Métodos Numéricos

Departamento de Computación, FCEyN, Universidad de Buenos Aires.

5 de mayo de 2016

Clase de hoy

- Clasificación de noticias
- Evaluación: Cross validation
- ¿Qué experimentar y cómo?
- Variantes para mostrar resultados

El problema

id	section	topic	text	title
3066966	Deportes	Deportes	“NUEVA YORK. - Novak . . . ”	”No es una utopía: . . . ”
3065926	“El Mundo”	Internacionales	“Los bomberos tratan de . . . “	”Declaran el estado de . . . ”
. . .	. . .	. . .	. . .	. . .
2496990	Tecnología	Tecnología	“La plataforma de video . . . ”	”La tiranía digital del . . . ”

El problema

id	section	topic	text	title
6431364	Procesados	????	“El presidente Mauricio . . . ”	”Panama Papers: la respuesta de Macri . . . ”

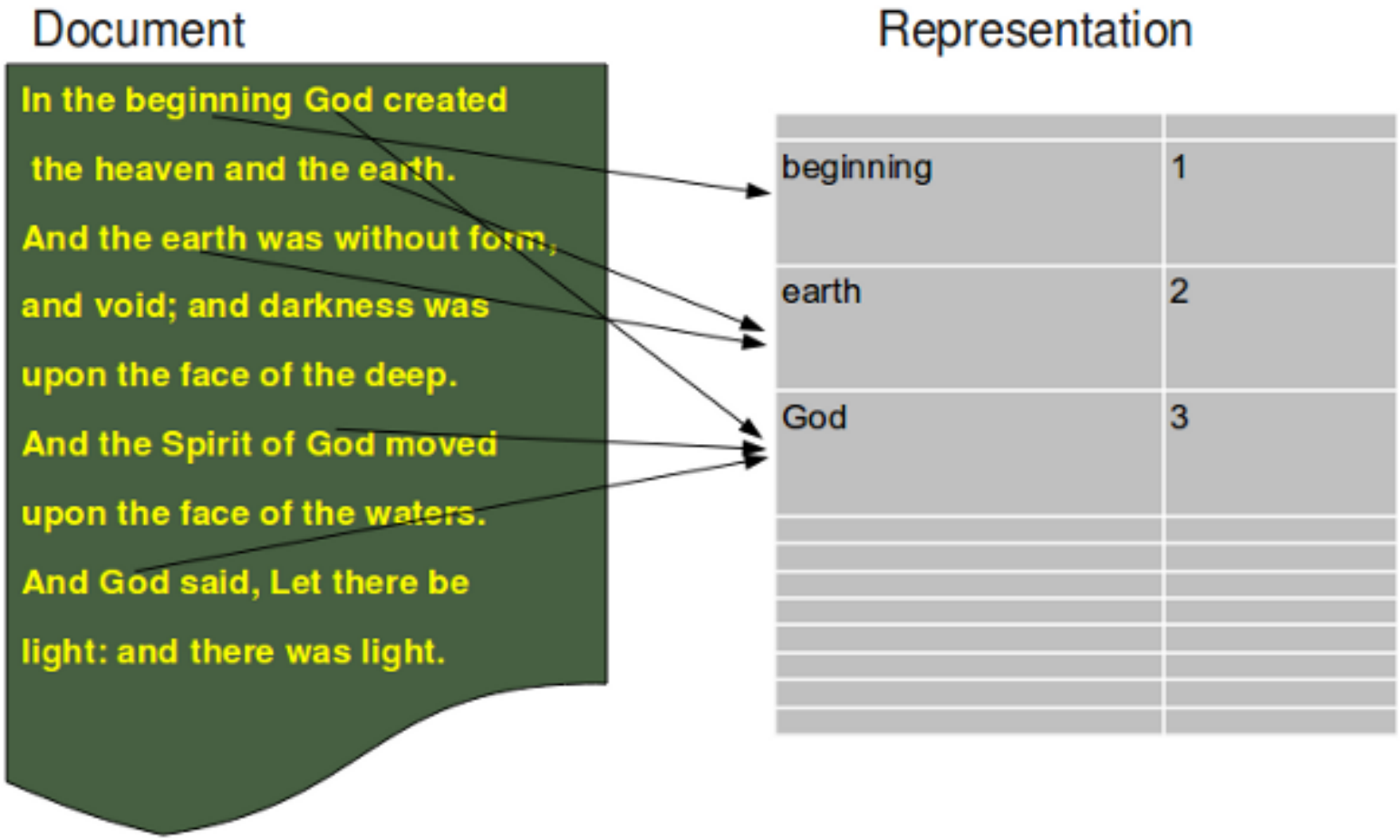
- En función del texto de las noticias decidir cuál categoría le corresponde
- Comparar con anteriores noticias usando:
  - Bag of words
  - N-grams
  - Term frequency–inverse document frequency (tf–idf)
  - Stemming

El problema

id	section	topic	text	title
6431364	Procesados	????	“El presidente Mauricio . . . ”	”Panama Papers: la respuesta de Macri . . . ”

- En función del texto de las noticias decidir cuál categoría le corresponde
- Comparar con anteriores noticias usando:
  - Bag of words
  - N-grams
  - Term frequency–inverse document frequency (tf–idf)
  - Stemming

