Reconocimiento de Patrones y Aprendizaje Automatizado

Práctica 4. Teoría de decisiones bayesiana

Profesores: Miguel Daniel Garrido Reyes
Ayudante: Melissa Vázquez González
Ayud. Lab.: Luis Emilio González Covarrubias

23 de febrero de 2024

Proporcionar un enlace a una carpeta en un repositorio de GitHub que contenga el cuaderno (notebook) donde se encuentran las respuestas al ejercicio siguiente.

Ejercicio

Dentro del archivo "spam_ham_dataset.csv" se encuentran datos recopilados sobre correos electronicos y la clasificacion de los mismos, si estos son o no son spam, se desea crear un clasificador que utilice el algoritmo de Naive Bayes (Bayes Ingenuo) para predecir si un correo es spam o no, resolver los siguientes incisos:

- Crear una función de Python llamada "preprocess_text" que reciba una cadena de texto, a la cual aplique:

  1. Tokenización
  2. Lower case
  3. Eliminar caracteres especiales y numeros
  4. Eliminar stop words
  5. Quitar cadenas vacias
  6. Aplicar lematización

  y regrese la lista con los tokens de esta cadena
  Ejemplo: Para la cadena:
  "Congratulations! You've won a free trip to Hawaii. Click here to claim your prize!"

1. Al aplicar tokenización:
   'Congratulations', '!', 'You', "'ve", 'won', 'a', 'free', 'trip', 'to', 'Hawaii', '.', 'Click', 'here', 'to', 'claim', 'your', 'prize', '!'
2. Lower case:
   'congratulations', '!', 'you', "'ve", 'won', 'a', 'free', 'trip', 'to', 'hawaii', '.', 'click', 'here', 'to', 'claim', 'your', 'prize', '!'
3. Elimiar caracteres especiales y numeros:
   'congratulations', '', 'you', 've', 'won', 'a', 'free', 'trip', 'to', 'hawaii', '', 'click', 'here', 'to', 'claim', 'your', 'prize', ''
4. Elimiar stop words:
   'congratulations', '', 'free', 'trip', 'hawaii', '', 'click', 'claim', 'prize', ''
5. Quitar cadenas vacias:
   'congratulations', 'free', 'trip', 'hawaii', 'click', 'claim', 'prize'
6. Aplicar lematización:
   'congratulation', 'free', 'trip', 'hawaii', 'click', 'claim', 'prize'

   La funcion debe retornar:
   'congratulation', 'free', 'trip', 'hawaii', 'click', 'claim', 'prize'

- Utiliza el notebook **2024_02_22_RPAA_Naive_Bayes.ipynb** para implementar el clasificador de correos electronicos de spam o no spam, aplicando la funcion "preprocess_text" del inciso anterior, para obtener el conjunto de palabras con el que se calcularan las probabilidades

- Utilizando las metricas vistas en clase, evaluar el modelo generado y determinar si mejora su desempeño respecto al modelo visto en clase (sin aplicar procesamiento de lenguaje natural)