Vector space model and Overlap metric for Author Identification

Notebook for PAN at CLEF 2013

Arun Jayapal and Binayak Goswami

Nuance Communications
jayapal.arunkumar@gmail.com, bnayok@gmail.com

Abstract This paper describes our entry for the Author Identification task at PAN 2013. The Author Identification task was performed using a combination of Vector Space Model [1] (VSM) and Similarity Overlap Metric [3] (SOM) on the character n-grams extracted from the documents related to an author and the document of question. A combination of the VSM and SOM provided an overall F-measure, precision and recall values of 0.576 each.

1 Introduction

Author identification is an important task used these days in different contexts such as plagiarism detection and forensic analysis. Being part of PAN 2013, for the Author Identification task, given a set of upto 10 documents, the objective was to classify whether the unknown document is written by the author who wrote the given set of documents. The task was performed using a combination of Vector Space Model [1] (VSM) and Similarity Overlap Metric [3] (SOM) on the character n-grams extracted from the documents related to an author and the document of question. A combination of the VSM and SOM provided an overall F-measure, precision and recall values of 0.576 each. The rest of the paper describes the approach and investigations carried out in more detail.

2 Author Identification

The author identification task is to identify whether the given document is written by the author of the known document(s). The initial challenge was to identify what features would help in identifying the author. Based on the following intuition, we identified that character 3-grams would help in identifying the author. For example in English, some authors like to write-up either in past tense, present tense, future tense or continuous tense. All these tenses in English can be identified using clear indicators such as continuous tense is always indicated by Jerund format ie., words with characters *'ing'* at the suffix. The tense will clearly indicate the style of the author. Further to this, an author might use descriptive words *(i.e., adjectives)* to represent something in their storyline. These details can be easily identified from a character 3-gram or 4-gram. The intuition was considered with expertise in English language only. Based on this intuition, we decided to use character n-gram model for the system.

2.1 Problem statement & Data

For the author identification task, we were provided with the following training data by PAN.

- 10 English, 20 Greek and 5 Spanish folders
- Each folder had upto 10 documents written by a known author
- Each folder also had a document written by an unknown author
- A gold standard was also provided for all the folders

The idea was to classify whether the document written by the unknown author is written by the known author or not.

2.2 Approach

The author identification task being closely related to intrinsic plagiarism detection, we considered author style as the only feature for classification. Therefore, we used a combination of VSM [1] and SOM [3]. The known author's documents and unknown author's document were processed before they were considered for vector space model. The processing involved converting all the spaces into underscore (_) and converting all newline characters to double underscore (__). On processing the documents, the following approaches were implemented.

Vector Space Model The known author's documents were split into 3-gram characters and considered as a vector and unknown author's document was split into 3-gram characters and was considered as a different vector. Based on the VSM, a similarity measure was computed which represented the cosine of the angle between two vectors. As mentioned in [1], following is the cosine value computation.

$$\begin{gathered} consider\ known\ documents\ vector\ be \overrightarrow{x} = (x_1, x_2, .., x_n) \\ and\ unknown\ document\ vector\ be \overrightarrow{y} = (y_1, y_2, .., y_n) \end{gathered}$$

$$\cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x}, \overrightarrow{y}}{|\overrightarrow{x}|.|\overrightarrow{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Similarity Overlap On initial evaluation, it was identified that just the cosine values produced by VSM were not enough to classify whether the unknown document was written by the author of the known documents or not. Therefore, we required another approach for the classification. The similarity overlap metric [2] attempted last year for the source retrieval task is used here to compute the 3-gram character similarity overlap between the known document and the unknown document.

$$consider\ known\ documents\ set\ X = set(x_1, x_2, .., x_n)$$

$$consider\ unknown\ document\ set\ Y = set(y_1, y_2, .., y_n)$$

$$Overlap(X,Y) = \frac{X \bigcap Y}{min(X,Y)}$$

The cosine and overlap values always range between 0 and 1, therefore we used thresold values to classify the unknown documents. The following table represents the threshold values used for classification.

Cosine	Overlap	Classify
>0.0009	>0.5	Yes
>0.09	_	Yes
_	>0.8	Yes
Otherwise	Otherwise	No

Based on the said threshold values, the unknown document was classified as known document or unknown document. Although the initial intuition was to use this methodology for English dataset, we attempted the same for all other datasets as well. The following results and discussion section provides the insight based on the results obtained for the said approach.

2.3 Results & Discussion

The baseline precision, recall and f-measure values were set at 0.500 for the author identification task. The results available at [4] are summarized in the following table. The overall precision and recall values stood at 0.576, which suggested that the overall approach carried out for author identification is a good start.

Data Set	F1	Precision	Recall
Overall	0.576	0.576	0.576
English	0.600	0.600	0.600
Greek	0.633	0.633	0.633
Baseline	0.500	0.500	0.500
Spanish	0.480	0.480	0.480

As we get into the scores for different language sets, the approach did not work well for spanish dataset. This suggests that we need to look for better language specific features to classify author. Although the approach was carried out based on intuition for English language, the results suggests that the same language features work better for Greek than for English. Since we have used two very simple approaches to identify the author, the time complexity of the system proves to be the best of all.

3 Conclusion

The author identification task was completed with better than baseline results for all the datasets, while the precision and recall values for spanish dataset were lower than baseline. Therefore the features for spanish need to be selected carefully understanding the language specification and flavour. Further to this, the current system produced promising results with a combination of vector space model and similarity overlap metric, which can further be experimented with diferent character grams. Moreover, we will be scrutinizing other language specific features which may be useful to identify the document's author.

References

1. G.Salton, A.Wong, C.S.Yang: A vector space model for automatic indexing. In: C.A.Montgomery (ed.) Communications of ACM. vol. 18, pp. 613 – 620 (November 1975)
2. Jayapal, A.: Similarity overlap metric and greedy string tiling at pan 2012: Plagiarism detection. In: Notebook for PAN at CLEF 2012 (2012)
3. Nawab, R.M.A., Stevenson, M., Clough, P.: University of sheffield lab report for pan at clef 2010 (2010)
4. PAN: Author identification. Online (June 2013), http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/pan13-ai-final-results.pdf