PROFILE-BASED APPROACH FOR AGE & GENDER IDENTIFICATION

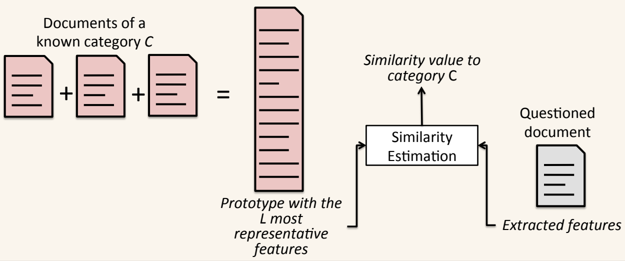
M. J. Garciarena-Ucelay[1], M. P. Villegas[1], D. G. Funez[1], L. C. Cagnina[1,2],
M. L. Errecalde[1], G. Ramírez-de-la-Rosa[3] & E. Villatoro-Tello[3]

[1]LIDIC Research Group, Universidad Nacional de San Luis, Argentina; [2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET);
[3]Language and Reasoning Research Group, Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, México.

INTRODUCTION

- The main goal of the ***Author Profiling*** is to distinguish, from a given text, among different authors' categories and not to identify the author itself.
- This task aims at *modeling* groups of authors through more general set of features.
- Such *features* will represent how different categories of authors employ language depending on its age, gender, native language, political preference, personality, etc.

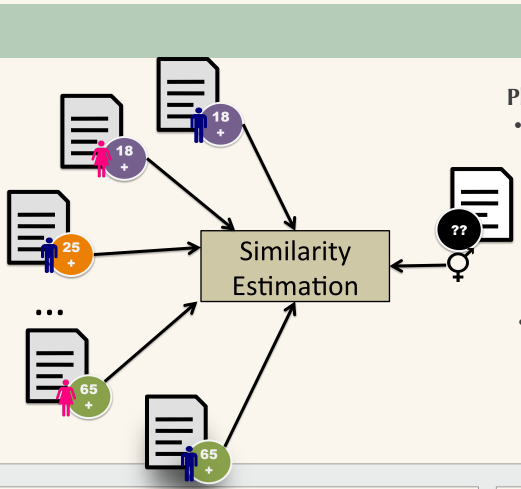
PROFILE-BASED APPROACH


Documents of a known category *C*
+
+
=
Prototype with the *L* most representative features
Similarity Estimation
Similarity value to category C
Questioned document
Extracted features

OUR METHOD

Phase One:

- *Unification*: all xml files concatenated in a single txt file, one per category.
- *Preprocessing of the txt*: remove tags and images.
- *N-grams extraction* with frequencies.
- *Sort* the n-grams by frequency.
- *Save the profile* of the category considering the L most frequent n-grams obtained in the previous step.



Phase Two:

- *Preprocessing of the txt*: remove tags and images.
- *L most frequent N-grams extraction* (unknown profile).
- *Compare* the unknown document profile (D) with the profile of each category (P_c).
- Return the category with the highest value of *Sim*.

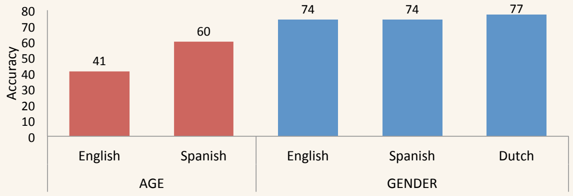
$$Sim = \sum_{x \in P_c \cap D} \left(\frac{2 \times (f_{P_c}(x) - f_D(x))}{f_{P_c}(x) + f_D(x)} \right)^2$$

EXPERIMENTS

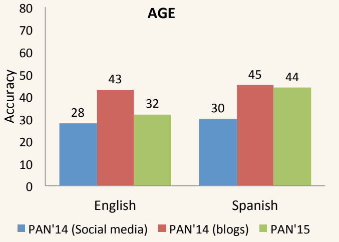
Tuning phase (L and n):

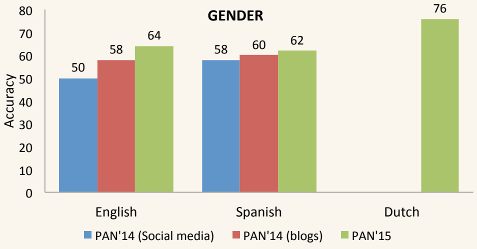
Intra genre:

- Best *L*=4000
- best *n*=3



Cross genre: Best *L*=8000, best *n*=3





CONCLUSIONS

- We presented a profile-based method for the Author Profiling task.
- Our proposal uses profiles of character 3-grams and lenght of 8000 terms.
- We performed experiments intra and cross genre scenarios.

As a future work, we plan:

- Test different features for the construction of the profiles.
- Use different similarity measures for comparing the profiles.