Práctico 3: Clasificación / Modelos de Lenguaje / HMM

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) Cuente la cantidad de TP, TN, FP, FN y calcule Precisión, Recall y Medida-F, convirtiendo el problema en tres problemas *one-vs-all.*
d) Calcule macro-Precisión, macro-Recall y macro-F.

Ejercicio 2

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

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

Un sistema de reconocimiento de texto ofrece las siguientes opciones para un texto escrito a mano, con sus respectivas probabilidades de certeza:

“Me pegué en la cabeza .” (P=0.30)
“Me pegué en la oreja .” (P=0.10)
“Me pegué en las cabeza .” (P=0.50)

Dispone de un corpus de 192685 palabras en 6030 oraciones y un vocabulario de 1320 palabras, con los siguientes conteos de ocurrencias de palabras y bigramas:

Me (5), pegué (0), en (4340), la (6412), cabeza (28), las (1832), oreja (0), “.” (5866), “<s>Me” (5), “Me pegué” (0), “pegué en” (0), “en la” (703), “la cabeza” (13), “en las” (190), “las cabeza” (0), “la oreja” (0), “cabeza .” (3), “oreja .”(0), “. </s>” (5842).

a) Utilice un modelo de bigramas para estimar la probabilidad de cada oración, utilizando un estimador de máxima verosimilitud en el corpus (utilice una técnica de suavizado adecuada).

b) Calcule, utilizando la Regla de Bayes, la mejor oración candidata de las tres presentadas por el sistema.

Ejercicio 4

Se desea aplicar un modelo HMM para resolver la asignación de categorías léxicas a un texto. Del análisis del corpus de entrenamiento, se estiman las siguiente probabilidades:

Probabilidades de transición $P(tag_j | tag_i)$

	j:	**0**	**1**	**2**	**3**	**4**
i		</s>	DET	V	ADJ	N
0	<s>	0	0,32	0,03	0,01	0,12
1	DET	0	0,03	0	0,09	0,8
2	V	0,05	0,25	0,15	0,05	0,06
3	ADJ	0,19	0,02	0,05	0,03	0,21
4	N	0,53	0,01	0,08	0,15	0,02

Probabilidades de emisión diferentes de 0:

P(*unas* | DET) =0,0020
P(*unas* | V) = 0,000045
P(*blancas* |ADJ) = 0,00029
P(*blancas* | N) = 0,000023
P(*velas* | N) = 0,000029
P(*velas*|V)=0,000045

a) Calcule la secuencia de tags más probables para la frase “Unas blancas velas”, y su probabilidad.

b) Calcule la probabilidad de la frase.

Ejercicio 5 (Diciembre de 2010)

Se desea aplicar un modelo HMM para resolver la asignación de categorías léxicas a un texto. Del análisis del corpus de entrenamiento, se estiman las siguiente probabilidades:

Probabilidades de transición $P(tag_j | tag_i)$

	j:	**0**	**1**	**2**	**3**	**4**
i		</s>	PREP	DET	V	NN
0	<s>	0	0,2	0,05	0,1	0,1
1	PREP	0	0	0,5	0,1	0,1
2	DET	0	0	0	0	0,1
3	V	0,3	0,1	0,1	0,2	0,2
4	NN	0,2	0,2	0,1	0,1	0,3

Probabilidades de emisión diferentes de 0:

P(*sobre* | PREP) =0,07
P(*sobre* | V) = 0,0012
P(*sobre* | N) = 0,00062
P(*el* | DET) = 0,4

a) Aplique el algoritmo de Viterbi para obtener la secuencia de tags más probable para la expresión “*sobre el sobre*”

b) ¿Cuál es la diferencia entre el algoritmo Viterbi y el algoritmo Forward para HMMs?

Ejercicio 6

Considere el problema de identificar el alcance de una especulación en un texto. Para esto, se modela el problema como un caso de clasificación secuencial con esquema FOL. Así, dada una oración («Los resultados indican que existe una relación entre A y B») y una marca de especulación («indican”) , se tiene cada token de la oración anotado con una clase F si es el comienzo del alcance y con L si es el final, anotando el resto de los tokens con la clase O.

a) Muestre las clase asociada a cada token de la oración «Los resultados indican que existe una relación entre A y B»
b) ¿Cómo es la relación entre Accuracy, Precision y Recall para este problema?

Ejercicio 7

Se desea obtener el POS tag de cada palabra de un documento solamente teniendo en cuenta el POS de la palabra anterior y la siguiente. Por lo tanto, el problema puede modelarse como el de obtener el POS que maximiza la siguiente probabilidad:

P(c|Ant=c1, Sig=c2)

siendo Ant y Sig variables aleatorias que representan a los eventos “La palabra anterior tiene clase c1” y “La palabra siguiente tiene clase c2”.

a) Indique cuál sería el clasificador construido utilizando el método Naïve Bayes
b) Suponga que se cuenta en el corpus de entrenamiento y se obtienen los siguientes valores (c(DET,NOM) indica la cantidad de bigramas DET NOM)):

c(<s>,DET) = 18
c(<s>, NOM) = 9
c(<s>, ADJ) = 3

c(DET, NOM) = 21
c(DET, ADJ) = 18
c(DET, DET) = 0
c(DET,</s>) = 0

c(NOM, NOM) = 9
c(NOM, ADJ) = 9
c(NOM, DET) = 21
c(NOM,</s>) =21

c(ADJ, NOM) = 21
c(ADJ, ADJ) = 0
c(ADJ, DET) = 0
c(ADJ,</s>) = 9

Calcule la clase para una palabra que aparece entre un determinante y un nombre, si suponemos que los únicos POS tags posibles son NOM, ADJ, DET. Justifique.

Ejercicio 8

Indique cómo se resolvería el problema anterior, si lo que se quiere es calcular la probabilidad de cada clase utilizando un modelo de Entropía Máxima, considerando como atributos, nuevamente, la clase anterior y siguiente de cada palabra a clasificar.