Author profiling

Lovre Mrčela, Marko Ratković, Ante Žužul

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{lovre.mrcela, marko.ratkovic, ante.zuzul}@fer.hr

Abstract

The goal of this project was to profile an author by analyzing a set of texts written by them, and then determining degree of each the Big Five personality traits. In addition, gender and age–group for each author are derived as well. The dataset was collected from twitter profiles, in English, Italian, Spanish and Dutch. Approach was based on *tf-idf*, considering occurences of trigrams.

1. Introduction

Author profiling deals with problem of describing someone's personality, by way of extracting information from their writing style. Personality can be described using five traits (the so-called *"Big Five personality traits"*), which are: extraversion, stability, agreeableness, conscientiousness and openness to experience. Degrees of each trait range from -0.5 (indicating the total opposite) to 0.5 (indicating the exact match).

Provided with degrees of the five traits, it is possible to determine author's gender and age–group, via means of classification based on a model trained on previously labelled data. In this project, we used the linear SVC and Gaussian naive Bayes models for the classification into gender and age–group, and the linear regression with squared error measure for determining the degrees of personality traits. The training set we used was a collection of twitter posts in English, Spanish, Italian and Dutch authors, ranging from around 35 authors in Dutch to 150 in English, each author's file containing about 100 posts.

2. Approach

This is the second section. In scientific papers this is usually (but not necessarily) the section in which related research is (briefly) described.

2.1. Text preprocessing

For the rest of the process to be optimal, some sort of text preprocessing needs to be done on the raw input data. The input data we use is given in *xml* format, so the first step in preprocessing was to parse the actual sentences from the *xml* structure. When that is done, following steps are also applied:

- *urls* to other sites are substituted with an URL tag,
- other user names (when referenced in replies) are substituted with a REPLY tag,
- all the text is converted to lower case,
- each three consecutive letters are grouped into trigrams, and
- weighted vector of trigrams is obtained by using *tf–idf* weighting scheme.

2.2. Gender and age–group classification

For the gender and age–group classification subproblem, following approaches were considered:

- logistic regression
- naive Bayes classifier
- decision tree classifier
- random forest classifier
- SVC (using *rbf*, linear, poly– or sigmoid kernels)

Results of using each approach are compared in the table 1. The best result was obtained by using SVC with linear kernel.

2.3. Personality traits regression

2.3.1. Sub-subsection example

This is a sub-subsection. If possible, it is better to avoid sub-subsections.

3. Results

The results.

4. Extent of the paper

The paper should have at least. The paper should have a minimum of 3 and a maximum of 5 pages plus an additional page for references.

5. Figures and tables

5.1. Figures

Here is an example on how to include figures in the paper. Figures are included in LaTeXcode immediately *after* the text in which these figures are referenced. Allow LaTeXto place the figure where it believes is best (usually on top of the page of at the position where you would not place the figure). Figures are references as follows: "Figure 1 shows ...". Use tilda (~) to prevent separation between the word "Figure" and its enumeration.

5.2. Tables

There are two types of tables: narrow tables that fit into one column and a wide table that spreads over both columns.

Figure 1: This is the figure caption. Full sentences should be followed with a dot. The caption should be placed *below* the figure. Caption should be short; details should be explained in the text.

Table 1: Comparison of results obtained by using different classifiers.

Heading1	Heading2
One	First row text
Two	Second row text
Three	Third row text
	Fourth row text

5.2.1. Narrow tables

5.3. Wide tables

Table 2 is an example of a wide table that spreads across both columns. The same can be done for wide figures that should spread across the whole width of the page.

