

Actividad 2.2. Descubriendo asociaciones de palabras

Alumno: Álvaro Miguel Toriz Proenza

27 de agosto de 2021

Introducción

La presente actividad está pensada para que entendamos y hagamos un ejercicio para descubrir asociaciones de palabras.

De acuerdo con Church y Hanks (1989), el estudio estadístico de las asociaciones de palabras tiene un gran número de aplicaciones potenciales como, por ejemplo, reducir el modelo de lenguaje tanto para el reconocimiento de voz como para reconocimiento óptico de caracteres (OCR), proporcionar claves de desambiguación para analizar estructuras sintácticas altamente ambiguas como compuestos de sustantivos, conjunciones y preposiciones, recuperar textos de grandes bases de datos (por ejemplo, periódicos, patentes), mejorar la productividad de los lingüistas computacionales en la compilación de léxicos de hechos léxico-sintácticos y mejorar la productividad de lexicografistas en la identificación de uso normal y convencional del lenguaje.

Desarrollo

In [1]:

```
# Load libraries
import json
import nltk
from nltk.corpus import stopwords
nltk.download('punkt')
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
import pandas as pd
import re
import string
```

```
[nltk_data] Downloading package punkt to
[nltk_data]      /home/alvaromigueltorizproenza/nltk_data...
[nltk_data]      Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]      /home/alvaromigueltorizproenza/nltk_data...
[nltk_data]      Package stopwords is already up-to-date!
```

In [2]:

```
# Get number of lines
num_lines = sum(1 for line in open('Datos Actividad 2.2.json'))
# Create a list to store bodys
mensajes = []
# Open file and read lines
with open('Datos Actividad 2.2.json') as file:
    lines = file.readlines()
# Deserialize to a Python object each line
    objeto = [json.loads(line) for line in lines]
# Extract just 'body' items
    for x in range(num_lines):
        mensajes.append(objeto[x]['_source']['body'])
```

In [3]:

```
# Replace accents
mensajes = [x.replace("á","a") for x in mensajes]
mensajes = [x.replace("é","e") for x in mensajes]
mensajes = [x.replace("í","i") for x in mensajes]
mensajes = [x.replace("ó","o") for x in mensajes]
# All letters will be lowercase
mensajes = [x.lower() for x in mensajes]
```

In [4]:

```
# Remove Twitter users (@xxxx) and hashtags (#xxxxxx) and URLs
mensajes = [re.sub("@[A-Za-z0-9_]+","", x) for x in mensajes]
mensajes = [re.sub("#[A-Za-z0-9_]+","", x) for x in mensajes]
```

In [5]:

```
# Create a regex pattern to match all special characters
pattern = r'[' + string.punctuation + ']'
# Remove special characters
mensajes = [re.sub(pattern, '', x) for x in mensajes]
```

In [6]:

```
# Remove stop words and show a sample
def removeStopwords(tweets):
    text_tokens = word_tokenize(tweets)
    return [ word for word in text_tokens if word not in stopwords.words('spanish') ]
mensajes = [removeStopwords(x) for x in mensajes]
print(mensajes[36])
```

```
['ingratos', 'estan', 'viniendo', 'dia', 'hoy', 'cortar', 'gas', 'plena', 'contingencia']
```

In [9]:

```
# A test to find bigrams
mensajes_36 = mensajes[36]
palabras=zip(mensajes_36, mensajes_36[1:])
bigramas = [p for p in palabras]
print(bigramas)
```

```
[('ingratos', 'estan'), ('estan', 'viniendo'), ('viniendo', 'dia'), ('dia', 'hoy'), ('hoy', 'cortar'), ('cortar', 'gas'), ('gas', 'plena'), ('plena', 'contingencia')]
```

In [10]:

```
# Find all bigrams
bigramas = []
for tweet in mensajes:
    juntas=zip(tweet, tweet[1:])
    single_tweet_bigrams = [p for p in juntas]
    bigramas.append(single_tweet_bigrams)
print(type(bigramas))
```

```
<class 'list'>
```

