Experiments in Authorship-Link Ranking and Complete Author Clustering

Valentin Zmiycharov[1], Dimitar Alexandrov[1], Hristo Georgiev[1],
Yasen Kiprov[1], Georgi Georgiev[1], Ivan Koychev[1], and Preslav Nakov[2]

[1] FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria {valentin.zmiycharov, dimityr.alexandrov, hristo.i.georgiev}@gmail.com,
{yasen.kiprov, g.d.georgiev}@gmail.com, koychev@fmi.uni-sofia.bg
[2] Qatar Computing Research Institute, HBKU, Doha, Qatar pnakov@qf.org.qa

INTRODUCTION & OBJECTIVES

Multiple independent tasks:

- Up to 100 documents
- Single-authored
- Same language
- Same genre
- # of distinct authors – unknown

Objectives:

- **Complete author clustering result**: Each cluster should contain all documents found in the collection by a specific author. The clusters should be non-overlapping, i.e., each document should belong to exactly one cluster.
- **Authorship-link ranking result**: A list of document pairs ranked according to a real-valued score in [0,1], where higher values denote higher confidence that the pair of documents are written by the same author.

Training set:

- 3 languages (English, Dutch, Greek)
- 2 genres (Articles, Reviews)
- 18 folders (3 for each pair)

OUR APPROACH

Feature:

A function with 2 variables: document1 and document2. It returns a real number which is the difference between those 2 documents.

Features list:

- **Average sentence length**

Returns the absolute difference between the average words count per sentence of the documents.

- **Function words ratio**

Function words or stop words refer to the most common words in a language. Even though usually function words are filtered out before processing of natural language data in this particular task they are quite useful. This feature calculates the number of stop words divided by the number of sentences of each document. Then it returns the absolute difference of those values.

- **Type-token ratio**

Returns the absolute difference between the number of unique tokens divided by the total number of tokens per document.

- **Features, based on part of speech**

For each part of speech we consider important (nouns, adjectives and verbs) we do the following:

1. Calculate the count of the part for each sentence.
2. Order the calculated numbers in ascending order.
3. Calculate a vector of five elements including the minimum value, the first quartile, the median, the third quartile and the maximum value of the row.
4. The feature returns the Euclidean distance between the 2 vectors.

CONTACT

https://github.com/pan-webis-de/zmiycharov16

You can find me at:

- valentin.zmiycharov@gmail.com
- https://github.com/v-zmiycharov
- https://www.linkedin.com/in/zmiycharov

OUR SOLUTION

Authorship-link results:

- Input parameters: features results for each pair of documents
- Without normalization
- Classification problem - LibSVM
- 6 different classifiers

Complete author clustering:

- Relies entirely on authorship-links results
- 4 steps:

1. Generate clusters of 2 documents for each pair that is considered by the same author.
2. Add a document to cluster if it is similar to more than the half of the documents in the cluster.
3. Find documents which exist in more than one cluster and remove all except for the most relevant one.
4. Create clusters of one document for non-clustered documents.

RESULTS & ANALYSIS

Feature results on training set:

Feature name	Avg difference for same author	Avg difference for different authors
Average sentence length	14.67	22.55
Function words ratio	0.14	0.15
Type-token ratio	0.04	0.07
Nouns ratio	4.26	4.70
Adjectives ratio	2.21	2.50
Verbs ratio	4.57	4.62
Postag conjunction	2.06	2.47

PAN Contest evaluation:

- Clustering: BCubed F-score
- Authorship links: Mean Average Precision
- 11 teams
- 7 successful submissions
- 4th place: clustering
- 7th place: authorship links

Language	Genre	F-Cubed	R-Cubed	P-Cubed	Av-Precision
english	articles	0.77326	0.71429	0.84286	0.0030303
english	articles	0.64987	0.50408	0.91429	0.0024442
english	articles	0.88479	0.91429	0.85714	0
english	reviews	0.80802	0.725	0.9125	0
english	reviews	0.91233	0.9	0.925	0
english	reviews	0.65973	0.525	0.8875	0
dutch	articles	0.77824	0.73684	0.82456	0
dutch	articles	0.86833	0.87719	0.85965	0
dutch	articles	0.6529	0.52632	0.85965	0
dutch	reviews	0.85953	0.88	0.84	0
dutch	reviews	0.64029	0.51	0.86	0.0020102
dutch	reviews	0.75232	0.71	0.8	0
greek	articles	0.7483	0.71429	0.78571	0.010809
greek	articles	0.6378	0.50952	0.85238	0.003261
greek	articles	0.85619	0.88571	0.82857	0.0024691
greek	reviews	0.80194	0.75714	0.85238	0.0051692
greek	reviews	0.88102	0.92857	0.8381	0.011905
greek	reviews	0.66584	0.57143	0.79762	0.018752