Ejercicio 1: Escritura predictiva y n-gramas

- Considera el siguiente corpus de frases para un sistema de escritura predictiva de un dispositivo móvil

<s> Yo voy bien </s>

<s> Voy a casa </s>

<s> Yo voy a comer a casa </s>

<s> Voy de cráneo </s>

<s> Voy a estudiar </s>

<s> Ya voy yo </s>

<s> Vamos a comer </s>

<s> Quiero comer </s>

- Utilizando bigramas y considerando las palabras tal y como aparecen (sin coger su raíz o su lexema) indica la siguiente palabra para las siguientes frases
  - <s> Ya voy
  - <s> Voy

Ejercicio 1: Escritura predictiva y n-gramas

<s> Yo voy bien </s>

<s> Voy a casa </s>

<s> Yo voy a comer a casa </s>

<s> Voy de cráneo </s>

<s> Voy a estudiar </s>

<s> Ya voy yo </s>

<s> Vamos a comer </s>

<s> Quiero comer </s>

- Usando el modelo de bigramas la siguiente palabra para las frases “<s> Ya voy” y “<s> Voy” será la misma
  - En el modelo de bigramas n=2, y la última palabra solamente depende de las n-1 palabras anteriores
  - Es decir debemos encontrar la palabra Y que maximice **P(Y | voy)**
- El considerar las palabras tal y como aparecen hace que no consideremos las distintas apariciones del verbo ir (voy, vamos) con distintos tiempos verbales como si fueran la misma palabra.

Ejercicio 1: Escritura predictiva y n-gramas

<s> Yo voy bien </s>

<s> Voy a casa </s>

<s> Yo voy a comer a casa </s>

<s> Voy de cráneo </s>

<s> Voy a estudiar </s>

<s> Ya voy yo </s>

<s> Vamos a comer </s>

<s> Quiero comer </s>

- Calculamos la probabilidad de los posibles P(Y | voy)
  - P(bien | voy) = 1/6
  - **P(a | voy) = 3/6**
  - P(de | voy) = 1/6
  - P(yo | voy) = 1/6

Para ambas frases la siguiente palabra sería “a”

“Ya voy a” y “Voy a”

Estimación optima por máxima verosimilitud

$$P(w_n|w_{n-1}) = \frac{frec(w_{n-1}, w_n)}{frec(w_{n-1})}$$

Ejercicio 1: Escritura predictiva y n-gramas

<s> Yo voy bien </s>

<s> Voy a casa </s>

<s> Yo voy a comer a casa </s>

<s> Voy de cráneo </s>

<s> Voy a estudiar </s>

<s> Ya voy yo </s>

<s> Vamos a comer </s>

<s> Quiero comer </s>

- Para el modelo de trigramas la probabilidad a maximizar es distinta en cada frase P(Y | <s>, voy) y P(Y | ya, voy)
  - P(Y | <s>, voy)
    - P(de | <s>, voy) = 1/3
    - **P(a | <s>, voy) = 2/3**

Las frases quedarían

"Ya voy yo" y "Voy a"

  - P(Y | ya, voy)
    - **P(yo | ya, voy) = 1/1**

Estimación optima por máxima verosimilitud

$$P(w_n|w_{n-1}) = \frac{frec(w_{n-1}, w_n)}{frec(w_{n-1})}$$

Ejercicio 1: Escritura predictiva y n-gramas

<s> Yo voy bien </s>

<s> Voy a casa </s>

<s> Yo voy a comer a casa </s>

<s> Voy de cráneo </s>

<s> Voy a estudiar </s>

<s> Ya voy yo </s>

<s> Vamos a comer </s>

<s> Quiero comer </s>

- Para determinar la frase que se generaría automáticamente con el modelo de bigramas vamos a ahorrarnos los cálculos
  - Al ser tan pocas frases se puede contar a ojo cuál es la opción más probable
    1. P(Y | <s>) → Y= Voy
    2. P(Y | voy) → Y= a
    3. P(Y | a) → Y= casa | comer
    4. P(Y | casa) → Y= </s>
    5. P(Y | comer) → Y= </s>

Hay dos frases equiprobables:

“Voy a casa”
“Voy a comer”

Ejercicio 2: Recuperación de información

- Considera la siguiente tabla de frecuencias de aparición de términos en varios documentos.