Bag of words

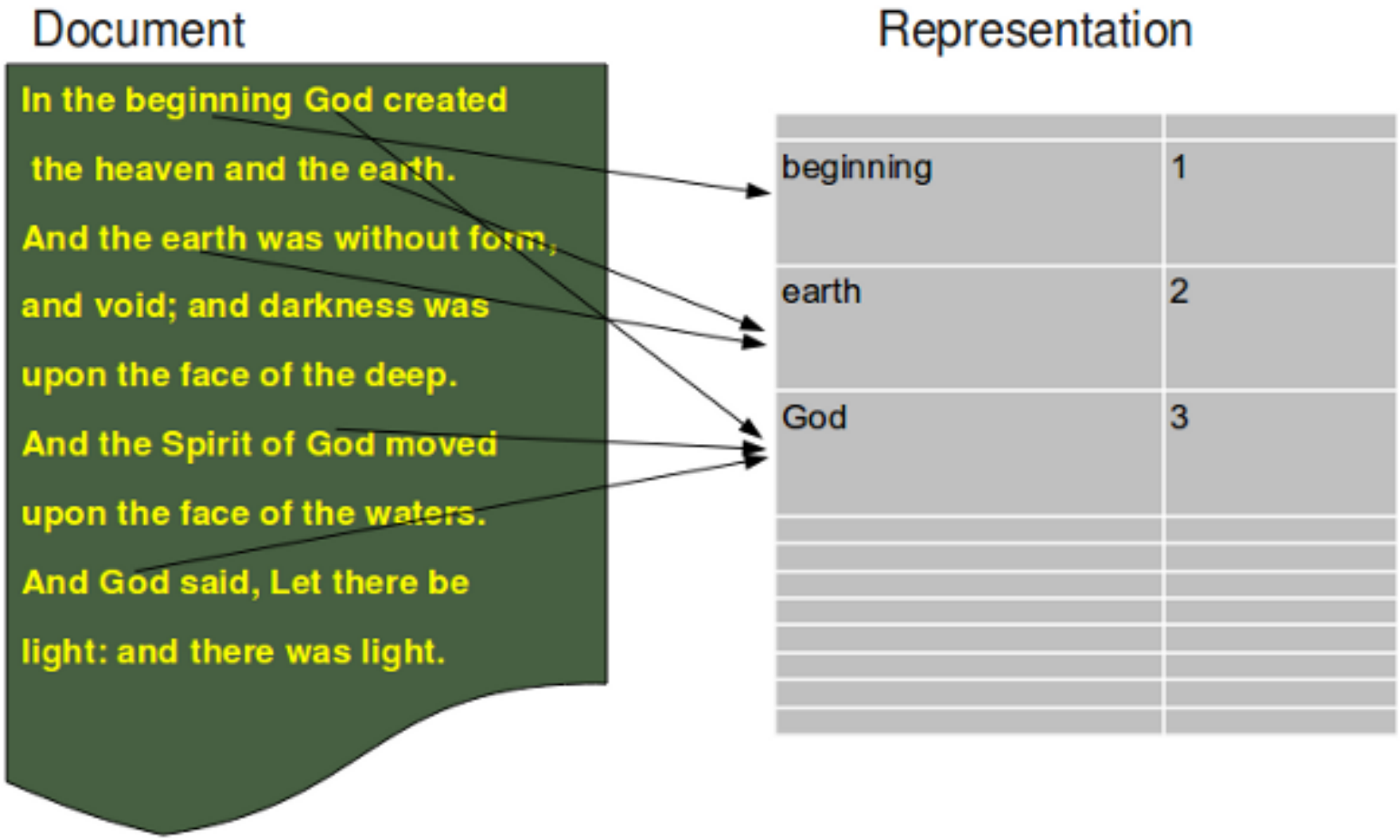


1

¿Virtudes? ¿Problemas?

[1] http://www.python-course.eu/text_classification_python.php

Bag of words



[1]

- ¿Virtudes? ¿Problemas?

[1] http://www.python-course.eu/text_classification_python.php

N-grams

- La probabilidad de un "grama" está dada por los $n-1$ anteriores
- $P(g_n|g_{n-1}, g_{n-2}, \ldots, g_{n-N+1}) = \frac{\#(g_n, g_{n-1}, g_{n-2}, \ldots, g_{n-N+1})}{\#(g_{n-1}, g_{n-2}, \ldots, g_{n-N+1})}$
- Bigramas, trigramas, etc.
- En nuestro caso, se puede hacer sobre palabras o sobre caracteres
  - Otras aplicaciones: detección de idioma (letras), cadenas de proteínas, fonemas/palabras en el contexto de procesamiento del habla
- ¿Virtudes? ¿Problemas?

N-grams

- La probabilidad de un "grama" está dada por los $n-1$ anteriores
- $P(g_n|g_{n-1}, g_{n-2}, \ldots, g_{n-N+1}) = \frac{\#(g_n, g_{n-1}, g_{n-2}, \ldots, g_{n-N+1})}{\#(g_{n-1}, g_{n-2}, \ldots, g_{n-N+1})}$
- Bigramas, trigramas, etc.
- En nuestro caso, se puede hacer sobre palabras o sobre caracteres
  - Otras aplicaciones: detección de idioma (letras), cadenas de proteínas, fonemas/palabras en el contexto de procesamiento del habla
- ¿Virtudes? ¿Problemas?

N-grams

- La probabilidad de un "grama" está dada por los $n-1$ anteriores
- $P(g_n|g_{n-1}, g_{n-2}, \ldots, g_{n-N+1}) = \frac{\#(g_n, g_{n-1}, g_{n-2}, \ldots, g_{n-N+1})}{\#(g_{n-1}, g_{n-2}, \ldots, g_{n-N+1})}$
- Bigramas, trigramas, etc.
- En nuestro caso, se puede hacer sobre palabras o sobre caracteres
  - Otras aplicaciones: detección de idioma (letras), cadenas de proteínas, fonemas/palabras en el contexto de procesamiento del habla
- ¿Virtudes? ¿Problemas?

Term frequency–inverse document frequency (tf–idf)

- tf-idf: cuán importante es una palabra en un documento respecto del conjunto de todos los documentos
- Dados d un documento en D y t un término,
- term frequency: $tf(t) = \frac{\#\text{apariciones de } t \text{ en } d}{\#\text{términos en } d}$
- inverse document frequency: $idf(t) = \log \frac{|D|}{\#\text{documentos que contienen a } t}$
- $tfidf(t) = tf(t).idf(t)$
- ¿Virtudes? ¿Problemas?

