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

Exploración del Dataset PAN-AP-2013

Francisco Rangel
Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Objetivos de la exploración

- Entender propiedades de los datos
- Encontrar patrones
- Depurar
- Comunicar resultados

Descarga del Dataset

- Training:

  http://www.uni-weimar.de/medien/webis/corpora/corpus-pan-labs-09-today/pan-13/pan13-data/pan13-author-profiling-training-corpus-2013-01-09.zip

- Test:

  http://www.uni-weimar.de/medien/webis/corpora/corpus-pan-labs-09-today/pan-13/pan13-data/pan13-author-profiling-test-corpus2-2013-04-29.zip

Características del dataset

- Colección con decenas/cientos de miles de autores
- Obtenido de redes sociales
- Gran variedad de temas
- Conversaciones sobre sexo vs. predadores sexuales
- Dificultad para etiquetar la información
- Personas reales vs. robots (chatbots)
- Multilingue: inglés + español

¿Big Data?

Criterio de selección

- Posts agrupados por autor
- Se mantienen autores con pocos posts
- Se parte (chunk) autores con más de 1000 palabras
- Se equilibra por sexo
- Grupos de edad no equilibrados:
  - 10s (13-17)
  - 20s (23-27)
  - 30s (33-47)
- Partición aleatoria en tres datasets:
  - Training
  - Early Bird (10%)
  - Testing (+20%)
- Incorporación de casos especiales
  - Predadores sexuales (0.0012%)
  - Conversaciones sobre sexo entre adultos

Estadísticas básicas del dataset

IDIOMA EDAD SEXO			NÚMERO DE AUTORES		
			TRAINING	EARLY BIRDS	TEST
EN	10s	HOMBRE	8.600	740	888
		MUJER	8.600	740	888
	20s	HOMBRE	(72) 42,828	3.840	(32) 4,576
		MUJER	(25) 42,875	3.840	(10) 4,576
	30s	HOMBRE	(92) 66,708	6.020	(40) 7,184
		MUJER	66.800	6.020	7.224
∑			***236.600***	***21.200***	***25.440***
ES	10s	HOMBRE	1.250	120	144
		MUJER	1.250	120	144
	20s	HOMBRE	21.300	1.920	2.304
		MUJER	21.300	1.920	2.304
	30s	HOMBRE	15.400	1.360	1.632
		MUJER	15.400	1.360	1.632
∑			***75.900***	***6.800***	***8.160***

Predadores sexuales
Conversaciones sobre sexo entre adultos

Formato de ficheros

- Una vez descomprimidos el dataset de training y el de test, la estructura es la siguiente:
- Un fichero de verdad: truth-en.txt / truth-es.txt con el formato:
  - HASH-MD5:::gender:::age
- Una carpeta por idioma: en / es con un fichero xml por cada autor:
  - En training [HASH-MD5]_[lang]_[age]_[gender].xml
  - En test lang, age y gender están substituidos por XX
- El formato del fichero xml se muestra a la derecha

```
<author lang="lang_code" gender="gender_code" age_group="age_group">
        <conversations count="number_of_conversations_in_file">
                <conversation id="UUID">
                        [Original HTML Content of the conversation]
                </conversation>

                <conversation id="UUID">
                        [Original HTML Content of the conversation]
                </conversation>

                ....

        </conversations>
</author>
```

Estadísticas básicas del dataset

- ¿Qué nos interesa al explorar?
  - Número de documentos por clase
  - Número de palabras por documento / clase
  - Distribuciones de palabras/documentos/autores por /documento/autor/clase...
  - Palabras ¿muy extrañas?, ¿muy frecuentes?, ¿muy comunes?...
  - ...

Material de trabajo

- Script Stats:
  - https://github.com/autoritas/RD-Lab/blob/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/src/Stats.zip
- Ficheros de distribución
  - https://github.com/autoritas/RD-Lab/tree/master/doc/projects/Text%20Mining%20II%20-%20Big%20Data/data
- Una ayuda en R:
  - data <- read.csv("PAN-AP-2013-test-es-distribution.csv")
  - pergender <- split(data, data$Gender)
  - perage <- split(data, data$Age)
  - pergenderandage <- split(data, list(data$Gender, data$Age))