Facultad de Ingeniería
departamento de informatica
Inteligencia artificial
Tarea - Clasificación: Bayes ingenuos y árboles de decisión

Parte 1 - 80%

Observación: Para solucionar esta práctica, solo se permitirán las bibliotecas numpy y pandas.

Esta práctica está relacionada con la etapa de clasificación del proceso KDD. Consiste en el procesamiento de un conjunto de tweets con el fin de identificar automáticamente al usuario que emitió el tweet mediante el clasificador Naive Bayes. Los datos a tratar se encuentran en el fichero tuits bayes.txt, que se puede descargar de Internet. Este archivo contiene 1.349 tweets en el siguiente formato:

"id de estado", "nombre de pantalla", "texto"

Tienes que realizar las siguientes tareas:

Tareas a realizar

1. Cargue el archivo en un marco de datos.

2. Realice el proceso de limpieza de la siguiente manera:

   (a) Eliminar enlaces URL (b)

   Eliminar palabras cuya longitud sea igual o inferior a 2 caracteres (c) Eliminar cadenas de

   caracteres que carezcan completamente de caracteres alfabéticos
   caracteres

(d) Quitar tildes (e) Eliminar las

entradas de aquellos usuarios que tengan 5 o menos tweets

3. Para cada usuario, extraiga el 80% de sus tweets para usarlos como datos de entrenamiento y reserve el 20% restante para usarlos como datos de prueba.
4. Calcula las probabilidades previas de que un usuario publique un tweet.
5. Entrene el sistema (esto equivale a calcular la probabilidad condicional).

ciudades)

6. Pruebe el sistema y cree una tabla que resuma todas las pruebas. ¿Qué puedes decir sobre el desempeño del clasificador Naive Bayes?

Parte 2 - 20%

Elaborar una presentación sobre Árboles de Decisión (DT). Se deben considerar aspectos como la definición, tipos de DT, ejemplos y aplicaciones.