Term frequency–inverse document frequency (tf–idf)

- tf-idf: cuán importante es una palabra en un documento respecto del conjunto de todos los documentos
- Dados d un documento en D y t un término,
- term frequency: $tf(t) = \frac{\#\text{apariciones de } t \text{ en } d}{\#\text{términos en } d}$
- inverse document frequency: $idf(t) = \log \frac{|D|}{\#\text{documentos que contienen a } t}$
- $tfidf(t) = tf(t).idf(t)$
- ¿Virtudes? ¿Problemas?

Term frequency–inverse document frequency (tf–idf)

- tf-idf: cuán importante es una palabra en un documento respecto del conjunto de todos los documentos
- Dados d un documento en D y t un término,
- term frequency: $tf(t) = \frac{\#\text{apariciones de } t \text{ en } d}{\#\text{términos en } d}$
- inverse document frequency: $idf(t) = \log \frac{|D|}{\#\text{documentos que contienen a } t}$
- $tfidf(t) = tf(t).idf(t)$
- ¿Virtudes? ¿Problemas?

Stemming

- ¿Tiene sentido considerar separadamente palabras como *investigadora*, *investigador*, *investigadoras*, *investigadores*, *investigar*, *investigaron*, *investigación*, etc.?
- Reemplazar palabras por su raiz antes de usar, por ejemplo, bag-of-words
- ¿Virtudes? ¿Problemas?

Stemming

- ¿Tiene sentido considerar separadamente palabras como *investigadora*, *investigador*, *investigadoras*, *investigadores*, *investigar*, *investigaron*, *investigación*, etc.?
- Reemplazar palabras por su raiz antes de usar, por ejemplo, bag of words
- ¿Virtudes? ¿Problemas?

Stemming

- ¿Tiene sentido considerar separadamente palabras como *investigadora*, *investigador*, *investigadoras*, *investigadores*, *investigar*, *investigaron*, *investigación*, etc.?
- Reemplazar palabras por su raiz antes de usar, por ejemplo, bag of words
- ¿Virtudes? ¿Problemas?

Evaluación

- ¿Cómo sé si mi clasificador funciona bien?
- ¿Y si sufre de overfitting[2] (sobreajuste)?
- Paliativo: datos de entrenamiento y datos de validación

[2] Mal y pronto: poco error sobre el conjunto de entrenamiento pero poca generalización.

Evaluación

- ¿Cómo sé si mi clasificador funciona bien?
- ¿Y si sufre de overfitting[2] (sobreajuste)?
- Paliativo: datos de entrenamiento y datos de validación

[2] Mal y pronto: poco error sobre el conjunto de entrenamiento pero poca generalización.

Evaluación

- ¿Cómo sé si mi clasificador funciona bien?
- ¿Y si sufre de overfitting[2] (sobreajuste)?
- Paliativo: datos de entrenamiento y datos de validación

[2] Mal y pronto: poco error sobre el conjunto de entrenamiento pero poca generalización.

Validación y cross-validation[4]

- Evaluar el modelo en los datos de entrenamiento puede darnos una impresión errónea.
- Entrenamiento (100-p) % Validación p % (p.e. 20 %).
- Separando los datos al azar para evitar tomar patrones en las divisiones.
- ¿Y si al azar no funciona tan bien? → k-Fold Cross Validation
  1. Desordenar los datos
  2. Separar en k folds del mismo tamaño
  3. Para $i = 1 \ldots k$:
     Entrenar sobre todos los folds menos el i y validar sobre el i

[3] Tomada de clase de Aprendizaje Automático.

[4] Diapo fuertemente basada en las de Aprendizaje Automático

Validación y cross-validation[4]

- Evaluar el modelo en los datos de entrenamiento puede darnos una impresión errónea.
- Entrenamiento (100-p) % Validación p % (p.e. 20 %).
- Separando los datos al azar para evitar tomar patrones en las divisiones.
- ¿Y si al azar no funciona tan bien? → k-Fold Cross Validation
  1. Desordenar los datos
  2. Separar en k folds del mismo tamaño
  3. Para $i = 1 \ldots k$:
     Entrenar sobre todos los folds menos el i y validar sobre el i

[3] Tomada de clase de Aprendizaje Automático.

[4] Diapo fuertemente basada en las de Aprendizaje Automático

Validación y cross-validation[4]

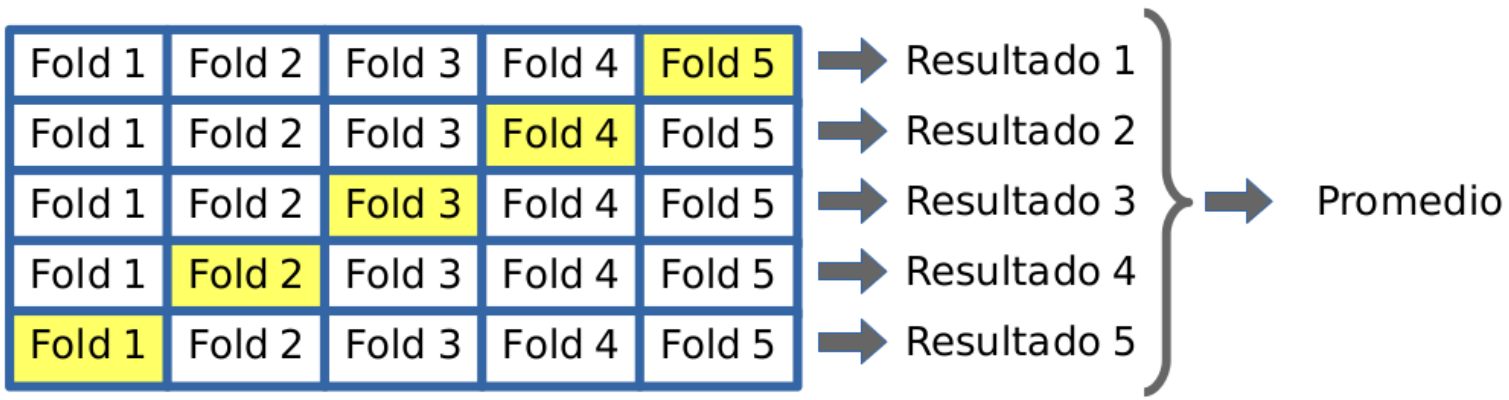
- Evaluar el modelo en los datos de entrenamiento puede darnos una impresión errónea.
- Entrenamiento (100-p) % Validación p % (p.e. 20 %).
- Separando los datos al azar para evitar tomar patrones en las divisiones.
- ¿Y si al azar no funciona tan bien? → k-Fold Cross Validation
  1. Desordenar los datos
  2. Separar en k folds del mismo tamaño
  3. Para $i = 1 \ldots k$:
     Entrenar sobre todos los folds menos el i y validar sobre el i

[3] Tomada de clase de Aprendizaje Automático.

