Assigment 3

Manuel Montes-y-Gómez , PhD - INAOE - Mexico

Catedra Internacional de Ingeniería

Universidad Nacional de Colombia

Consists in assign a text of unknown authorship to one candidate author, given a set of candidate authors for whom text samples are available.

We will use a corpus texts from 5 Mexican Poets Efraín Huerta, Octavio Paz, Jaime Sabines, Rosario Castellanos, Rubén Bonifaz.

Authorship Attribution using Word Sequences. Rosa María Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez and Paolo Rosso. CIARP 2006.

I suggest using: Python and Weka. You can use the language and classification tool of your preference.

The purpose is two determine what is more important for discriminating among these authors: content or style.

1. You have to evaluate representations. One has to be based on the BoW representation (baseline result). For both classifiers use the same classifier.

2. Do 10CFV and report macro F1 results

3. Deadline July 5. Send a report via Drobox request with the introduction of the experiment and the description of used representations, experimental setup (dataset, classifier configuration, evaluation measures) and results and their discussion. Send all the files needed in a .zip named nlp-assign3-nombreapellido1-nombreapellido2.zip. Groups of 2.The data set is available in the web page or the course.