Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features

Maria De-Arteaga, Sergio Jimenez, George Dueñas, Sergio Mancera and Julia Baquero

Universidad Nacional de Colombia. Bogotá, Colombia [mdeg|sgjimenezv|geduenasl|samanceran|jmbaquerov]@unal.edu.co

Introduction

In our study, each document is represented as a vector, where each feature adds one unit to the dimension. These features include stylistic, corpus-extracted and lexicon-based attributes, which are relevant to distinguish gender and age-range of authors.

Official Results

	Total	Gender	Age
Spanish	0.3145	0.5627	0.5429
English	0.2450	0.4998	0.4885
Baseline	0.1650	0.5000	0.3333

6th place in the Spanish corpus

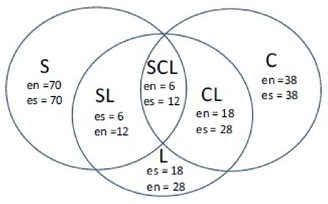

Fig. 1: Number of features by category.

Stylistic Features

Character-based: Densities of letters [a-z], uppercase and lowercase characters, punctuation marks, and special characters [*#@%&].
Word-based: Average word length and densities of hapaxlegomenas, dislegomenas until 5-legomenas.
Stylistic lexicons: Features obtained from lexicons indicative of style, e.g. stop words and dictionaries.

Lexicon-based Features

Content: Cooking, legal, love-sex and sports.
Style: Badwords, Internet, stopwords, and spanish/english dictionaries.
Emotions: Sidorov's sentiment analysis lexicon.
Characteristic words: built with T-test.
• We calculate *density*, *weighted density*, *collection and document frequency entropy*.

Unsupervised Corpus Statistics

This set of 6 features is built from statistics gathered from the training corpus, ignoring the demographic categories *age* and *gender* associated to each document. Features involving the probability of a given word are approximated using both the collection frequency and the document frequency (this second one yields a value that is proportional to a probability distribution).

$$P_f(w) = \frac{f(w)}{W} \qquad P_{df}(w) = \frac{df(w)}{D}$$

IR features Measures the informative character of the words in a given document.

$TF.IDF(w,d) = \frac{\sum_{w \in d} tf(w,d) \cdot idf(w)}{len(d)}$.

Entropy

$H(d) = \sum_{w \in d} P(w) \cdot log_2(P(w))$.

Kullback-Leibler divergence Measures the information loss when a document probability distribution Q is used to aproximate the 'true' corpus distribution P.

$$Q_d(w) = \frac{tf(w,d)}{len(d)} \qquad P_d(w) = \frac{f(w)}{\sum_{v \in d} f(v)}$$

$P||Q(d) = \sum_{w \in d} P_d \cdot \ln(\frac{P_d(w)}{Q_d(w)})$

Cross entropy Similarly to the KL-divergence, compares P and Q measuring the ability of the former for predicting the latter.

$H(P_d, Q_d) = -\sum_{w \in d} P_d(w) \cdot \log_2(Q_d(w))$.

Supervised Corpus Statistics

Unlike the previous set of features, this collection of features was built taking into account the age and gender of the authors of the training documents.
Bayes score Using Bayes Theorem, calculates the probability of a demographic cateogry given a word and aggregates this values.

$BS_C(d) = \sum_{w \in d} P(C|w)$

Gender score Aggregates the differences between the probabilities of a word w estimated in the corpus of documents written by males and females.

$GS(d) = \sum_{w \in d}(P(w|male) - P(w|female))$

Experimental Results

Using the first 20,000 documents of the training set, we carried out experiments by using ten-fold cross validation. For different groupings of features, the avarage of ten random folds is reported with its corresponding standard deviation. Given the consistency between our Spanish and English experiments, we suspect there might be a bug in the english system we submitted.

Three Main Feature Sets

Feature Set	Gender *en*	Age *en*	Gender *es*	Age *es*
Statistic	0.8393(0.0005)	0.7860(0.0013)	0.8038(0.0007)	0.7866(0.0004)
Lexicon	0.5933(0.0010)	0.6198(0.0003)	0.6261(0.0007)	0.6446(0.0006)
Stylistic	0.5502(0.0012)	0.6048(0.0003)	0.5981(0.0008)	0.6336(0.0009)
All	0.8477(0.0023)	0.7809(0.0002)	0.8202(0.0013)	–

Supervised and Unsupervised Features

Feature Set	Gender *en*	Age *en*	Gender *es*	Age *es*
Supervised	0.8432(0.0003)	0.7968(0.0006)	0.8155(0.0007)	0.7941 (0.0005)
Unsupervised	0.5487(0.0012)	0.6075(0.0006)	0.5990(0.0005)	–

Subcategories in the "Statistics"Feature Set

	Gender *en*	Age *en*	Gender *es*	Age *es*
Bayes	0.7951(0.0004)	0.7382(0.0015)	0.7696(0.0002)	0.7677(0.0003)
Cross entropy	0.5527(0.0008)	0.5891(0.0006)	0.5376(0.0006)	0.5624(0.0004)
Kullback	0.5485(0.0005)	0.6034(0.0003)	0.5896(0.0005)	0.5952(0.0007)
tt Lexicons	0.5863(0.0006)	0.6204(0.0004)	0,6240(0.0005)	0.6377(0.0003)
Word given X	0.5416(0.0007)	0.6165(0.0003)	0.6152(0.0007)	0.5979(0.0003)

Supervised KL-divergence Analogous to the unsupervised KL-divergence, measures the information loss when a document frequency distribution is used to aproximate the probability distribution of the sub-corpus of a demographic category.
Supervised cross entropy It compares the document probability distribution with the probability distribution of the sub-corpus of a demographic category, as described for the unsupervised feature.

Supervised lexicon extraction using T-test The Student's t-test allows us to determine the most characteristic words of each demographic category.

$$T_g = \frac{P_f(w|male) - P_f(w|female)}{\sqrt{\frac{s^2_{male}}{D_{male}} + \frac{s^2_{female}}{D_{female}}}}$$

$$T_A = \frac{P_f(w|A) - P_f(w)}{\sqrt{\frac{s^2_A}{D_A} + \frac{s^2}{D}}}.$$