[4] Diapo fuertemente basada en las de Aprendizaje Automático

Validación y cross-validation[4]

- Evaluar el modelo en los datos de entrenamiento puede darnos una impresión errónea.
- Entrenamiento (100-p) % Validación p % (p.e. 20 %).
- Separando los datos al azar para evitar tomar patrones en las divisiones.
- ¿Y si al azar no funciona tan bien? → k-Fold Cross Validation
  1. Desordenar los datos
  2. Separar en k folds del mismo tamaño
  3. Para $i = 1 \dots k$:
     Entrenar sobre todos los folds menos el i y validar sobre el i



3

[3] Tomada de clase de Aprendizaje Automático.

[4] Diapo fuertemente basada en las de Aprendizaje Automático

¿Qué experimentar y cómo?

- Entender el problema
- Visualizar los resultados. ¿Qué medidas de performance podré usar?
  - Exactitud (accuracy): porcentaje de instancias bien clasificadas
    - A favor: es fácil de entender y reportar
    - En contra: puede ser engañosa. 95 % parece muy bueno pero ¿y si hay 2 clases y el 98 % del total pertenece a una?

¿Qué experimentar y cómo?

- Entender el problema
- Visualizar los resultados. ¿Qué medidas de performance podré usar?
  - Exactitud (accuracy): porcentaje de instancias bien clasificadas
    - A favor: es fácil de entender y reportar
    - En contra: puede ser engañosa. 95 % parece muy bueno pero ¿y si hay 2 clases y el 98 % del total pertenece a una?

¿Qué experimentar y cómo?

- Entender el problema
- Visualizar los resultados. ¿Qué medidas de performance podré usar?
  - Exactitud (accuracy): porcentaje de instancias bien clasificadas
    - A favor: es fácil de entender y reportar
    - En contra: puede ser engañosa. 95 % parece muy bueno pero ¿y si hay 2 clases y el 98 % del total pertenece a una?

¿Qué experimentar y cómo?

- Entender el problema
- Visualizar los resultados. ¿Qué medidas de performance podré usar?
  - Exactitud (accuracy): porcentaje de instancias bien clasificadas
    - A favor: es fácil de entender y reportar
    - En contra: puede ser engañosa. 95 % parece muy bueno pero ¿y si hay 2 clases y el 98 % del total pertenece a una?

¿Qué experimentar y cómo?

- Entender el problema
- Visualizar los resultados. ¿Qué medidas de performance podré usar?
  - Exactitud (accuracy): porcentaje de instancias bien clasificadas
    - A favor: es fácil de entender y reportar
    - En contra: puede ser engañosa. 95 % parece muy bueno pero ¿y si hay 2 clases y el 98 % del total pertenece a una?

Matriz de confusión

Matriz de Confusión:
(Clasificación binaria)

tp: true positives
tn: true negatives
fp: false positives
fn: false negatives

	SPAM (predicho)	**NO SPAM (predicho)**
SPAM (real)	2739 **tp**	56 **fn**
NO SPAM (real)	4 **fp**	1042 **tn**

[5]

Definiciones

- Documento recuperado = Positivo predicho
  Por ejemplo mail clasificado como spam por el modelo
- Documento relevante = Positivo real
  Por ejemplo mail clasificado como spam por el usuario

- $Precision = \frac{tp}{tp+fp}$ De los recuperados, qué porcentaje son relevantes
- $Recall = \frac{tp}{tp+fn}$ De los relevantes, qué porcentaje son recuperados

[5] Tomada de clase de Aprendizaje Automático.

Matriz de confusión

Matriz de Confusión:
(Clasificación binaria)

tp: true positives
tn: true negatives
fp: false positives
fn: false negatives

	SPAM (predicho)	**NO SPAM (predicho)**
SPAM (real)	2739 **tp**	56 **fn**
NO SPAM (real)	4 **fp**	1042 **tn**

[5]

Definiciones

- Documento recuperado = Positivo predicho
  Por ejemplo mail clasificado como spam por el modelo
- Documento relevante = Positivo real
  Por ejemplo mail clasificado como spam por el usuario

- $Precision = \frac{tp}{tp+fp}$ De los recuperados, qué porcentaje son relevantes
- $Recall = \frac{tp}{tp+fn}$ De los relevantes, qué porcentaje son recuperados

[5] Tomada de clase de Aprendizaje Automático.

Matriz de confusión

Matriz de Confusión:
(Clasificación binaria)

tp: true positives
tn: true negatives
fp: false positives
fn: false negatives

	SPAM (predicho)	**NO SPAM (predicho)**
SPAM (real)	2739 **tp**	56 **fn**
NO SPAM (real)	4 **fp**	1042 **tn**

5

Definiciones

- Documento recuperado = Positivo predicho
  Por ejemplo mail clasificado como spam por el modelo
- Documento relevante = Positivo real
  Por ejemplo mail clasificado como spam por el usuario

- $Precision = \frac{tp}{tp+fp}$ De los recuperados, qué porcentaje son relevantes
- $Recall = \frac{tp}{tp+fn}$ De los relevantes, qué porcentaje son recuperados

[5] Tomada de clase de Aprendizaje Automático.

