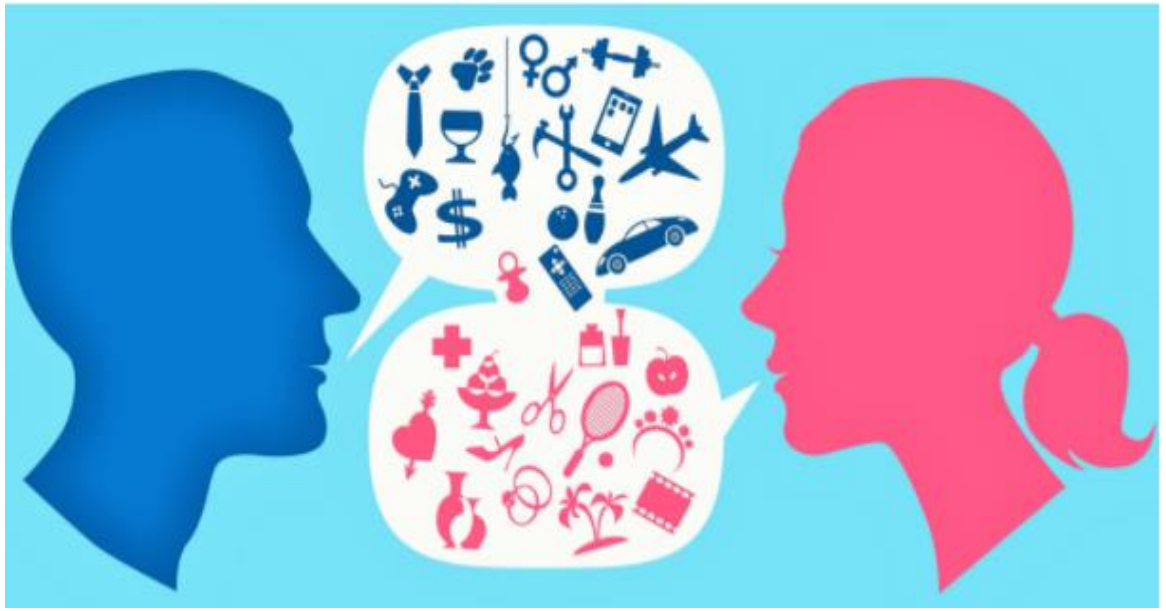
Text Mining en Social Media
Máster Big Data Analytics

Introducción

- A vueltas con el código y herramientas de desarrollo
  - UT-8
  - Library stringi
  - Stri_trans_general(corpus.raw, “Latin-ASCII”)
  - Instalar paquetes para que funcione el entorno.
  - Transformación de tipos
  - ¿Será Windows?
- Análisis de los datos y posibles estrategias para los dos casos.
  - ¿Cómo lo abordamos?

Genere (female, male)

- Planteamiento clase con dos valores posibles
- Caso base
  - Preprocesado: minúsculas, eliminación de números, palabras vacías..
  - Funciones:
    - GenerateVocabulary
    - GenerateBow
  - Vocabulario: 1000 palabras
  - Accuracy: 0,6643
- Preprocesado añadido
  - UT-8
  - Quitar acentos
- Modelo SVM Support Vector Machine



- Estrategia1: Genere female
  - Vocabulario 1003 palabras
  - Función GenerateVocabulariyGenero(…., genero)
  - Accuracy: 0,67
- Estrategia 2: Genere male
  - Vocabulario 1000 palabras
  - Función GenerateVocabulariyGenero(…., genero)
  - Accuracy: 0,6821
- Estrategia 3: Unión de diferencias
  - Vocabulario: 379 palabras
  - Accuracy: 0,6957
- Estrategia 4: Anterior más la intersección de vocabularios
  - Vocabulario: 1191 palabras
  - Accuracy: 0,6907
- Estrategia 5: Vocabulario mejor (estrategia 3) más lista de palabras
  - Vocabulario: 401 palabras
  - Función: add.unique
  - Accuracy: 0,6979
- el, la, este, poco, grande, excelente, soy, somos, voy, hago, lindo, dios, pueblo, seguidores, trabajar, facebook, hijo, no, pero, ni, mal, mierda, not, para, con, !!!, !!, :), :(

Hombres	Mujeres
Determinantes	Pronombres
Adjetivos	Negaciones
	Verbos presente

Variety (colombia, argentina, spain, venezuela, peru, chile, mexico)

- Planteamiento clase con siete valores posibles
- Caso base
  - Preprocesado: minúsculas, eliminación de números, palabras vacías..
  - Funciones:
    - GenerateVocabularyVariedad(….., variedad)
    - GenerateBow
  - Vocabulario: 1000 palabras
  - Accuracy: 0,7721
- Preprocesado añadido
  - UT-8
  - Quitar acentos
- Modelo
  - SVM
  - RandomForest

- Estrategia 1
  - Palabras propias de cada país
    - 7 vocabularios
  - Palabras exclusivas de cada país
    - País x – diferencias con el resto.
  - Hacer la unión de los anteriores
  - Accuracy: 0,8393

- Estrategia 2
  - Mismo diccionario que estrategia anterior pero con randomForest (50 árboles)
  - Accuracy: 0,8507

- Estrategia 3
  - Palabras propias de cada país
    - 7 vocabularios
  - De cada diccionario nos quedamos con las 100 palabras más frecuentes, y generamos un nuevo diccionario que tendrá 700 palabras.
  - Accuracy: 0,7271

Conclusiones

- Planificación optimista
  - Errores en código y herramientas.
  - Tiempo de procesado
- No por más palabras en la bolsa mejor resultado.
- Nos ha faltado probar con más modelos y realizar presentaciones gráficas para ver datos anómalos por ejemplo.
- Añadir más variables explicativas con importancia para el modelo. Por ejemplo, hemos visto que las mujeres utilizan más emoticonos, o sacar la media de la longitud de los tuits de cada autor, seleccionar palabras determinantes de cada país (ahorita, pibe, boludo,…)