Text Mining II - Diploma en Big Data (2014-2015)

Construcción del baseline PAN-AP-2013

Francisco Rangel
Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Objetivo

- Disponer de un método comunmente usado en el estado del arte que nos permita:
  - Tener una representación base de los datos
  - Obtener unos resultados preliminares que marcarán el mínimo a conseguir
  - Poder comparar nuestra representación con esos resultados preliminares

Modelo de representación

- Selección del modelo de representación
  - Bolsa de palabras
- Preprocesado del dataset
  - Eliminación de html
  - Agrupamiento de todos los posts de un mismo autor
- Selección de características
  - Los 1000 términos más frecuentes
  - NO vamos a eliminar las palabras vacías

Representación de los documentos

- Training vs. test
  - Las características (1000 términos más frecuentes) se deben obtener del corpus de training y utilizarse para generar las representaciones tanto del corpus de training como de test
  - Vamos a representar la bolsa de palabras como la frecuencia relativa de cada palabra de la bolsa en el documento (una alternativa sería la representación binaria indicando la presencia/ausencia de cada palabra de la bolsa en el documento, y otra representación alternativa más elaborada sería la basada en pesos tf/idf)

Aprendizaje automático

- ¿Qué herramienta de aprendizaje vamos a utilizar?
  - Determinará el formato de la representación (pe. ficheros arff en Weka)
- Aprendizaje
  - Vamos a aprender un modelo a partir de la representación del training
- Evaluación
  - Vamos a evaluar con la representación del test
  - Por comparabilidad con PAN, usaremos accuracy (ratio de aciertos y totales)

Material

- **Script para generación de baselines**
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/src/GenerateBOWBaseline.zip
- **Bolsas de palabras**
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/bow-en.txt
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/bow-es.txt
- **Representación del corpus en formato arff (Weka)**
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-training-es-gender.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-training-es-age.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-training-en-gender.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-training-en-age.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-test-es-gender.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-test-es-age.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-test-en-gender.arff.zip
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data/pan-ap-13-test-en-age.arff.zip

Big Data

- Hemos generado la baseline con las 1000 palabras más frecuentes. Aún asi, esto supone unos ficheros arff de tamaños considerables:
  - Training-es: 377.1MB
  - Training-en: 1.6Gb
- Se hace preciso:
  - Hacer un sampling de los datos
  - Aplicar técnicas de Big Data:
    - En *"Suraj Maharjan, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In Advances in Artificial Intelligence. Iberamia, pages 95–107, 2014."* los autores usan MapReduce para modelar los documentos con aproximadamente 3 millones de características