In [11]:

```
# Show bigrams
print(bigramas)
```

```
[([('cinismo', 'puro'), ('puro', 'corte'), ('corte', 'masivo'), ('masivo', 'falta'), ('falta', 'pago'), ('pago', 'recibos'), ('recibos', 'gas'), ('gas', 'natural'), ('natural', 'pandemia'), ('pandemia', 'mexico'), ('mexico', 'parte'), ('parte', 'apoyo'), ('apoyo', 'iniciativa'), ('iniciativa', 'privada'), ('privada', 'sociedad'), ('sociedad', 'httpstcoiynv8hiq2c')], [('cinismo', 'puro'), ('puro', 'corte'), ('corte', 'masivo'), ('masivo', 'falta'), ('falta', 'pago'), ('pago', 'recibos'), ('recibos', 'gas'), ('gas', 'natural'), ('natural', 'pandemia'), ('pandemia', 'mexico'), ('mexico', 'parte'), ('parte', 'apoyo'), ('apoyo', 'iniciativa'), ('iniciativa', 'privada'), ('privada', 'sociedad')], [['gracias', 'mande'], ('mande', 'dm')], [['buenas', 'tardes'], ('tardes', 'cortaron'), ('cortaron', 'servicio'), ('servicio', 'enblas'), ('enblas', 'oficinas'), ('oficinas', 'pagar'), ('pagar', 'no')]]
```