6. Math expressions and formulas

Math expressions and formulas that appear within the sentence should be writen inside the so-called *inline* math environment: $2 + 3$, $\sqrt{16}$, $h(x) = \mathbf{1}(\theta_1 x_1 + \theta_0 > 0)$. Larger expressions and formulas (e.g., equations) should be written in the so-called *displayed* math environment:

$$b_k^{(i)} = \begin{cases} 1 & \text{ako } k = \text{argmin}_j \|\mathbf{x}^{(i)} - \mu_j\| \\ 0 & \text{inače} \end{cases}$$

Math expressions which you reference in the text should be written inside the *equation* environment:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} b_k^{(i)} \|\mathbf{x}^{(i)} - \mu_k\|^2 \tag{1}$$

Now you can reference equation (1). If the paragraphs continues right after the formula

$$f(x) = x^2 + \varepsilon \tag{2}$$

like this one does, then use the command *noindent* after the equation to prevent the indentation of the row starting the paragraph.

Multiletter words in the math environment should be written inside the command *mathit*, otherwise LaTeXwill insert spacing between the letters to denote the multicplication of values denoted by symbols. For example, compare $Consistent(h, \mathcal{D})$ and $Consistent(h, \mathcal{D})$.

If you need a math symbol, but you don't know the command for it in LaTeX, try *Detexify*.[1]

7. Referencing literature

References to other publications should be written in brackets with the last name of the first author and the year of publication, e.g., (Chomsky, 1973). Multiple references are written in sequence, one after another, separated by semicolon and without whitespaces in between, e.g., (Chomsky, 1973; Chave, 1964; Feigl, 1958). References are typically written at the end of the sentence and necessarily before the sentence punctuation.

If the publication is authored by more than author, only the name of the first author is written, after which abbreviation *et al.*, meaning *et alia*, i.e., and others is written as in (Johnson et al., 1976). If the publication is authored by only two authors, then the last names of both authors are written (Johnson and Howells, 1974).

If the name of the author is incorporated into the text of the sentence, it should be out of the brackets (only the year should be in the brackets). E.g., "Chomsky (1973) suggested that . . . ". The difference is whether you reference the publication or the author who wrote it.

The list of all literature references is given alphabetically at the end of the paper. The form of the reference depends on the type of the bibliographic unit: conference papers, (Chave, 1964), books (Butcher, 1981), journal articles (Howells, 1951), doctoral dissertations (Croft, 1978) and book chapters (Feigl, 1958).

All of this is produced for you automatically by using BibTeX. Sve ovo dobivate automatski ako. In the file `tar2014.bib` insert the BibTeX entries, and then reference them via their symbolic names.

8. Conclusion

Conclusion is the last enumerated section of the paper. Conclusion should not exceed half of the column and is typically be split into 2–3 paragraphs.

Acknowledgements

If suited, before inserting the literature references you can include the Acknowledgements section in order to thank those who helped you in any way to deliver the paper, but are not co-authors of the paper.

References

Judith Butcher. 1981. *Copy-editing*. Cambridge University Press, 2nd edition.

K. E. Chave. 1964. Skeletal durability and preservation. In J. Imbrie and N. Newel, editors, *Approaches to paleoecology*, pages 377–87, New York. Wiley.

N. Chomsky. 1973. Conditions on transformations. In S. R. Anderson and P. Kiparsky, editors, *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

W. B. Croft. 1978. *Organizing and searching large files of document descriptions*. Ph.D. thesis, Cambridge University.

[1] `http://detexify.kirelabs.org/`

Table 2: Wide-table caption

Heading1	Heading2	Heading3
A	A very long text, longer that the width of a single column	128
B	A very long text, longer that the width of a single column	3123
C	A very long text, longer that the width of a single column	−32

F. Feigl, 1958. *Spot Tests in Organic Analysis*, chapter 6. Publisher publisher, 5th edition.

W. W. Howells. 1951. Factors of human physique. *American Journal of Physical Anthropology*, 9:159–192.

G. B. Johnson and W. W. Howells. 1974. Title title title title title title title title title. *Journal journal journal*.

G. B. Johnson, W. W. Howells, and A. N. Other. 1976. Title title title title title title title title title. *Journal journal journal*.