PAN 2016 shared task
Author identification

Can a clustering sandwich help a neural network

in establishing authorship links?

Olivier Louwaars (s2814714), Department of Information Science

Rijksuniversiteit Groningen

Abstract

Faced with the problem of having an unknown number of authors writing an unknown number of documents that might but do not necessarily have the same topic, we came up with a pipeline of a K-means clustering algorithm informing a neural network to establish document similarity and which then informs a Meanshift algorithm that outputs the final clusters (each cluster representing an author and enveloping one or more documents). The data provided by PAN consists of 18 problems that each consist of 50-100 documents from various authors. Every problem has either news articles or reviews, and is written in English, Dutch or Greek. Documents are 15-20 sentences long, summing up to 1000-2000 sentences per problem. This is not enough training data for a neural network, so character based features were created for training. The most promising feature for this task were character based skipgrams as proposed by [], pairing every character with either a neighboring one (positive sample) or one further away (negative sample). The embeddings thus created informed the neural network with underlying information regarding character sequences and structures. Besides the lack of data, the task itself limited feature selection to character-only as well. If words or word n-grams would have been used, the system would be tricked easily into clustering on topic instead of author. As all documents within a problem have one genre, it is not unimaginable that documents from different authors have the same topic, and are therefore grouped together. Preventing the system from topic clustering and aiming it at author clustering was one of the biggest challenges in this task. Although the task is to cluster all documents of a single author, the chosen approach was to train the neural network with every possible document pair, like the PAN 2015 task for author identification. The baseline thus created was very high (94%), resulting in a default decision to most frequent class (negative) for all samples. The K-means clusterer was added to see if the number of pairs could be cut back, as a set of 50 documents already results in over 1200 possible pairs. K-means was run with an iterative setting of [1:n-1] clusters, after which the total of document clusters was counted. If two documents were never clustered together, they were stripped from the input for the neural network. This lead to a 50% reduction in document pairs, but unfortunately also a 25% reduction of correct pairs. Although this improved the baseline slightly, the data was still to biased for the neural network to be able to detect correct pairs in the training data. This lead to a useless output of the network, which gave the Meanshift no additional features to work with. Based on only character based preprocessed data, as suggested by [], Meanshift output had a precision of 0.12 on average.

Contents

1. Introduction

With the publishing and sharing of documents being accessible to everybody, the need to verify what was written by who becomes apparent. Not only to prevent plagiarism, but also to prevent texts from being attributed to an author they do not belong to. Since 2011, PAN[1] contains a shared task regarding automatic authorship identification. Where in recent years the task focused on determining whether or not a certain document belongs to a set of known documents of an author, the 2016 task is to cluster documents per author, without knowing the number of contributing authors. This task asks for a radically different approach, and offers new technical usages. If the results of this shared task are satisfactory, the method can be applied on, for example, a portfolio of documents of students, to see if the author of all documents is the same. Additionally, the task also comprises a second step, in which links have to be established between the different documents within a cluster/author. This second step is comparable with the earlier shared tasks, as it is a one on one comparison of documents. Depending on the similarity of two documents, the certainty of the author of both can be established. With this second step included in the method, the first could be used as a first separator to make an initial division in a large collection of documents. This saves a lot of time as otherwise every document in the set would have to be compared to all the others. To make sure both steps are executed properly and no work is duplicated, the following research question will be the foundation of this research:

“Can an artificial neural network cluster documents per author?”

To back up the main questions, the following sub questions can be formulated:

- What features are important for the initial clustering per author?
- What features are important for the one on one comparison?
- Can the system be prevented from clustering based on obvious but wrong patterns such as topic?

In this thesis, a solution for the PAN 2016 shared task regarding author identification, clustering, and ranking will be explored. The proposed sol

1.1. Problem

[1] http://pan.webis.de/

2. Method

2.1. Literature

2.2. Approach

3. Results

4. Conclusion