se'), ('nose', 'hacer'), ('hacer', 'cuenta'), ('cuenta', '081080622')], [(['hola', 'algui
n']), ('alguien', 'puede'), ('puede', 'asesorar'), ('asesorar', 'app'), ('app', 'deseo'),
('deseo', 'desligar'), ('desligar', 'cuenta'), ('cuenta', 'correo'), ('correo', 'favor')],
[(['reporte', 'cl3999611']), ('cl3999611', '23'), ('23', 'marzo'), ('marzo', 'llamando'),
('llamando', 'dia'), ('dia', 'dia'), ('dia', '12'), ('12', 'dias'), ('dias', 'habiles'),
('habiles', '14'), ('14', 'dias'), ('dias', 'calendario'), ('calendario', 'solicitud'),
('solicitud', 'atencion'), ('atencion', 'parte'), ('parte', 'uds'), ('uds', 'hoy'), ('ho
y'), ('7'), ('7', 'abril'), ('abril', 'aun'), ('aun', 'pueden'), ('pueden', 'presentar')],
[(['paso', 'vas']), ('vas', 'robar'), ('robar', 'dinero'), ('dinero', 'cuentame')], [(['monte
rrey', 'nl']), ('nl', 'cliente'), ('cliente', '097276944'), ('097276944', 'villas'), ('vill
as', 'col'), ('col', 'real'), ('real', 'cumbres'), ('cumbres', '1er'), ('1er', 'sector'),
('sector', 'monterrey'), ('monterrey', 'nl'), ('nl', '64234'), ('64234', 'ojala'), ('ojal
a', 'sirva'), ('sirva', 'solo'), ('solo', 'palabras')], [], [(['mas', 'detalle']), ('detall
e', 'quieres'), ('quieres', 'mande'), ('mande', 'reporte'), ('reporte', 'fecha'), ('fech
a', 'reporte'), ('reporte', 'cliente'), ('cliente', 'direccion'), ('direccion', 'casa'),
('casa', 'mas'), ('mas', 'detalle'), ('detalle', 'necesitas'), ('necesitas', 'claro')],
[(['mantendran', 'informados']), ('informados', 'sinicos'), ('sinicos', 'si'), ('si', 'esta
n'), ('estan', 'cortando'), ('cortando', 'momento'), ('momento', 'servicio'), ('servicio',
'empleados'), ('empleados', 'inmunes'), ('inmunes', 'covid19'), ('covid19', 'ustedes'),
('ustedes', 'aparte'), ('aparte', 'rateros'), ('rateros', 'cobro'), ('cobro', 'abusivo'),
('abusivo', 'servicio'), ('servicio', 'incensibles')], [(['si', 'dices']), ('dices', 'verda
d'), ('verdad', 'fernando'), ('fernando', 'carrizales'), ('carrizales', 'tendria'), ('tend
ria', 'esperar'), ('esperar', '23'), ('23', 'marzo'), ('marzo', 'vengan'), ('vengan', 'rev
isar'), ('revisar', 'suministro'), ('suministro', 'gas'), ('gas', 'domicilio')], [(['pues',
'familia']), ('familia', 'constituyentes'), ('constituyentes', 'san'), ('san', 'nicolas'),
('nicolas', 'corto'), ('corto', 'servicio'), ('servicio', 'gas'), ('gas', 'falta'), ('falt
a', 'pago'), ('pago', 'estan'), ('estan', 'pasando'), ('pasando', 'mal'), ('mal', 'crisi
s'), ('crisis', '650'), ('650', 'falta'), ('falta', '350'), ('350', 'reconexion'), ('recon
exion', 'buscan'), ('buscan', 'algun'), ('algun', 'apoyo'), ('apoyo', 'wsp'), ('wsp', '811
0074884'), ('8110074884', 'tm')], [(['nah', 'paguen']), ('paguen', 'mejor'), ('mejor', 'compr
an'), ('compran', 'tanquesito'), ('tanquesito', 'veran'), ('veran', 'ahorran')], [(['pues',
'familia']), ('familia', 'constituyentes'), ('constituyentes', 'san'), ('san', 'nicolas'),
('nicolas', 'corto'), ('corto', 'servicio'), ('servicio', 'gas'), ('gas', 'falta'), ('falt
a', 'pago'), ('pago', 'estan'), ('estan', 'pasando'), ('pasando', 'mal'), ('mal', 'crisi
s'), ('crisis', '650'), ('650', 'falta'), ('falta', '350'), ('350', 'reconexion'), ('recon
exion', 'buscan'), ('buscan', 'algun'), ('algun', 'apoyo'), ('apoyo', 'wsp'), ('wsp', '811
0074884'), ('8110074884', 'tm')], [(['apoya', 'ayd']), ('ayd', 'cobres'), ('cobres', 'apoy
e'), ('apoyen', 'renuncien'), ('renuncien', 'sua'), ('sua', 'prerrogativas'), ('prerrogati
vas', 'apoyo'), ('apoyo', 'sociedad')], [(['tendra', 'prorroga']), ('prorroga', 'pago'), ('p
ago', 'contingencia'), ('contingencia', 'sanitaria'), ('sanitaria', 'coronavirus')], [(['me
didor', 'hora']), ('hora', 'marca'), ('marca', '5613'), ('5613', 'recibo'), ('recibo', 'dic
e'), ('dice', '5685'), ('5685', 'lectura'), ('lectura', 'real'), ('real', 'buenas'), ('bue
nas', 'tomas'), ('tomas', 'lectura'), ('lectura', 'literal'), ('literal', 'fraude'), ('fra
ude', 'robo')], [(['cuantos', 'asi']), ('asi', 'llevan'), ('llevan', 'dos'), ('dos', 'mes
e'), ('meses')], [(['pues', 'familia']), ('familia', 'constituyentes'), ('constituyentes', 'san'), ('sa
n', 'nicolas'), ('nicolas', 'corto'), ('corto', 'servicio'), ('servicio', 'gas'), ('gas',
'falta'), ('falta', 'pago'), ('pago', 'estan'), ('estan', 'pasando'), ('pasando', 'mal'),
('mal', 'crisis'), ('crisis', '650'), ('650', 'falta'), ('falta', '350'), ('350', 'reconnex
ion'), ('reconexion', 'buscan'), ('buscan', 'algun'), ('algun', 'apoyo'), ('apoyo', 'ws
p'), ('wsp', '8110074884'), ('8110074884', 'tm')], [(['pues', 'familia']), ('familia', 'const
ituyentes'), ('constituyentes', 'san'), ('san', 'nicolas'), ('nicolas', 'corto'), ('cort
o', 'servicio'), ('servicio', 'gas'), ('gas', 'falta'), ('falta', 'pago'), ('pago', 'esta
n'), ('estan', 'pasando'), ('pasando', 'mal'), ('mal', 'crisis'), ('crisis', '650'), ('65
0', 'falta'), ('falta', '350'), ('350', 'reconexion'), ('reconexion', 'buscan'), ('busca
n', 'algun'), ('algun', 'apoyo'), ('apoyo', 'wsp'), ('wsp', '8110074884'), ('8110074884',
'tm')], [(['cinismo', 'puro']), ('puro', 'corte'), ('corte', 'masivo'), ('masivo', 'falta'),
('falta', 'pago'), ('pago', 'recibos'), ('recibos', 'gas'), ('gas', 'natural'), ('natural',
'pandemia'), ('pandemia', 'mexico'), ('mexico', 'parte'), ('parte', 'apoyo'), ('apoy
o', 'iniciativa'), ('iniciativa', 'privada'), ('privada', 'sociedad')], [(['abusivos', 'est
an']), ('estan', 'cortando'), ('cortando', 'servicio'), ('servicio', 'solo'), ('solo',
'7'), ('7', 'dias'), ('dias', 'atraso'), ('atraso', 'pago'), ('pago', 'ningun'), ('ningu
n', 'aviso'), ('aviso', 'ahora'), ('ahora', 'voy'), ('voy', 'tener'), ('tener', 'pagar'),
('pagar', '300'), ('300', 'reconectararlo'), ('reconectararlo', 'ustedes'), ('ustedes', 'empre
sas'), ('empresas', 'mas'), ('mas', 'abusivas'), ('abusivas', 'pais')], [(['pues', 'app'),
('app', 'apesta'), ('apesta', 'pidio'), ('pidio', 'crear'), ('crear', 'contraseña'), ('con
traseña', 'registro'), ('registro', 'llene'), ('llene', 'campos'), ('campos', 'cero'), ('c
ero', 'llene')]]