Más medidas

- $Sensitivity = \frac{tp}{tp+fn}$
  Porcentaje de pacientes enfermos correctamente diagnosticados[6]
- $Specificity = \frac{tn}{tn+fp}$
  Porcentaje de pacientes sanos correctamente diagnosticados

F-measures

- Media armónica: $F_1 = 2\frac{precision.recall}{precision+recall}$
- Fórmula general: $F_\beta = (1+\beta^2)\frac{precision.recall}{\beta^2.precision+recall}$
- F_2 enfatiza recall mientras que $F_{0,5}$ enfatiza precision

[6] Notar que Sensitivity es Recall.

Más medidas

κ de Cohen

Indica cuánto concuerdan dos clasificadores sobre los mismos datos. Es más robusta que un cálculo de porcentaje de acuerdo ya que tiene en cuenta el acuerdo por casualidad.

$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$

donde p_o es el acuerdo relativo entre los clasificadores y p_e es la probabilidad hipotética de acuerdo por casualidad.

$\kappa \leq 1$ y valores cercanos a 1 indican un buen nivel de acuerdo mientras que valores negativos indican lo contrario.

Ejemplo

		A	
		Sí	No
B	Sí	17	7
	No	5	14

$p_o = \frac{17+14}{43} = 0{,}721$

$p_e = P(\text{Sí}|A) * P(\text{Sí}|B) + P(\text{No}|A) * P(\text{No}|B)$

$p_e = \frac{22}{43} * \frac{24}{43} + \frac{21}{43} * \frac{19}{43} = 0{,}501$

$\kappa = \frac{p_o - p_e}{1 - p_e} = 0{,}441$

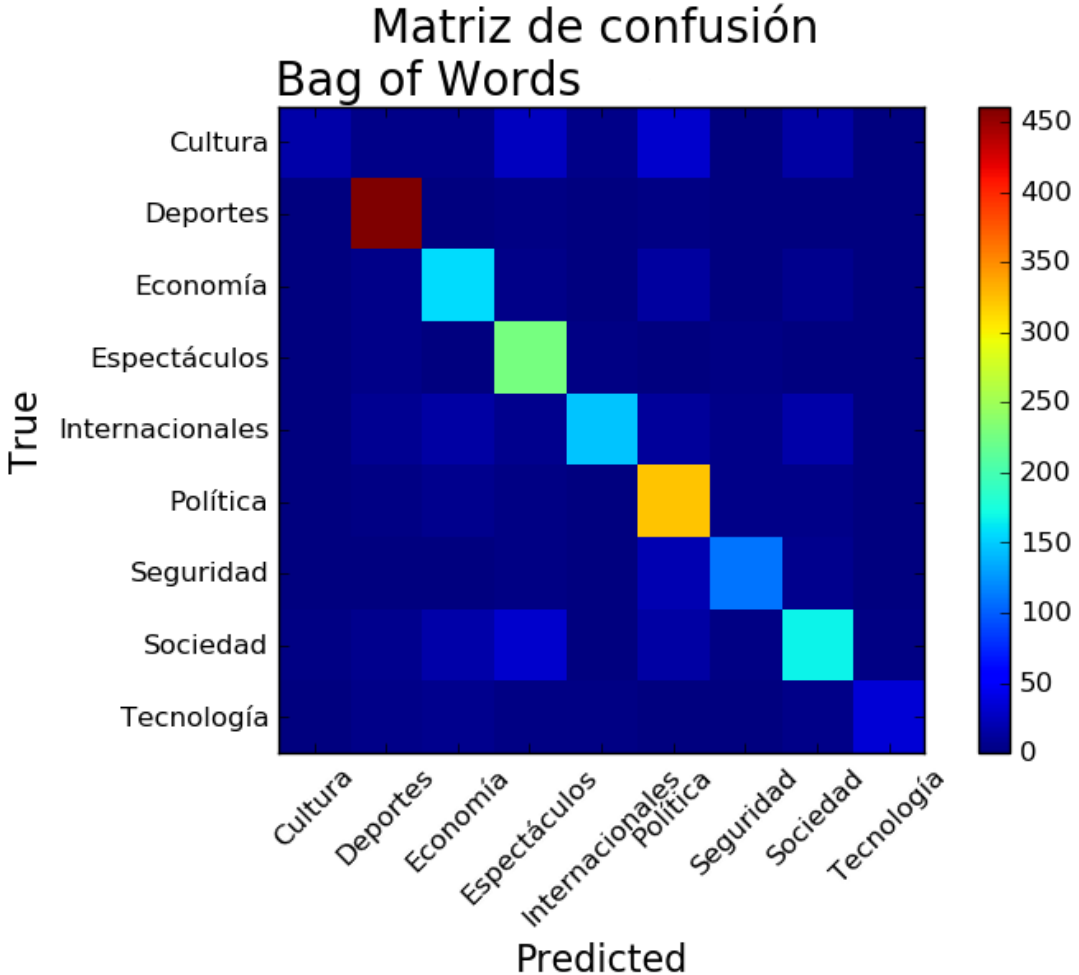
Resultados

- Corremos el clasificador sobre los datos y obtenemos 82,4 % de accuracy. Nada mal.

- ¿Algo interesante para destacar?

Resultados

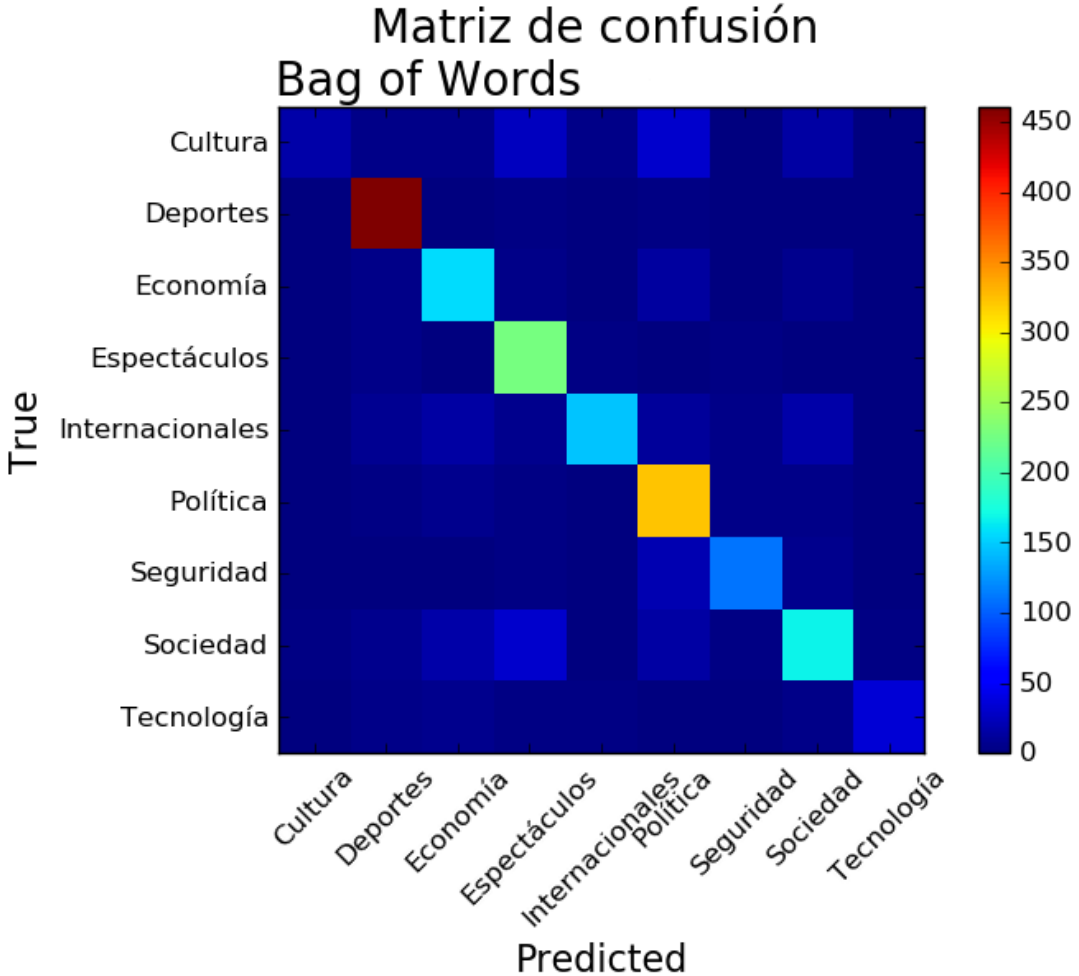
- Corremos el clasificador sobre los datos y obtenemos 82,4 % de accuracy. Nada mal.



- ¿Algo interesante para destacar?

Resultados

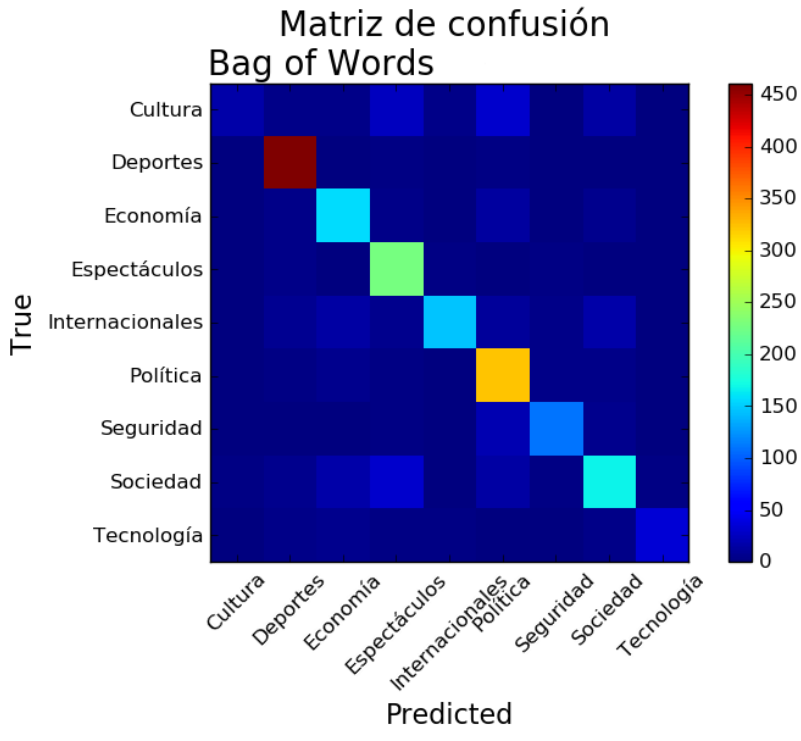
- Corremos el clasificador sobre los datos y obtenemos 82,4 % de accuracy. Nada mal.



- ¿Algo interesante para destacar?

Resultados - Problemas

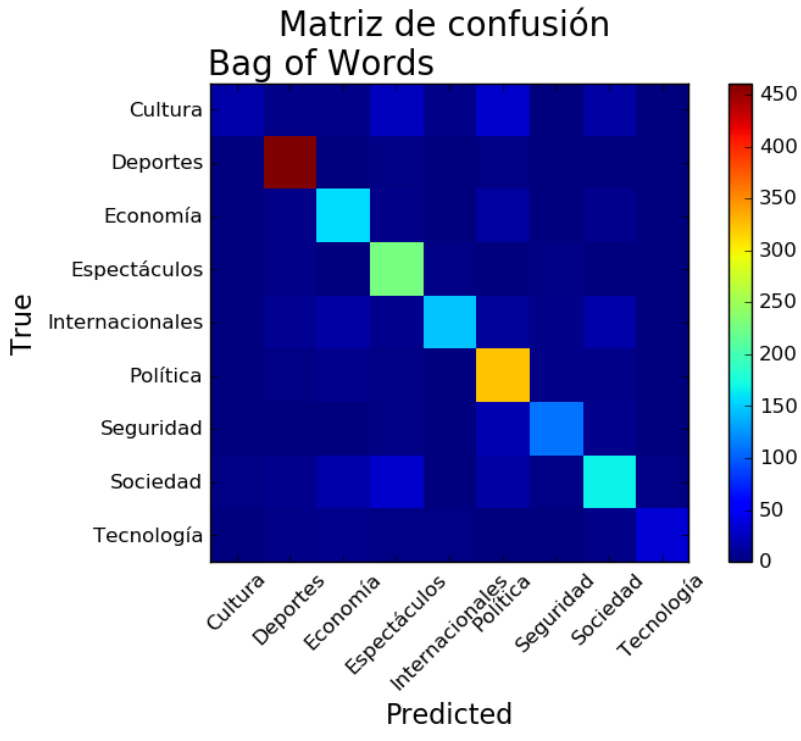
- Muchas instancias de “Cultura” son clasificadas como otras categorías, principalmente como “Política”
- Algo similar sucede con “Sociedad”
- Tal vez esas categorías tienen muchas palabras en común.
  ¿Y si sacamos las stopwords[7] del texto? Accuracy: 88,8 %



[7] Las palabras más usadas en un lenguaje, como por ejemplo artículos y preposiciones, no aportan información relevante al tipo de texto pero pueden influir negativamente en la clasificación.

Resultados - Problemas

- Muchas instancias de “Cultura” son clasificadas como otras categorías, principalmente como “Política”
- Algo similar sucede con “Sociedad”
- Tal vez esas categorías tienen muchas palabras en común. ¿Y si sacamos las stopwords[7] del texto? Accuracy: 88,8 %

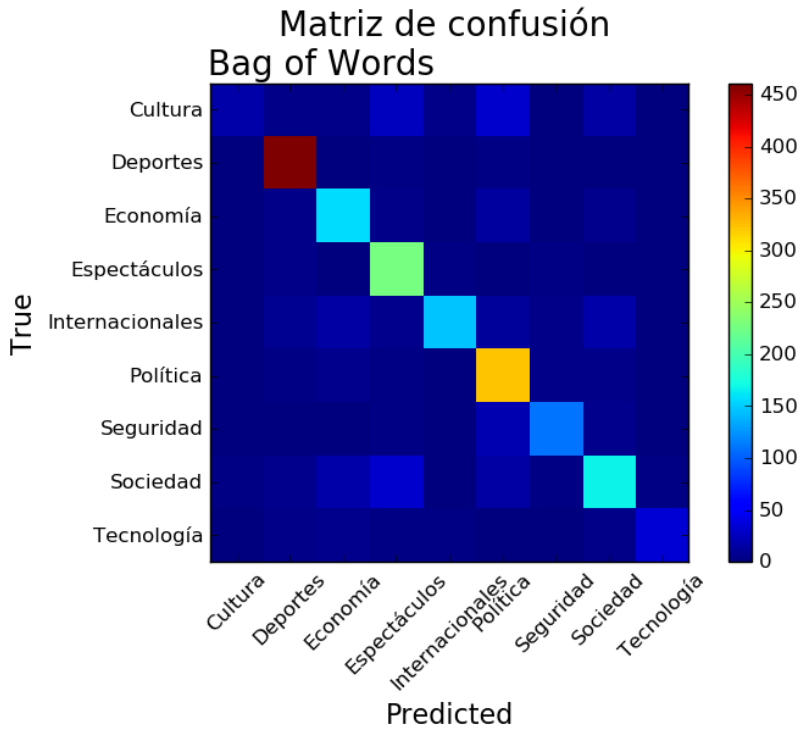


[7] Las palabras más usadas en un lenguaje, como por ejemplo artículos y preposiciones, no aportan información relevante al tipo de texto pero pueden influir negativamente en la clasificación.

Resultados - Problemas

- Muchas instancias de “Cultura” son clasificadas como otras categorías, principalmente como “Política”
- Algo similar sucede con “Sociedad”
- Tal vez esas categorías tienen muchas palabras en común.

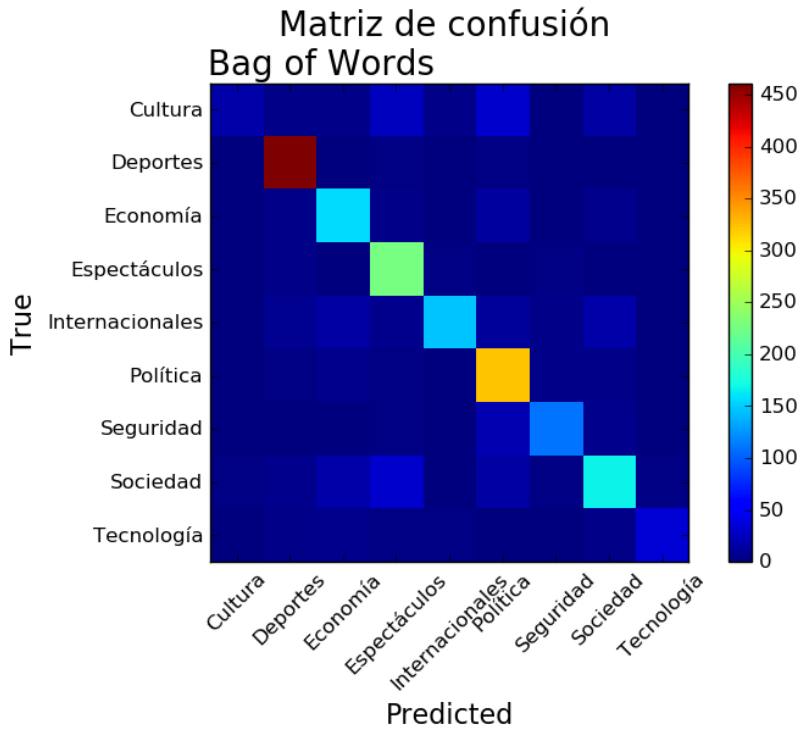
¿Y si sacamos las stopwords[7] del texto? Accuracy: 88,8 %



[7] Las palabras más usadas en un lenguaje, como por ejemplo artículos y preposiciones, no aportan información relevante al tipo de texto pero pueden influir negativamente en la clasificación.

Resultados - Problemas

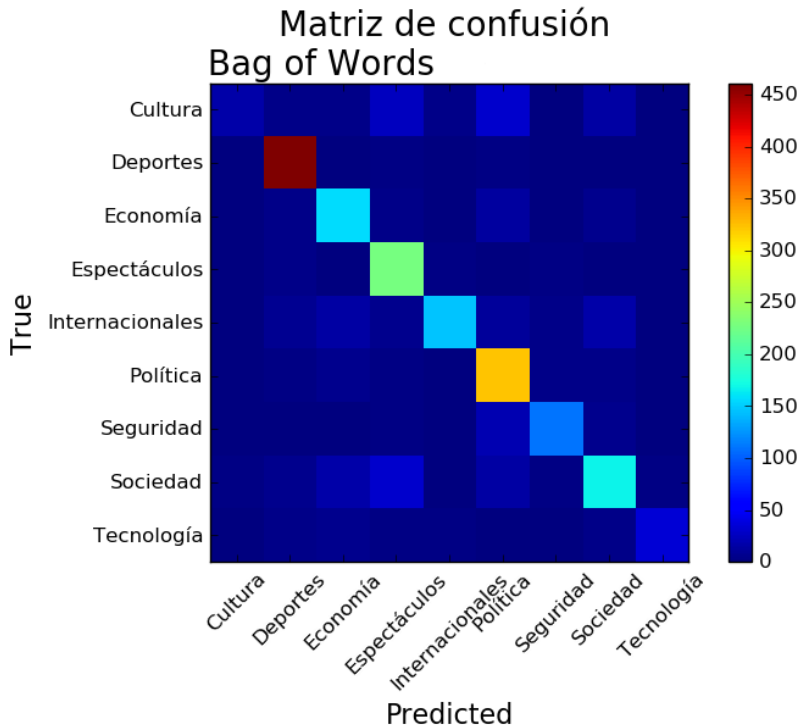
- Muchas instancias de “Cultura” son clasificadas como otras categorías, principalmente como “Política”
- Algo similar sucede con “Sociedad”
- Tal vez esas categorías tienen muchas palabras en común. ¿Y si sacamos las stopwords[7] del texto? Accuracy: 88,8 %

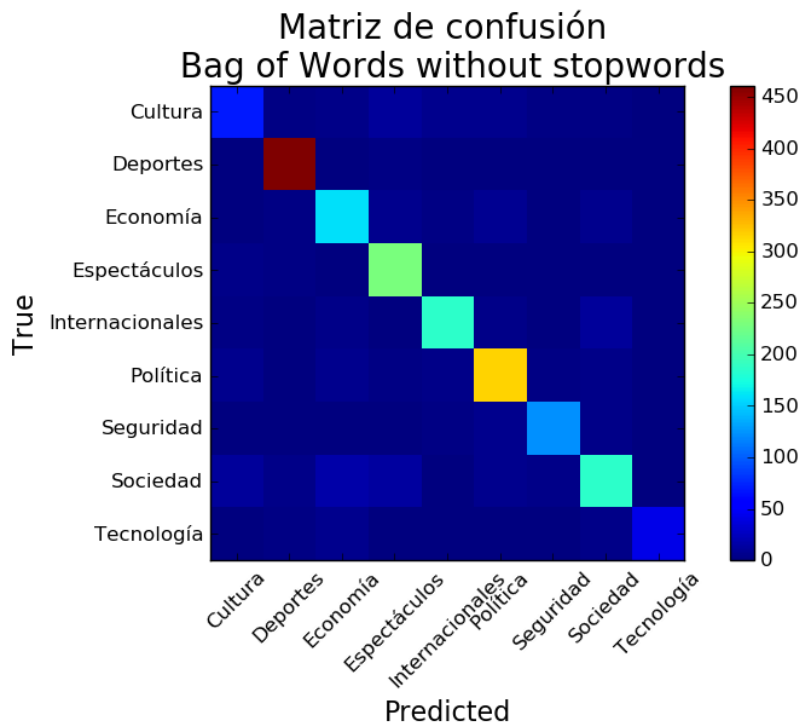


[7] Las palabras más usadas en un lenguaje, como por ejemplo artículos y preposiciones, no aportan información relevante al tipo de texto pero pueden influir negativamente en la clasificación.

Resultados - Problemas

- Muchas instancias de "Cultura" son clasificadas como otras categorías, principalmente como "Política"
- Algo similar sucede con "Sociedad"
- Tal vez esas categorías tienen muchas palabras en común. ¿Y si sacamos las stopwords[7] del texto? Accuracy: 88,8 %





[7] Las palabras más usadas en un lenguaje, como por ejemplo artículos y preposiciones, no aportan información relevante al tipo de texto pero pueden influir negativamente en la clasificación.

Resumen

- La experimentación no es sólo reportar resultados. En base a los resultados se gana entendimiento y se repiensa el problema y esto permite iterar nuevamente con experimentos.
- Es importante elegir una manera adecuada para mostrar los resultados. Ciertas características pueden quedar ocultas detrás de medidas mentirosas.
- Siempre recordando los límites en términos de tiempo que hay en los TPs.

Resumen

- La experimentación no es sólo reportar resultados. En base a los resultados se gana entendimiento y se repiensa el problema y esto permite iterar nuevamente con experimentos.
- Es importante elegir una manera adecuada para mostrar los resultados. Ciertas características pueden quedar ocultas detrás de medidas mentirosas.
- Siempre recordando los límites en términos de tiempo que hay en los TPs.

Resumen

- La experimentación no es sólo reportar resultados. En base a los resultados se gana entendimiento y se repiensa el problema y esto permite iterar nuevamente con experimentos.
- Es importante elegir una manera adecuada para mostrar los resultados. Ciertas características pueden quedar ocultas detrás de medidas mentirosas.
- Siempre recordando los límites en términos de tiempo que hay en los TPs.