Práctico 3: Clasificación supervisada de documentos

Ejercicio 1

Se tiene un conjunto de documentos que se quieren clasificar. Para esto, se utiliza un método de clasificación sobre el corpus de entrenamiento (generado eligiendo documentos del corpus original, al azar, hasta un 80% del total). Los documentos pueden clasificarse en: Nacional, Internacional, Deportes, y cada uno tiene solamente una categoría posible.

Al evaluar sobre el corpus de evaluación, se obtienen los siguientes valores:

Documento	Clase original	Predicción
d1	Nacional	Internacional
d2	Deportes	Internacional
d3	Nacional	Nacional
d4	Deportes	Nacional
d5	Deportes	Deportes
d6	Deportes	Deportes
d7	Deportes	Nacional
d8	Internacional	Internacional
d9	Internacional	Deportes
d10	Internacional	Deportes
d11	Deportes	Deportes
d12	Nacional	Nacional
d13	Deportes	Deportes
d14	Deportes	Deportes
d15	Internacional	Nacional
d16	Internacional	Nacional
d17	Internacional	Internacional
d18	Internacional	Internacional
d19	Internacional	Internacional
d20	Deportes	Nacional

a) Construya la matriz de confusión
b) Calcule la *accuracy* del clasificador
c) Para cada clase, cuente la cantidad de TP, TN, FP, FN y calcule Precisión, Recall y Medida-F.
d) Calcule macro-Precisión, macro-Recall y macro-F.

Ejercicio 2

Se desea utilizar un clasificador Naïve Bayes con atributos tipo bag of words para analizar reviews de películas. Suponga que tiene el siguiente corpus:

Review	Clasificación
Buenísima. Entrenida y bien lograda.	+
Escenas traídas de los pelos. No la recomiendo.	-
No me gustó la película.	-
Horrible. Me aburrí como un hongo.	-
Muy buena la película. La super recomiendo.	+
Muy linda película.	+
Me gustó. La recomiendo totalmente.	+
No la recomiendo. Es un divague.	-
Una historia que es un mamarracho.	-

El preprocesamiento incluye pasar a minúsculas, tokenización simple separando por espacios, y eliminación de stopwords y de símbolos de puntuación.

¿Cuál sería el resultado del clasificador para los siguientes ejemplos?

> *Muy buena, la recomiendo.*
> *jaaa un mamarracho*
> *Una verdadera pérdida de tiempo. Linda para dormir la siesta.*
> *Fui con mi familia, pasamos genial*

Ejercicio 3

Considere el problema de identificar los siguientes elementos en un texto: predicados de opinión (elemento que indica la presencia de una opinión) y fuentes de opinión (entidad que expresa la opinión). Se modela el problema como un caso de clasificación secuencial con esquema BIO. Se utilizan las etiquetas B-fuente, I-fuente, B-predicado, I-predicado, O.

Dados los siguientes ejemplos:

> *El presidente de Ancap, Raul Sendic, dijo a Montevideo Portal, que le "gustaría" seguir al frente del organismo, pero "si el presidente cree que tengo que ir al ministerio de Industria, no voy a tener dificultad en aceptar".*
>
> *"La garrafa está costando 100 pesos menos que el año pasado", agregó Sendic.*
>
> *Según el jerarca, los costos se mantendrán estables por el resto del año.*

a) Muestre la clase asociada a cada token para cada oración. Los predicados son: *dijo, gustaría, cree, agregó, según*. Las fuentes son: *El presidente de Ancap, Raúl Sendic*; *el presidente*; *Sendic*; *el jerarca*.
b) ¿Cómo mediría la performance de un sistema automático para este problema?

Ejercicio 4

Resuelva el notebook publicado en este link. Utilice los archivos que se publicaron en el módulo "Clasificación supervisada de documentos" del EVA.