# Text mining

## Por tus tweets te definirán

Toni Avia
José Antonio Cano
Josep Carrasco
José Antonio Miras (Pepe)

# ¿De qué datos disponemos?

Disponemos de los siguientes datos.

- 100 tweets de 2800 usuarios anonimizados en formato XML para training
- Dataset con el ID del usuario más su género y variedad de castellano para training
- 100 tweets de 1400 usuarios anonimizados para test
- Dataset con el ID del usuario más su género y variedad de castellano para test

# Exploración de datos

Para poder trabajar con los datos, en primer lugar tenemos que realizar un tratamiento de los 2800 xml para train y los 1400 xml para test, de manera que consigamos un dataset con una línea por autor.

El set de datos se basa en una **bolsa de palabras**

vocabulary = [si,q,gracias,vía,hoy,ser,día,mejor,bien,así]

bow_*** = [[female,id,0,7,6,4...]...]

# ¿cueces o enriqueces?
# Siempre hay que enriquecer

A las variables bow... podemos añadir

- Longitud media de los tweets 2º derivada (máximo, mínimo, ¿cuantiles?..)
  - Por número de caracteres por tweet
  - Por número de palabras por tweet
  - Por número de frases por tweet (identificando por puntos)
- Se considera como punto de partida que la longitud del tweet sí puede estar relacionada con el género

# ¿cueces o enriqueces? Menciones

A las variables bow... podemos añadir

- Número de menciones por el autor a otros usuarios
  - Puede que dependiendo del género se hagan más o menos menciones
- Estudio del género de los influencers que siguen las menciones del autor
  - Construir un dataset de los influencers con su género
- Las relaciones con otros usuario pueden ser relevantes para la determinación del género

# ¿cueces o enriqueces? Hashtags y emojis

A las variables bow... podemos añadir

- Número de emojis 😁😂
- Número de hashtag por autor

# ¿cueces o enriqueces?
# Tipos Palabras

A las variables bow... podemos añadir

- Contador por tipo de palabras (verbos, nombre, determinantes...)
- Identificar verbos en primera persona
- Con los dataset de entrenamiento, construir una bolsa de palabras por género

# ¿cueces o enriqueces?
# URL

A las variables bow... podemos añadir

- Extensión del dominio de las urls usadas en los twitts (.es .co... )
  - una variable con la cantidad de veces que aparece cada extensión en los tweets del autor
- Tener una bolsa de palabras de los temas más comunes para analizar las urls que se están compartiendo
- Puede ser una propiedad muy interesante las url que menciona el autor según su género

# ¿cueces o enriqueces?
# Siempre hay que enriquecer

A las variables bow... podemos añadir

- Longitud media de los tweets 2º derivada (máximo, mínimo, ¿cuantiles?..)
  - Por número de caracteres por tweet
  - Por número de palabras por tweet
  - Por número de frases por tweet (identificando por puntos)
- Se considera como punto de partida que la longitud del tweet sí puede estar relacionada con la variedad

# ¿cueces o enriqueces?
# Menciones

A las variables bow... podemos añadir

- Número de menciones por el autor a otros usuarios
  - localización del perfil de la persona con la que el autor se relaciona
  - y localización del perfil de los amigos del mencionado (2º nivel de relación)
- Estudio de la variedad de los influencers que siguen las menciones del autor
  - Hay que construir un dataset con influencers de cada país
- Las relaciones con otros usuario pueden ser relevantes para la determinación de la variedad

# ¿cueces o enriqueces? Hashtags y emojis

A las variables bow... podemos añadir

- Contenido de hashtags del autor
  - Comparar los hashtags con las tendencias de cada país
- Podría haber un problema dado que los tweets no son actuales
- Número de emojis

# ¿cueces o enriqueces?
# Tipos de Palabras

A las variables bow... podemos añadir

- Contador por tipo de palabras (verbos, nombre, determinantes...)
- Identificar verbos en primera persona
- Con los dataset de entrenamiento, construir una bolsa de palabras por variedad

# ¿cueces o enriqueces?
# URL

A las variables bow... podemos añadir

- Extensión del dominio de las urls usadas en los twitts (.es .co... )
  - una variable con la cantidad de veces que aparece cada extensión en los tweets del autor
- Tener una bolsa de palabras de los periódicos digitales, para analizar las urls que se están compartiendo
- Puede ser una propiedad muy interesante las url que menciona el autor según su variedad de lenguaje

# Medidas Accuracy sin tocar el dataset

| | n=50 | | | | n=100 | | | |
|---|---|---|---|---|---|---|---|---|
| | GÉNERO | | VARIEDAD | | GÉNERO | | VARIEDAD | |
| MODELO | train | test | train | test | train | test | train | test |
| SVM linear | **0.6553** | 0.6657 | 0.3550 | **0.3764** | **0.6542** | **0.6779** | 0.4675 | **0.4893** |
| KNN | 0.6092 | 0.5893 | 0.2571 | 0.2807 | 0.5861 | 0.6114 | 0.3067 | 0.3414 |
| Random F. | **0.6621** | **0.6664** | **0.3596** | **0.3600** | **0.6632** | 0.6714 | **0.4764** | **0.5043** |
| Naive Bayes | 0.6028 | 0.6364 | 0.2392 | 0.2814 | 0.6171 | 0.6257 | 0.3221 | 0.3443 |
| LDA | 0.6473 | **0.6721** | **0.3626** | 0.3586 | 0.6443 | **0.6743** | **0.4710** | 0.4693 |
| Red Neuronal | 0.6496 | 0.6600 | 0.2803 | 0.2829 | 0.6325 | 0.6629 | 0.3710 | 0.3843 |

# GRACIAS POR SU ATENCIÓN