	A	B	C	D	E
naranja	30	0	30	10	0
limón	0	10	10	0	10
kiwi	0	30	10	30	10
fresa	0	20	10	0	10
manzana	0	10	0	0	10
pera	10	0	0	0	0
piña	10	0	0	0	10

Ejercicio 2: Recuperación de información

- Calcula el ranking de documentos que recuperaríamos usando la similitud del coseno para la consulta “naranja” y usando como representación de los términos del documento lo siguiente:
  - La presencia del término en el documento (variable binaria)
  - La frecuencia del término en el documento
  - El peso TF-IDF del término en el documento

Ejercicio 2: Rec. de información - binaria

	A	B	C	D	E
naranja	30	0	30	10	0
limón	0	10	10	0	10
kiwi	0	30	10	30	0
fresa	0	20	10	0	10
manzana	0	10	0	0	10
pera	10	0	0	0	0
piña	10	0	0	0	10

- Transformamos la tabla en binaria (presencia/ausencia) e ignoramos los documentos que no tienen el término naranja, porque nunca serán recuperados por nuestro buscador

	A	C	D
naranja	1	1	1
limón	0	1	0
kiwi	0	1	1
fresa	0	1	0
manzana	0	0	0
pera	1	0	0
piña	1	0	0

Ejercicio 2: Rec. de información - binaria

	Cons	A	C	D
naranja	1	1	1	1
limón	0	0	1	0
kiwi	0	0	1	1
fresa	0	0	1	0
manzana	0	0	0	0
pera	0	1	0	0
piña	0	1	0	0

Similitud del coseno entre la consulta c y el documento d_j

$$cos(c, d_j) = \frac{c \cdot d_j}{|c||d_j|} = \frac{\sum_{i=1}^{M} c_i \cdot w_{i,j}}{\sqrt{\sum_{i=1}^{M} c_i^2}\sqrt{\sum_{i=1}^{M} w_{i,j}^2}}$$

- Debemos calcular la similitud del coseno entre la consulta y cada documento
  - Mostramos los cálculos que se realizan para entender cómo funciona

	A	C	D
Prod. Esc. $c \cdot d_j$	1	1	1

	c	A	C	D
modulo	1	$\sqrt{3}$	2	$\sqrt{2}$

	A	C	D
$\boldsymbol{cos(c, d_j)}$	$1/\sqrt{3}$	1/2	$1/\sqrt{2}$

El ranking de relevancia

D > A > C

Los documentos son más relevantes en tanto en cuanto aparecen menos términos en ellos

Ejercicio 2: Rec. de información - frecuencias

- En este caso, usamos la tabla de frecuencias sin transformarla

	Cons	A	C	D
naranja	1	30	30	10
limón	0	0	10	0
kiwi	0	0	10	30
fresa	0	0	10	0
manzana	0	0	0	0
pera	0	10	0	0
piña	0	10	0	0

Similitud del coseno entre la consulta c y el documento d_j

$$cos(c, d_j) = \frac{c \cdot d_j}{|c||d_j|} = \frac{\sum_{i=1}^{M} c_i \cdot w_{i,j}}{\sqrt{\sum_{i=1}^{M} c_i^2}\sqrt{\sum_{i=1}^{M} w_{i,j}^2}}$$

- Debemos calcular la similitud del coseno entre la consulta y cada documento
  - Mostramos los cálculos que se realizan para entender cómo funciona

	A	C	D
Prod. Esc. $c \cdot d_j$	30	30	10

	c	A	C	D
modulo	1	$\sqrt{1100}$	$\sqrt{1200}$	$\sqrt{1000}$

	A	C	D
$\boldsymbol{cos(c, d_j)}$	$30/\sqrt{1100}$	$30/\sqrt{1200}$	$10/\sqrt{1000}$

El ranking de relevancia

A > C > D

La frecuencia del término es se ve matizada por su frecuencia relativa en el documento

Ejercicio 2: Rec. de información – TF-IDF

	IDF	A	B	C	D	E
naranja	Log(5/4)	30	0	30	10	0
limón	Log(5/4)	0	10	10	0	10
kiwi	Log(5/4)	0	30	10	30	0
fresa	Log(5/4)	0	20	10	0	10
manzana	Log(5/3)	0	10	0	0	10
pera	Log(5/2)	10	0	0	0	0
piña	Log(5/3)	10	0	0	0	10

Calculamos la IDF de cada término en el corpus, que nos modera más su peso cuanto más popular es en el corpus

- Transformamos la tabla de frecuencias a TF-IDF con $w_{t,d} = tf_{t,d} \times \overbrace{\log\left(\frac{N}{df_t + 1}\right)}^{\text{IDF}}$
  - Solamente transformamos los documentos que nos interesan

	A	C	D
naranja	6.69	6.69	2.23
limón	0.00	2.23	0.00
kiwi	0.00	2.23	6.69
fresa	0.00	2.23	0.00
manzana	0.00	0.00	0.00
pera	9.16	0.00	0.00
piña	5.11	0.00	0.00

Ejercicio 2: Rec. de información – TF-IDF

	A	C	D
naranja	6.69	6.69	2.23
limón	0.00	2.23	0.00
kiwi	0.00	2.23	6.69
fresa	0.00	2.23	0.00
manzana	0.00	0.00	0.00
pera	9.16	0.00	0.00
piña	5.11	0.00	0.00

