UniNE at CLEF 2016: Author Profiling

Mirco Kocher, Jacques Savoy

University of Neuchâtel, Switzerland

Swiss National Science Foundation

Task

Given a short set of tweets, predict an author's age and gender in a different genre.

Language	Goal	Genre Training	Genre Testing
Dutch	Gender	Twitter	Reviews
English	Gender, Age	Twitter	Blogs
Spanish	Gender, Age	Twitter	Blogs

- 5 age groups: 18-24, 25-34, 35-49, 50-64, and ≥65
- *TIRA* platform for evaluation
  - Encapsulated system with restricted data access
  - Fair comparison of the time needed to produce an answer
- Total 22 different participants

Strategy

- Select 200 most frequent terms from the query text

Term	Q	…	Ada	Alan	Ken	Tim	Vint	…
the	0.09		0.08	0.10	0.08	0.07	0.07	
of	0.06		0.06	0.06	0.04	0.04	0.03	
to	0.04		0.03	0.03	0.06	0.07	0.06	
.	0.03		0.04	0.03	0.01	0.05	0.01	
#tag	0.02		0.00	0.01	0.00	0.00	0.01	
and	0.01		0.01	0.01	0.01	0.03	0.02	
…								

≤200

Number of samples in training set

- L_1-norm to select five nearest neighbors
  - $\Delta(Q, Ada)$: 0.05 -> 2nd neighbor
  - $\Delta(Q, Alan)$: 0.03 -> 1st neighbor
  - $\Delta(Q, Ken)$: 0.09 -> 3rd neighbor
  - $\Delta(Q, Tim)$: 0.13 -> 5th neighbor
  - $\Delta(Q, Vint)$: 0.11 -> 4th neighbor
  - … …
- Determine gender and age with majority voting
  - Select nearest if no unique majority exists

Rank	Gender	Age
1 (Alan)	Male	35-49
2 (Ada)	Female	35-49
3 (Ken)	Male	≥65
4 (Tim)	Male	50-64
5 (Vint)	Male	≥65

  - Predict: Male 35-49

Results

Language	Joint	Gender	Age	Runtime
Dutch	0.5040	0.5040	-	00:02:27
English	0.2564	0.5769	0.4103	00:01:18
Spanish	0.1964	0.5357	0.3393	00:00:30

- Joint = accuracy of age and gender
- No age prediction in Dutch required
- Number of samples and text length is important
- Performance loss due to different genres
  - Gender accuracy drops by ~30%
  - Age prediction 50% less reliable

Evaluation

- Explanation of our proposed assignment (*e.g.* in English)
  - Usually, the relative frequency differences with very frequent words such as *when*, *is*, *in*, *that*, *to*, or *it* can explain the decision
- Difference in writing style
  - Female author tend to use more pronouns. Male authors seem to use more determiners and big words (more than 6 characters)
  - First person singular pronouns and full stops most common among young writers, but few first person plural pronouns and big words.
- Overall
  - Dutch gender prediction low for most participants

Conclusion

- Simple supervised approach can solve the profiling problem
- Reduced set of comprehensible features explains the decision
- No language dependent adjustments or parameter training
- Most frequent terms tend to select most discriminative features
- Difference in writing style exists independent of the genre

Acknowledgments

This research was supported, in part, by the NSF under Grant #200021_149665/1.

Mirco Kocher, Jacques Savoy – University of Neuchâtel – Rue Emile-Argand 11 – 2000 Neuchâtel, Switzerland – {Mirco.Kocher, Jacques.Savoy}@unine.ch

http://style.unine.ch/