A Single Author Style Representation for the Author Verification Task

Cristhian Mayor, Josue Gutierrez, Angel Toledo, Rodrigo Martinez, Paola Ledesma , Gibran Fuentes, and Ivan Meza
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México (UNAM)

GrupoGolem
http://golem.iimas.unam.mx

We present our experience implementing three approaches for the 'PAN 2014 Author Identification' task using the same representation for the author's style. Two of our approaches extend previous successful approaches: naive Bayes and impostor methods. The third approach is based on original research on sparse representation for text documents. We present results withthe official development and test corpora.

PAN 2014 Author Identification Task

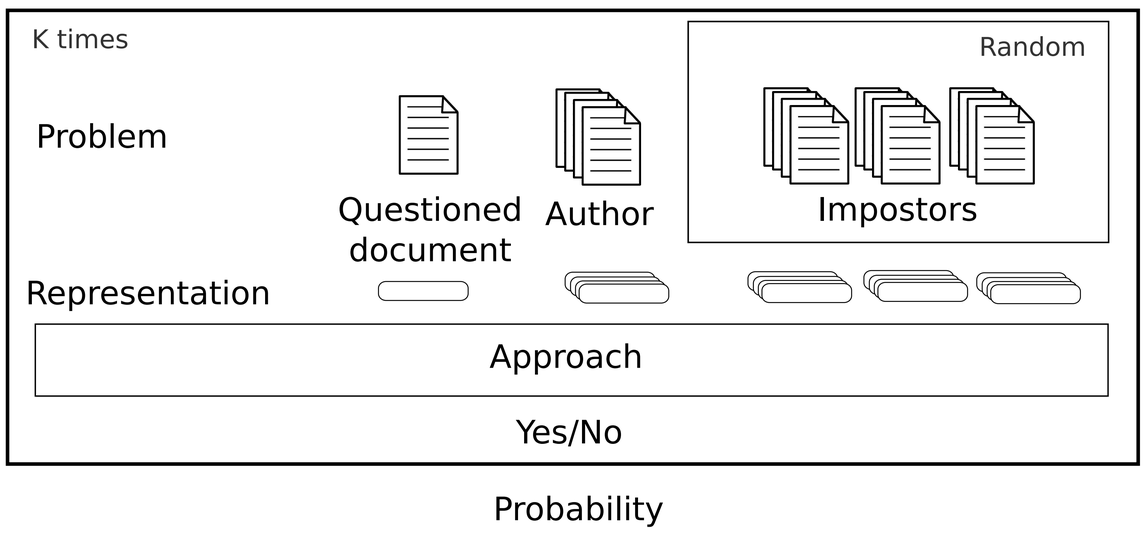
Given a small set (no more than 5, possibly as few as one) of "known"documents by a single person and a "questioned" document, the task is to determine whether the questioned document was written by the same person who wrote the known document set

Documents Representation

We experiment with a single style representation based on the following features:

- **Bag of words**
- **Bigram**
- **Trigram**
- **Punctuation**
- **Prefix**
- **Sufix**
- **Prefix bigram**
- **Sufix trigram**
- **Stop words**
- **Stop words bigram**
- **Words per sentence**

Elements of the Approaches



Impostors Approach

This method consists on iteratively compare the vector distance between the **author's document** to the **questioned document** versus the distances between several **impostor documents** to the questioned document. With these distances a score is built up based on how many times the author and questioned documents are closer than the impostor and questioned documents.

For this approach we followed the description of the method by Seidman (2013). We modified the method to work on more than one set of features and instead of using impostors from the web we used the training corpus as source of impostors. Additionally, we extended the approach to produce a probability as output based on **repetition** of the algorithm since the document instances were randomly sampled.

Naive Bayes Approach

This method consists of sampling from the author and the impostor style representations, two document instances for each. A probability score is then calculated using the common terms between the **questioned** document and the **author's documents**. On the other hand, an alternative score is calculated between the questioned document and **the impostor documents**. These scores are derived using Bayes. The purpose of the score is to capture the probability that the document was created by the same author. If the score for the author is higher than the impostor, we consider it as evidence of authorship. We iterate **k times** over this method to calculate the probability of authorship

Sparse Representation Approach

This methodology has been successfully applied to the face recognition task, in which the identity of a face image has to be determined from a set of known faces. We adapted this methodology to the authorship verification task. The method consists on identifying the components that contribute to the questioned document from samples of candidate documents from a set of authors. The rationale is that the biggest contribution of com-ponents should be elements from a single candidate author. In order to identify the componentsthe method proposes the following l-minimization

$$\textbf{minimize} \quad x_0 = \text{argmin} \, ||x||_1$$
$$\textbf{subject to} \quad Ax = y$$

Where y is the **questioned document**, A is the matrix of n samples from different m candidate authors (**impostors**), and x is the variable to minimize which represent the contribution from each candidate. So that by multiplying the samples times the contribution of the candidates we could generate the **questioned document**. From the resulting variable x_0 we can quantify the residuals given by Ax versus y and decide which author contributes with more components. We adapt this method to produce a probability as result by iterating ***k* times** over the full method.

Results

Approach	AUC	C@1	Score
impostor	62%	56%	35%
n-gram	64%	57%	36%
Sparse	72%	68%	48%

Approach	AUC	C@1	Score
Dutch reviews	93%	88%	82%
Dutch essays	57%	52%	30%
English essays	57%	56%	32%
English novels	66%	61%	41%
Greeks articles	82%	75%	62%
Spanish news	75%	71%	54%
Overall	70%	65%	50%

Conclusions

The system had a relatively good performance in general. It reached the best performance for ***Dutch reviews***.