(['calentar', 'agua'], ['agua', 'resistencia']), [(['tramites', 'interminables'], ('interminables', 'documentos')), ('documentos', 'inútiles'), ('inútiles', 'tiempos'), ('tiempos', 'espera'), ('espera', 'larguisimos'), ('larguisimos', ' posible'), ('possible', 'dia'), ('dia', 'hoy'), ('hoy', 'despues'), ('despues', 'dos'), ('dos', 'semanas'), ('semanas', 'intentos'), ('intentos', 'inútiles'), ('inútiles', 'podido'), ('podido', 'reconectar'), ('reconectar', 'gas'), ('gas', 'peor'), ('peor', 'ser'), ('ser', 'monopolio'), ('monopolio', 'puedo'), ('puedo', 'contratar'), ('contratar', 'compañia')], [(['cuentan', 'dos'], ('dos', 'semanas'), ('semanas', 'puedan'), ('puedan', 'reconectar'), ('reconectar', 'servicio'), ('servicio', 'medio'), ('medio', 'pandemia'), ('pandemia', 'poder'), ('poder', 'utilizar'), ('utilizar', 'gas'), ('gas', 'van'), ('van', 'resolver')], [], [(['duda', '¿esta'], ('¿esta', 'vez'), ('vez', 'distribuyeron'), ('distribuyeron', 'recibos'), ('recibos', 'verdad'), ('verdad', 'nunca'), ('nunca', 'llego'), ('llego', 'casa'), ('casa', 'primera'), ('primera', 'vez'), ('vez', 'llega')], [(['hola', 'dan'], ('dan', 'naturgy'), ('naturgy', 'mexico'), ('mexico', 'servirte'), ('servirte', 'puedes'), ('puedes', 'contactarnos'), ('contactarnos', 'via'), ('via', 'dm'), ('dm', 'si'), ('si', 'dudas'), ('dudas', 'acerca'), ('acerca', 'recibos'), ('recibos', 'gusto'), ('gusto', 'haremos'), ('haremos', 'revision'), ('revision', 'correspondiente'), ('correspondiente', 'solo'), ('solo', 'debes'), ('debes', 'enviarnos'), ('enviarnos', 'siguentes'), ('siguentes', 'datos'), ('datos', 'número'), ('número', 'cuenta'), ('cuenta', 'nombre'), ('nombre', 'titular'), ('titular', 'localidad'), ('localidad', 'saludoslg')], [(['gas', 'natural'], ('natural', 'deberia'), ('deberia', 'omitir'), ('omitir', 'cobro'), ('cobro', 'bimestre'), ('bimestre', 'apoyo'), ('apoyo', 'pandemia'), ('pandemia', 'justo'), ('justo', 'despues'), ('despues', 'cobra'), ('cobra', 'merma'), ('merma', 'cobra'), ('cobra', 'exigir'), ('exigir', 'apoye'), ('apoye', 'mexico')], [(), (), [(['buenos', 'dia'], ('dias', 'cdmx'), ('cdmx', 'cuenta'), ('cuenta', '045491054'), ('045491054', '¿hay'), ('¿hay', 'algún'), ('algún', 'problema'), ('problema', 'zona'), ('zona', 'baja'), ('baja', 'presion'), ('presion', 'estufa'), ('estufa', 'calentador'), ('calentador', 'agua')]]

```
In [12]: # Convert bigrams list to bigrams dataframe  
df = pd.DataFrame(bigrams)
```

```
In [13]: # Show dataframe  
df
```

