Práctica VII

Aspectos de personalidad

Especificaciones

- En equipo de 3 a 4 personas realice lo siguiente:
  - Cargue el corpus de entrenamiento del concurso pan15
  - Con el corpus de entrenamiento cree un conjunto de desarrollo utilizando el **80% de los datos para entrenamiento** y el **20% restante para pruebas**. Es importante que los datos se revuelvan (shuffle) antes de realizar la separación y que asigne una semilla fija mediante la variable **random_state = 0**
  - Las tareas que se deben resolver son la clasificación de género y la clasificación de edad, para cada tarea debe hacer lo siguiente:
    - Utilizando el conjunto de entrenamiento (80% de los datos) realice las siguientes pruebas
      - Aplique el algoritmo de LSA para crear representaciones vectoriales del texto con distinto número de tópicos
      - Proponga distintos conjuntos de características como las utilizadas en el ejemplo base de la práctica (debe ser al menos uno distinto)
      - Una la representación creada por LSA y los conjuntos de características propuestos
      - Utilice las características unidas para entrenar un modelo de regresión logística
      - Utilice el modelo entrenado para clasificar las instancias del conjunto de pruebas (20% de los datos)
      - Calcule la exactitud del modelo
    - Después de probar con diferente número de tópicos, conjuntos de características y unión de características, seleccione el modelo que mejor resultado obtuvo en el conjunto de pruebas

Especificaciones

- Cargue el conjunto de prueba del concurso pan15
- Utilizando los modelos seccionados en el proceso de entrenamiento realice la clasificación de género y edad de las instancias en el conjunto de pruebas

Evidencias

- Código fuente
- Reporte donde se incluya
  - Una tabla con los números de tópicos y características probadas en el conjunto de desarrollo y los valores de exactitud (accuracy), precisión, recall y F-measure obtenidos por prueba en cada tarea de clasificación
  - La configuración de tópicos y características que mejores resultados obtuvo en el conjunto de desarrollo en cada tarea de clasificación
  - Los valores de exactitud (accuracy), precisión, recall y F-measure obtenidos por los modelos en cada tarea de clasificación en el conjunto de prueba
  - Matriz de confusión de los valores predichos vs los reales en el conjunto de prueba de cada tarea de clasificación