Learning from Data – Final Project
Author Profiling

General remarks

Differently than in the other assignments, for this final project we are not specifying exactly how you should do tackle the task you're given. In other words: the *way* you approach this problem is up to you, and it is an important part of the assignment itself, as it is normally the case that given a (learning) problem you will have to identify your best strategy to deal with it. Grading will be determined by your model, your report, and the final presentation you will give for everyone in class.

Deadline for submission on Nestor: 26th January 2015, end of day.
What you have to hand in by the deadline:

- script with your system (or systems, if you produce more than one to cope with different languages). Your script should take a training directory and a test directory as arguments. You can assume that training and test directories are under the same parent directory. Please, remember that the output of your system should be a file like `truth.txt` that you find in the training folders, excluding the final five columns (personality scores). You should output one `truth.txt` per language, thus a total of four `truth.txt` files, one per language subdirectory.
- report with system description (see Section 2 for details. Please, make sure to hand in a **pdf** file following the usual report template)

After submission, your systems will be run on test data which we have withheld, and you can report results on your presentation, together with a description of what you have done and your system.

Deadline for uploading your presentation on Nestor: January 28th, end of day (pdf file).

Presentations' schedule: TBA (Week of January 30th, 2017)

1 Task: Author profiling

"Author profiling, [...], distinguishes between classes of authors, rather than individual authors. Thus, for example, profiling is used to determine an author's gender, age, native lan-

guage, personality type, etc. Author profiling is a problem of growing importance in a variety of areas, including forensics, security and marketing. For instance, from a forensic linguistics perspective, being able to determine the linguistic profile of the author of a suspicious text solely by analyzing the text could be extremely valuable for evaluating suspects. Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products. Here we consider the problem of author profiling in social media, with particular focus on the use of everyday language and how this reflects basic social and personality processes." (Rangel et al., 2013)

Author profiling is organised as a yearly *shared task* within PAN: `pan.webis.de`.

At this website you also find information on all systems that participated in this task, and you might get some ideas on how to tackle this problem: `http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/about.html`.

These papers describe previous shared tasks on author profiling, and you're encouraged to read them.

- 2016: `http://www.uni-weimar.de/medien/webis/publications/papers/stein_2016l.pdf`
  (Rangel et al., 2016)
- 2015: `http://www.sensei-conversation.eu/wp-content/uploads/2015/09/15-pan@clef.pdf`
  (Pardo et al., 2015)
- 2014: `http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-RangelEt2014.pdf`
  (Rangel et al., 2014)
- 2013: `http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf`
  (Rangel et al., 2013).

While you are obviously encouraged to check the attempts of others, you are also advised not to spend too much time checking/reading all the papers, as it would be far too much time consuming. Also please remember that any inspiration you get is welcome, but you should always mention where you got it from by citing the appropriate sources. In the report, mention which techniques/resources you're using from previous work, and what instead is more like a new idea you had.

1.1 Task summary

Given a tweet, the task is to determine age and gender of the author.

Gender is a binary classification task, where the values are F and M, while age is a multi class problem where class values are {18-24, 25-34, 35-49, 50-xx.}

1.2 Training data

The training corpus comprises a collection of tweets which have been annotated for gender, age, and a series of personality traits that we are going to ignore, for the moment.

This is a multilingual task, so that you're asked to do author profiling for four different languages (English, Italian, Spanish, Dutch). Training data is organised in different directories, per language, all under a `training` directory:

- `training/dutch`
- `training/english`
- `training/italian`
- `training/spanish`

In each of them, you will find a set of XML documents plus a `truth.txt` file. The XML documents are the tweets, while the `truth.txt` file contains the gold labels. Gold values are associated to tweets via ids, which are the names of the XML files. The gold labels we are interested in are those concerning gender and age (the first two columns after the id, separated by `:::`)

Please, note that data was collected automatically and it isn't entirely noise-free, especially as far as the language is concerned: it is possible that a tweet in the English dataset isn't actually in English. This is noise that is to be accounted for in real settings, so it isn't such a bad thing, but you should bear this in mind.

2 What you have to do

- You have to produce **one or more author profiling models**, that is a model that given a tweet will output the age and gender of the author (age is one of four age groups). We will run your model(s) on test data which you haven't seen before. Please, make sure that the output of your system is a `truth.txt` file, in the same format as the truth file you're given with training data. Your script should also output accuracy, precision, recall, and f-score (and if you wish a confusion matrix) when compared to gold labels, but do not print them on the `truth.txt` file. You can decide to tune your system by setting aside a development set (using standard splits as we've done for assignments), or via cross-validation.

  Structure of test data The structure of the test directory is identical to that of the training directory: `test/dutch`, `test/english`, `test/italian`, `test/spanish`. Your system will produce one `truth.txt` file per subdirectory.

- This is a multilingual task, so that you will have to develop author profiling for more than one language (four). Whether you want to develop a system that is language-independent or you want to develop four different models is entirely up to you.
- You can obviously use all the support from scikit-learn and NLTK for this, and any other library you find useful. You can use any features and any learning algorithms you like. You are encouraged to experiment with several different features. You are also welcome to incorporate more data, if you have it and wish to do so. Anything you use will have to be mentioned in your report.
- You are also asked to produce a **report**. The report should contain the explanation

of how you tackled this problem, a description of the features you used, any feature selection method you applied, the algorithm(s) you chose to learn your model, including parameter tuning and setting, any additional data/resources you incorporated, and how well you do on training data in terms of accuracy, precision, recall, f-score. You should also justify your choices explaining why you selected a certain approach, certain features, the learning algorithm, and so on. Please, refer to the literature (see above) if you use and/or implement what other people have already done.

- Finally, you are asked to produce a **presentation** in which you will explain to the others what you have done, and why. You will have 15 minutes for this, including questions (think in terms of 10+5). Please, bear in mind that the presentation will contribute to the final grading as well. Before the presentation you will be given your results on the test set (we will run your system as soon as we get it, so if there are no glitches, you will get results on test data right away after submission), so that you can refer to those as well.

References

Pardo, F. M. R., F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans (2015). Overview of the 3rd author profiling task at PAN 2015. In L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan (Eds.), *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, Volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rangel, F., P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans (2014). Overview of the author profiling task at pan 2014. In *Proceedings of CLEF 2014*.

Rangel, F., P. Rosso, M. Koppel, E. Stamatatos, and G. Inches (2013). Overview of the author profiling task at pan 2013. In *Proceedings of CLEF 2013*.

Rangel, F., P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein (2016). Overview of the 4th author profiling task at pan 2016: cross-genre evaluations.