Unsupervised Decomposition of Multi-Author Document

Sayantan Sengupta and **Kautsya Kanu**
Indian Institute of Technology Delhi

Introduction

Here we propose a new unsupervised method for decomposing a multi-author document in to authorial components. We have assumed that we have no prior information about the authors and the documents, except the number of authors of the document. The key idea is to exploit the differences of the grammatical writing styles of the authors and use this information to build paragraph clusters. This is a difficult problem in many levels. Its easy to decompose based on topics and contexts, which is often known as text segmentation in literature. So it gets difficult to distinguish if multiple authors have written on the same topic. Quantifying the difference of the grammatical writing styles of authors is another big challenge. As there is no prior information/access to the authors written texts, supervised classification approaches cant be applied directly. On top of this, the number of author is not known in general of a random a document/article in general (in case of plagiarism). So fixing the number of clusters is another big task. So considering the above constraints, this paper focus more on the feature selection part of the texts which is the most important part of the whole unsupervised clustering, as good features will lead to more precise clustering of the correct sentences to their respective clusters. The traditional studies on text segmentation, as shown in Choi (2000), Brants et al. (2002), Misra et al. (2009) and Henning and Labor (2009), focus on dividing the the text into significant components such as words, sentences and topics rather than authors. There are almost no approaches, as those in Schaalje et. al. (2013), Segarra et al (2014) and Layton et al. (2013) deal with documents written by a single author only. Koppel et al. (2011) has considered the segmentation of a document according to multi-authorship, this approach requires manual translations and concordance to be available beforehand. Hence their document can only be applied on particular types of documents such as Bible books. Akiva and Koppel (2013) tried to come up with a solution. Their method relies on distance measurement to increase the precision and accuracy of the clustering and classification process. The performance is degraded when the number of authors increases to more than two.

Baseline

The latest state of the art technique used in this area is described below: Given a multi-author document written by l authors, it is assumed that every author has written consecutive sentences, and every sentence is completely written by only one of the authors. The approach goes through the following steps:

- Divide the document into segments of fixed length.
- Represent the resulted segments as vectors using an appropriate feature set which can differentiate the writing styles among authors (words occurring at least 3 times in the text).
- Cluster the resulted vectors into l clusters using an appropriate clustering algorithm targeting on high recall rates(GMM with iterative EM algorithm).
- Re-vectorize the segments using a different feature set to more accurately discriminate the segments in each cluster.
- Apply the segment Elicitation procedure, which identifies the vital segments from each clusters to improve the precision rates.
- Re-vectorize all selected segments using another feature set that can capture the differences in the writing styles of all the sentences in a document.
- Train the classifier using a Naive Bayesian model.
- Classify each sentence using the learned classifier.

Data Set

The data sets we have used to evaluate our model is:

- 690 blogs written by Gary Becker and Richard Posner.
- 1,182 New York Times articles written by Maureen Dowd, Gail Collins, Thomas Freidman and Paul Krugman.
  Each data set has its own set of challenges, since each author has written a lot of different topics and some topics are taken by both authors.

The resulting table is shown below: