Tecnológico de Monterrey

Aprendizaje automático

Preprocesamiento de texto

Dra. Yuridiana Alemán

¿Cómo representar un texto?

Tecnológico de Monterrey

Datos numéricos

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
```

Texto

NWH_Ingles.txt

```
1  1   I would definitely recommend this movie, another win for Marvel, they do an excellent job keeping the super hero formula fresh, but definitely not the best I've seen from Spiderman  Positivo
2  2   The Spider-Man movie series is thriving harder than before with "No Way Home". With no doubts of being a true classic within the MARVEL movie lineup.   Positivo
3  3   This was a movie. Not the concept that I expected, but all the action I loved from the other two movies. Seeing the 3 Spider-Men actors get together was fantastic. Seeing the set up for everything. Watching Tom's Spider-Man beat the demons that once corrupted the other 2 was a nice way to end this movie.   Positivo
4  4   Absolutely amazing! Great cameos and strory. Can't wait to see Tom Holland back in action!  Positivo
5  5   The best movie of my life, in my opinion is better than avengers endgame, watch the three spidermans fighting together and the classic villains is awesome. Positivo
6  6   This movie is only hype....There are so plots holes...  Negativo
```

Procesamiento de texto

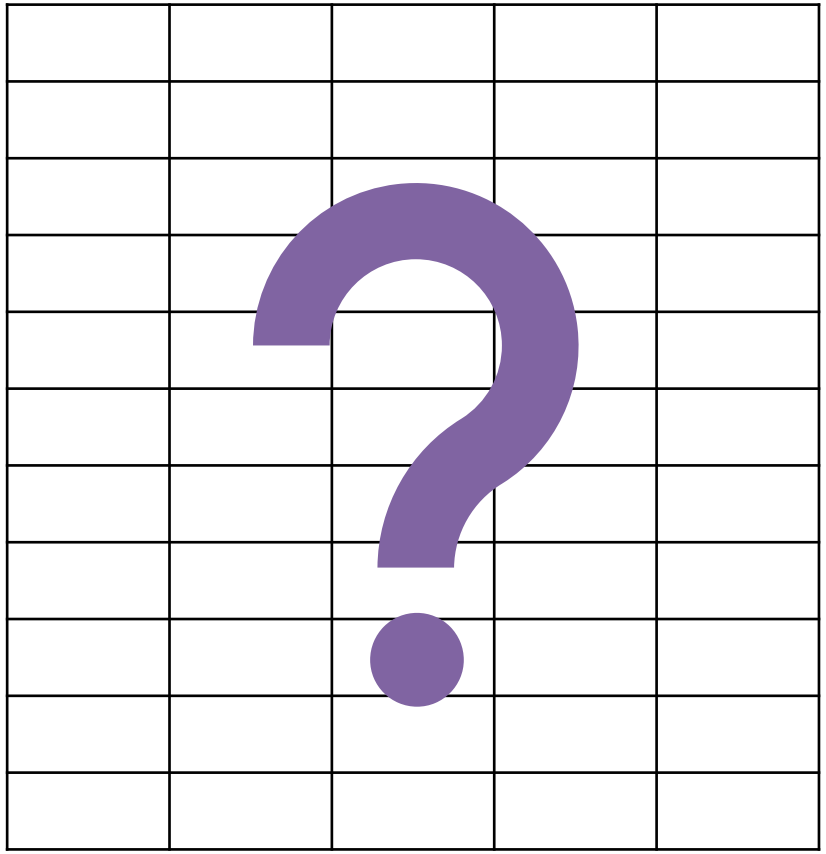
Tecnológico de Monterrey

Texto

NWH_Ingles.txt

```
1  1   I would definitely recommend this movie, another win for Marvel, they do an excellent job keeping the super hero formula fresh, but definitely not the best I've seen from Spiderman Positivo
2  2   The Spider-Man movie series is thriving harder than before with "No Way Home". With no doubts of being a true classic within the MARVEL movie lineup.   Positivo
3  3   This was a movie. Not the concept that I expected, but all the action I loved from the other two movies. Seeing the 3 Spider-Men actors get together was fantastic. Seeing the set up for everything. Watching Tom's Spider-Man beat the demons that once corrupted the other 2 was a nice way to end this movie.   Positivo
4  4   Absolutely amazing! Great cameos and strory. Can't wait to see Tom Holland back in action!  Positivo
5  5   The best movie of my life, in my opinion is better than avengers endgame, watch the three spidermans fighting together and the classic villains is awesome. Positivo
6  6   This movie is only hype....There are so plots holes...  Negativo
```

Representación numérica



¿Cómo representar un texto?

Tecnológico de Monterrey

Características

- ✓ Vocabulario (Bolsa de palabras)
- ✓ Lemas
- ✓ Categorías gramaticales
- ✓ Stemming
- ✓ Sentimientos

Representación

- ✓ Binaria
- ✓ Frecuencia de términos
- ✓ Frecuencia inversa (Tf-Idf)

	Características
Instancias	Representación

EXTRACIÓN DE CARACTERÍSTICAS

	Características
Instancias	Representación

NLTK

- https://www.nltk.org/
- Es una Plataforma para trabajar con datos del lenguaje humano (sobre todo para el idioma Inglés)
- Integra un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, lematización, etiquetado, análisis y razonamiento semántico, entre otras opciones

Tokenización

Tecnológico de Monterrey

- ✓ Divide un texto en una lista de subcadenas (usualmente palabras)

```
>>> from nltk.tokenize import word_tokenize
>>> s = '''Good muffins cost $3.88\nin New York.  Please buy me
... two of them.\n\nThanks.'''
>>> word_tokenize(s)
['Good', 'muffins', 'cost', '$', '3.88', 'in', 'New', 'York', '.',
'Please', 'buy', 'me', 'two', 'of', 'them', '.', 'Thanks', '.']
```

```
>>> from nltk.tokenize import wordpunct_tokenize
>>> wordpunct_tokenize(s)
['Good', 'muffins', 'cost', '$', '3', '.', '88', 'in', 'New', 'York', '
'Please', 'buy', 'me', 'two', 'of', 'them', '.', 'Thanks', '.']
```

```
>>> from nltk.tokenize import wordpunct_tokenize
>>> wordpunct_tokenize(s)
['Good', 'muffins', 'cost', '$', '3', '.', '88', 'in', 'New', 'York', '
'Please', 'buy', 'me', 'two', 'of', 'them', '.', 'Thanks', '.']
```

Tokenización

Tecnológico de Monterrey

- ✓ Divide un texto en una lista de subcadenas (usualmente palabras)

```
from nltk.tokenize import word_tokenize as wt

def Tokeniza(doc):
    docto=codecs.open('Iniciales/'+doc+'.txt','r')
    Salida=codecs.open('Documentos/'+doc+'-Tokens.txt','w')
    for x in docto.readlines():
        if len(x)>1:
            datos=x.split('\t')
            Texto=wt(datos[1].lower())
            CadFinal = " ".join(Texto)
            Salida.write(datos[0]+'\t'+CadFinal+'\t'+datos[2])
    docto.close()
    Salida.close()
```

LEMAS, STOPWORDS,CATEGORÍAS GRAMATICALES

Lematización

- ✓ El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra.

am, you, are, is → be

Estudiante, estudiamos, estudio→ estudiar

Lematización

Tecnológico de Monterrey

- ✓ WordNet

```
>>> import nltk
>>> nltk.download('wordnet')
```

```
>>> from nltk.stem.wordnet import WordNetLemmatizer
>>> lmtzr = WordNetLemmatizer()
>>> lmtzr.lemmatize('cars')
'car'
>>> lmtzr.lemmatize('feet')
'foot'
>>> lmtzr.lemmatize('people')
'people'
```

Lematización

Tecnológico de Monterrey

- ✓ CLIPS (https://www.clips.uantwerpen.be/clips.bak/pages/pattern)

```
00851429b21722a4d62f63a328c601ca    en  ey sorry i am late I printed directions
```

```
from pattern.en import parse, split
docto=codecs.open('Doctos/English.txt','r')
for x in docto.readlines():
    datos=x.split('\t')
    s = parse(datos[2], lemmata=True)
    print(s)
docto.close()
```

ey/NN/B-NP/O/ey sorry/VB/B-VP/O/sorry i/NN/B-NP/O/i am/VBP/B-VP/O/be late/RB/B-ADVP/O/late I/PRP/B-NP/O/i printed/VBP/B-VP/O/print directions/NNS/B-NP/O/direction

Stemming

Tecnológico de Monterrey

- ✓ Proceso de reducir la inflexión en las palabras a sus formas de raíz, incluso si la raíz en sí no es una palabra válida en el idioma (Derivar una palabra o una oración puede dar como resultado palabras que no son palabras reales).

```
# importing modules
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

ps = PorterStemmer()

sentence = "Programmers program with programming languages"
words = word_tokenize(sentence)

for w in words:
    print(w, " : ", ps.stem(w))
```

```
Programmers  :  program
program  :  program
with  :  with
programming  :  program
languages  :  languag
```

Stemming

Tecnológico de Monterrey

```
1  # -*- coding: utf-8 -*-
2  import codecs
3  import os, sys
4  import re
5  from nltk.stem import PorterStemmer
6
7  def Stemming(doc):
8      docto=codecs.open('Documentos/'+doc+'-Lemas.txt','r')
9      Salida=codecs.open('Documentos/'+doc+'-Sttemming.txt','w')
10     for x in docto.readlines():
11         if len(x)>1:
12             datos=x.split('\t')
13             Texto=datos[1].lower().split(' ')
14             CadFinal=""
15             for w in Texto:
16                 print(w, ": ", ps.stem(w))
17                 CadFinal=CadFinal+ps.stem(w)+' '
18             Salida.write(datos[0]+'\t'+CadFinal[1:]+'\t'+datos[2])
19     docto.close()
20     Salida.close()
21 ##############################################################
22 ps = PorterStemmer()
23 Stemming('NWH_Ingles')
24 Stemming('NWH_Espanol')
```

POS tagging

Tecnológico de Monterrey

- ✓ Part-of-speech tagging es el proceso que recibe como entrada texto en algún lenguaje y como salida regresa un conjunto de pares de la forma palabra-etiqueta gramatical (sustantivo, verbo, adjetivo, etc)

```
yuri@yuri-H110M-S2:~/Escritorio/ProcesamientoTexto$ python3 Lemas2.py
i/i/CCONJ
would/would/PROPN
definitely/definitely/ADJ
recommend/recommend/PROPN
this/this/PROPN
movie/movie/ADJ
,/,/PUNCT
another/another/PROPN
win/win/PROPN
for/for/PROPN
marvel/marvel/PROPN
,/,/PUNCT
they/they/ADJ
do/do/ADP
an/an/PROPN
excellent/excellent/PROPN
job/job/PROPN
keeping/keeping/PROPN
the/the/PROPN
super/super/PROPN
hero/hero/ADJ
formula/formula/PROPN
fresh/fresh/PROPN
,/,/PUNCT
but/but/NOUN
```

StopWords

Tecnológico de Monterrey

Palabras cerradas, palabras vacías.

- ✓ Términos extremadamente comunes que suelen aparecer en muchas ocasiones, generalmente son eliminadas en el preprocesamiento de texto.
- ✓ Palabras sin significado como artículos, pronombres, preposiciones.
- ✓ No hay una lista definitiva de StopWords
- ✓ No todas las herramientas de procesamiento de texto tienen disponibl
- ✓ No en todos los problemas es conveniente eliminarlas

a	*at*	*has*	*its*	*to*
an	*be*	*he*	*of*	*was*
and	*by*	*in*	*on*	*were*
are	*for*	*is*	*that*	*will*
as	*at*	*it*	*the*	*with*

StopWords

Tecnológico de Monterrey

```
00851429b21722a4d62f63a328c601ca    en  ey sorry i am late I printed directions
```

```
from nltk.corpus import stopwords

docto=codecs.open('Doctos/English.txt','r')

for x in docto.readlines():
    datos=x.split('\t')
    Text=datos[2].split(' ')
    NoStopWords=[x for x in Text if not x in stopwords.words('english') ]
    Out = (" ").join(NoStopWords)
    print(Out)
docto.close()
```

ey sorry late I printed directions

```
10016a25a9298bf98a494ed4de182e1b    es  somos esa clase de gente con la que nuestros padres, no quieren
```

clase gente padres quieren

```
reload(sys)
sys.setdefaultencoding("utf-8")
docto=codecs.open('Doctos/Spanish.txt','r')
for x in docto.readlines():
    datos=x.split('\t')
    Text=datos[2].split(' ')
    NoStopWords=[x for x in Text if not x in stopwords.words('spanish') ]
    Out = (" ").join(NoStopWords)
    print(Out)
docto.close()
```

Herramientas de procesamiento

Tecnológico de Monterrey

- WordNet
- POS Tagger Stanford
- Tree Tagger
- CLIPS Pattern
- Spycy
- Freeling

REPRESENTACIÓN

	Características
Instancias	Representación

Ocurrencia

Tecnológico de Monterrey

- ✓ Agrega un valor de atributo 0 cuando la palabra no aparece en la instancia, y 1 cuando aparece, sin importar las veces que lo haga

Id	ok	overall	top	10	dk	others	overrated	waste	money	Clase
1	0	0	0	0	0	0	0	0	0	Positivo
2	0	0	0	0	0	0	0	0	0	Positivo
3	0	0	0	0	0	0	0	0	0	Positivo
4	0	0	0	0	0	0	0	0	0	Positivo
5	0	0	0	0	0	0	0	0	0	Positivo
6	0	0	0	0	0	0	0	0	0	Negativo
7	0	0	0	0	0	0	0	0	0	Positivo
8	0	0	0	0	0	0	0	0	0	Positivo
9	0	0	0	0	0	0	0	0	0	Positivo
10	0	0	0	0	0	0	0	0	0	Positivo
11	0	0	0	0	0	0	0	0	0	Negativo
12	0	0	0	0	0	0	0	0	0	Negativo
13	0	0	0	0	0	0	0	0	0	Negativo
14	1	1	1	1	0	0	0	0	0	Negativo
15	0	0	0	0	1	1	1	1	1	Negativo

Frecuencia de términos

- ✓ La frecuencia de términos puede representarse como :

$$tf_{t,d} = Apariciones\ de\ t\ en\ d$$

Donde:

t= Término

d=Instancia

Id	would	definitely	recommend	movie	another	win	marvel	excellent	Clase
1	1	2	1	1	1	1	1	1	Positivo
2	0	0	0	2	0	0	1	0	Positivo
3	0	0	0	3	0	0	0	0	Positivo
4	0	0	0	0	0	0	0	0	Positivo
5	0	0	0	1	0	0	0	0	Positivo
6	0	0	0	1	0	0	0	0	Negativo
7	0	0	0	2	0	0	0	0	Positivo
8	1	0	0	0	0	0	1	0	Positivo
9	0	0	0	0	0	0	1	0	Positivo
10	0	0	0	0	0	0	0	0	Positivo
11	0	0	0	2	0	0	2	0	Negativo
12	1	0	0	3	0	0	0	0	Negativo
13	1	0	0	1	0	0	0	0	Negativo
14	0	1	0	4	0	0	2	0	Negativo
15	0	0	0	0	0	0	0	0	Negativo

Frecuencia inversa del documento

Asigna un peso a cada término

$$idf_t = log \frac{N}{df_t}$$

Donde:

N= Número de documentos

df_t=Frecuencia de documentos (número de documentos en los que aparece el término)

Término	df_t	idf_t
car	18,165	1.65
auto	6,723	2.08
insurance	19,241	1.62
best	25,235	1.50

N=806,791 documentos

Ponderación tf - idf

Tecnológico de Monterrey

Produce un peso para cada término en un documento:

$$tf\text{–}idf_{t,d} = tf_{t,d} \times idf_t$$

En otras palabras, asigna a un termino *t* un peso en un documento *d* que es:

- ✓ Muy alto cuando *t* ocurre muchas veces en pocos documentos.
- ✓ Bajo cuando el termino ocurre pocas veces en un documento, o en muchos documentos.
- ✓ Muy bajo cuando el termino ocurre en la mayoría de los documentos.

Ejemplo:

Tecnológico de Monterrey

	Doc1	Doc2	Doc3	Df_t
Carro	27	4	24	18,165
Casa	3	33	0	6,723
Flor	0	33	29	19,241
Internet	14	0	17	25,235

N=806,791

Ejemplo:

	Doc1	Doc2	Doc3	Df_t
Carro	27	4	24	18,165
Casa	3	33	0	6,723
Flor	0	33	29	19,241
Internet	14	0	17	25,235

N=806,791

$$idf_t = log\frac{N}{df_t}$$

$$idf_{carro} = \log\left(\frac{806791}{18165}\right) = \log(44.4244) = 1.648$$

Ejemplo:

	Doc1	Doc2	Doc3	Df_t
Carro	27	4	24	18,165
Casa	3	33	0	6,723
Flor	0	33	29	19,241
Internet	14	0	17	25,235

N=806,791

$$idf_t = log\frac{N}{df_t}$$

$$idf_{carro} = \log\left(\frac{806791}{18165}\right) = \log(44.4244) = 1.648$$

$$tf\text{–}idf_{t,d} = tf_{t,d} \times idf_t$$

$$tf - idf_{carro,Doc1} = 27 * 1.648 = 44.4858$$

Ejemplo:

Tecnológico de Monterrey

	Doc1	Doc2	Doc3	Df_t	
Carro	27	4	24		18,165
Casa	3	33	0		6,723
Flor	0	33	29		19,241
Internet	14	0	17	N=806,791	25,235

Palabra	Idf	tf - idf		
		Doc1	Doc2	Doc3
Carro	1.6476	44.4858	6.5905	39.5429
Casa	2.0792	6.2379	68.6167	0
Flor	1.6226	0	53.5468	47.0563
Internet	1.5048	21.0680	0	25.5825

Tf-Idf

Tecnológico de Monterrey

Id	would	definitely	recommend	movie	another	win	marvel	excellent	Clase
1	0.57403127	1.75012253	1.17609126	0.17609126	1.17609126	1.17609126	0.39794001	1.17609126	Positivo
2	0	0	0	0.35218252	0	0	0.39794001	0	Positivo
3	0	0	0	0.52827378	0	0	0	0	Positivo
4	0	0	0	0	0	0	0	0	Positivo
5	0	0	0	0.17609126	0	0	0	0	Positivo
6	0	0	0	0.17609126	0	0	0	0	Negativo
7	0	0	0	0.35218252	0	0	0	0	Positivo
8	0.57403127	0	0	0	0	0	0.39794001	0	Positivo
9	0	0	0	0	0	0	0.39794001	0	Positivo
10	0	0	0	0	0	0	0	0	Positivo
11	0	0	0	0.35218252	0	0	0.79588002	0	Negativo
12	0.57403127	0	0	0.52827378	0	0	0	0	Negativo
13	0.57403127	0	0	0.17609126	0	0	0	0	Negativo
14	0	0.87506126	0	0.70436504	0	0	0.79588002	0	Negativo
15	0	0	0	0	0	0	0	0	Negativo

¿Otras características?

Tecnológico de Monterrey

- ✓ De acuerdo con el conocimiento del dominio

¿Cómo determinar las características a utilizar en un algoritmo de aprendizaje?

Tecnológico de Monterrey

- ✓ ¿Se deben de agregar todas las características en un solo modelo?
- ✓ ¿Es posible tener diferentes resultados de clasificación utilizando el mismo clasificador pero diferentes conjuntos de características?
- ✓ Si un conjunto de características es el que tiene buenos resultados en el conjunto A, ¿Seguirá teniendo los mejores resultados en otro conjunto B?