
(ii) a clustering algorithm producing clusters of publications assumed to be written by the same author.

We shall use a similar approach here as well.

Methods

Data Preprocessing

The data set has been cleaned using various functions implemented in nltk **?** package and the steps for cleaning and preprocessing the data set is listed below:

- All features listed above were extracted from XML files and stored in arrays.
- The Titles of published papers were cleaned by removing stop words and PorterStemmer was used for stemming words . All words were converted to lower case for standardization.

Feature Extraction

The text data after noise removal is now converted to numerical data by checking the occurrence of each word in the sentence using Count Vectorizer.This is done for title , co-author names , journal and organization. For clustering we want to construct a distance matrix where Distance matrix[i][j] is the dissimilarity between data points i and j. Each point in the distance matrix is a 5 dimensional vector containing features described above. For Names of authors jaro distance is used and for rest normalized sum of product of count vectorizer is used.

Clustering

This distance matrix was then used for clustering the data points using Spectral clustering , K-means and Agglomerative clustering techniques. The suitable clusters were found by setting a cut-off . clusters similar to ground data were found.

The above models were implemented in jupyter notebooks Kluyver et al. (2016) using the scikit-learn Pedregosa et al. (2011) package in python Van Rossum and Drake (2009) language. Stackexchange has also served as a good resource for debugging LaTex and Python scripts. All the codes used for this work can be found at this GitHub repository. Any changes you wish to suggest can be reported on GitHub itself.

References

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Schmidt, B., editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press.

Louppe, G., Al-Natsheh, H., Susik, M., and Maguire, E. (2015). Ethnicity sensitive author disambiguation using semi-supervised learning. *ArXiv e-prints*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Wang, X., Tang, J., Cheng, H., and Yu, P. S. (2011). Adana: Active name disambiguation. In *ICDM'11*, pages 794–803.