Similitud del coseno entre la consulta c y el documento d_j

$$cos(c, d_j) = \frac{c \cdot d_j}{|c||d_j|} = \frac{\sum_{i=1}^{M} c_i \cdot w_{i,j}}{\sqrt{\sum_{i=1}^{M} c_i^2}\sqrt{\sum_{i=1}^{M} w_{i,j}^2}}$$

- Debemos calcular la similitud del coseno entre la consulta y cada documento
  - Mostramos los cálculos que se realizan para entender cómo funciona

	A	C	D
Prod. Esc. $c \cdot d_j$	6.69	6.69	2.23

	c	A	C	D
modulo	1	12.44	7.73	7.06

	A	C	D
$\boldsymbol{cos(c, d_j)}$	6.69/12.44	6.69/7.73	2.23/7.06

El ranking de relevancia

C > A > D

La frecuencia del término es se ve matizada primero por su frecuencia en el corpus y luego por su frecuencia relativa en el documento

Ejercicio3:Análisis de sentimiento y Naive Bayes

- Haz un clasificador de sentimiento de críticas de restaurantes a partir de 500 mensajes positivos y 600 negativos y la frecuencia de aparición de términos en los mensajes que aparece en la tabla.
- Utiliza Naive Bayes para determinar el sentimiento de un mensaje que tenga los siguientes términos:
  - mejor, gustar, volver, caro
  - mejor, gustar, lamentable

Ejercicio3:Análisis de sentimiento y Naive Bayes

- **En la tabla se muestra el número de mensajes positivos y negativos que contienen dicho término.**

término	***mensajes positivos***	***mensajes negativos***
fabuloso	50	0
mejor	200	60
gustar	300	300
volver	200	200
caro	50	150
lamentable	0	30

Ejercicio3:Análisis de sentimiento y Naive Bayes

- Partimos de 500 mensajes positivos y 600 negativos

término	*mensajes positivos*	*mensajes negativos*
fabuloso	50	0
mejor	200	60
gustar	300	300
volver	200	200
caro	50	150
lamentable	0	30

$$P(x_k = x_{ki} \,|y = y_i) = \frac{frec(x_k = x_{ki} \wedge y = y_i)}{frec(y = y_i)}$$

Probabilidad condicionada de "volver" para cada clase

$$P(volver|positivo) = \frac{200}{500} \qquad P(\text{NO } volver|positivo) = \frac{300}{500}$$

$$P(volver|negativo) = \frac{200}{600} \qquad P(\text{NO } volver|negativo) = \frac{400}{600}$$

Prob. cond. de presencia de término

	positivo	negativo
fabuloso	0.1	0
mejor	0.4	0.1
gustar	0.6	0.5
volver	0.4	0.33
caro	0.1	0.25
lamentable	0	0.05

Prob. cond. de ausencia de término

	positivo	negativo
fabuloso	0.9	1
mejor	0.6	0.9
gustar	0.4	0.5
volver	0.6	0.67
caro	0.9	0.75
lamentable	1	0.95

Ejercicio3:Análisis de sentimiento y Naive Bayes

- **Mensaje: mejor, gustar, volver, caro**

$$P(y = y_i | \boldsymbol{x} = \boldsymbol{x_i}) \propto P(y = y_i) \prod_{k=1}^{d} P(x_k = x_{ki} \,| y = y_i)$$

	Prob. clase	no fabuloso	mejor	gustar	volver	caro	no lamentable
positivo	0.45	0.9	0.4	0.6	0.4	0.1	1
negativo	0.55	1	0.1	0.5	0.33	0.25	0.95

Para cada fila hacemos el producto de todas estas probabilidades y calculamos la verosimilitud de cada clase

Clase	verosimilitud
Positivo	**0.003927**
Negativo	0.002159

Es más verosímil que sea positivo

Ejercicio3:Análisis de sentimiento y Naive Bayes

- **Mensaje: mejor, gustar, lamentable**

$$P(y = y_i | \boldsymbol{x} = \boldsymbol{x_i}) \propto P(y = y_i) \prod_{k=1}^{d} P(x_k = x_{ki} | y = y_i)$$

Prob. de clase $P(y = y_i) = \frac{frec(y=y_i)}{N}$

	Prob. clase	no fabuloso	mejor	gustar	no volver	no caro	lamentable
positivo	0.45	0.9	0.4	0.6	0.6	0.9	0
negativo	0.55	1	0.1	0.5	0.67	0.75	0.05

Para cada fila hacemos el producto de todas estas probabilidades y calculamos la verosimilitud de cada clase

Clase	verosimilitud
Positivo	0
Negativo	**0.0006818**

Es más verosímil que sea negativo

OJO: *El valor de positivo se va a cero por "lamentable" para evitarlo habría que usar algún tipo de alisado*