DCC-Universidad de Chile
CC5509 Reconocimiento de Patrones
Tarea 3: Clasificación de texto (recuperativa).

Profs. José M. Saavedra - Mauricio Cerda
Ayudante: Camila Álvarez

15 de Noviembre de 2017

1 Descripción y entrenamiento

En esta tarea se pide implementar un clasificador Bayesiano naive, para clasificar artículos de texto de la base de datos REUTERS-21578. Dichos artículos estan previamente manualmente clasificados en 90 clases. Para esta tarea se usará una version reducida de la base de datos ("Apte split"), y trabajar solamente con las 10 categorias más frecuentes: Earn, Acquisition, Money-fx, Grain, Crude, Trade, Interest, Ship, Wheat y Corn. Esta base reducida la pueden encontrar en: *http://disi.unitn.it/moschitti/corpora.htm.*

En este trabajo Ud. deberá entrenar su clasificador, y evaluarlo siguiendo la partición de Apte. En el entrenamiento, comience por:

- Eliminar signos de puntuación.
- Eliminar stopwords
- Construir las tablas de frecuencia por clase para el diccionario

Al construir las tablas de frecuencia utilice el suavizado de Laplace para evitar los problemas de palabras que no aparecen en una clase.

2 Evaluación

Evalue el rendimiento del clasificador, seleccionando métricas apropiadas. Incluya tambien una matriz de confusión. Seleccione algunos ejemplos errores de clasificación y describa el tipo de problemas encontrado.

3 Informe

Se debe presentar un informe en formato tipo *paper* , el que debe incluir:

1. Introducción
2. Desarrollo. Aquí describa todo lo referente a cómo se abordó problema y presentar una descripción detallada de sus programas.
3. Evaluación de Resultados. Analizar cómo afecta el valor de cada parámetro en el resultado.
4. Conclusiones

4 Datos importantes

- El trabajo se realiza en forma individual.
- Se dispondrá de 2 semanas para realizar esta tarea, la entrega es mediante u-cursos.
- No se aceptan atrasos.