Out[13]:	0	1	2	3	4	5	6	7	8	
0	(cinismo, puro)	(puro, corte)	(corte, masivo)	(masivo, falta)	(falta, pago)	(pago, recibos)	(recibos, gas)	(gas, natural)	(natural, pandemia)	(pander mex
1	(cinismo, puro)	(puro, corte)	(corte, masivo)	(masivo, falta)	(falta, pago)	(pago, recibos)	(recibos, gas)	(gas, natural)	(natural, pandemia)	(pander mex
2	(gracias, mande)	(mande, dm)	None	None	None	None	None	None	None	N
3	(buenas, tardes)	(tardes, cortaron)	(cortaron, servicio)	(servicio, enblas)	(enblas, oficinas)	(oficinas, pagar)	(pagar, nose)	(nose, hacer)	(hacer, cuenta)	081080€
4	(hola, alguien)	(alguien, puede)	(puede, asesorar)	(asesorar, app)	(app, deseo)	(deseo, desligar)	(desligar, cuenta)	(cuenta, correo)	(correo, favor)	N
...
1395	(gas, natural)	(natural, deberia)	(deberia, omitir)	(omitir, cobro)	(cobro, bimestre)	(bimestre, apoyo)	(apoyo, pandemia)	(pandemia, justo)	(justo, despues)	(desp col
1396	(gas, natural)	(natural, deberia)	(deberia, omitir)	(omitir, cobro)	(cobro, bimestre)	(bimestre, apoyo)	(apoyo, pandemia)	(pandemia, justo)	(justo, despues)	(desp col
1397	None	None	None	None	None	None	None	None	None	N
1398	None	None	None	None	None	None	None	None	None	N
1399	(buenos, dias)	(dias, cdmx)	(cdmx, cuenta)	(cuenta, 045491054)	(045491054, ¿hay)	(¿hay, algún)	(algún, problema)	(problema, zona)	(zona, baja)	(b pres

1400 rows × 33 columns

In [14]:

```
# Convert dataframe to just two columns
all_values = []
for column in df:
    this_column_values = df[column].tolist()
    all_values += this_column_values

df = pd.DataFrame(all_values)
df
```

Out[14]:

	0	1
0	cinismo	puro
1	cinismo	puro
2	gracias	mande
3	buenas	tardes
4	hola	alguien
...
46195	None	None
46196	None	None
46197	None	None
46198	None	None
46199	None	None

46200 rows × 2 columns

In [15]:

```
# Remove nulls
df = df.dropna()
df
```

Out[15]:

	0	1
0	cinismo	puro
1	cinismo	puro
2	gracias	mande
3	buenas	tardes
4	hola	alguien
...
44088	situacion	ajena
44091	situacion	ajena
44785	exploto	hdspm
45488	ajena	cliente
45491	ajena	cliente

17419 rows × 2 columns

In [16]:

```
# Add size column to get bigrams frequency
df = df.groupby(df.columns.tolist(),as_index=False).size()
df
```

Out[16]:

	0	1	size
0	0	8	1
1	0	ahi	1
2	0	criterio	1
3	0	funciona	1
4	0	van	1
...
9977	●●●menos	rateros	4
9978	💡	ayudan	1
9979	♂	tener	1
9980	♀	viva	1
9981	😊	puedo	1

9982 rows × 3 columns

In [17]:

```
# Find the most significant word associations
df = df.sort_values(by ='size', ascending=False)
df.head(13)
```

Out[17]:

	0	1	size
3983	gas	natural	67
5881	naturgy	mexico	53
1245	buen	dia	53
4960	llamanos		55
2196	cortaron	gas	37
1262	buenos	dias	33
1252	buenas	tardes	33
7246	puede	salir	33
6799	pesimo	servicio	32
6897	plena	contingencia	31
9659	via	dm	31
3863	fuga	gas	30
8404	servicio	gas	29

Conclusiones

De acuerdo con Fano (1961, como se citó en Church y Hanks, 1989), si dos puntos (palabras), x y y, tienen probabilidades P(x) y P(y), entonces su información mutua, I(x,y) se define como:

$$I(x,y) = \log_2 P(x,y) / P(x) P(y)$$

Entonces, la información mutua compara la probabilidad de observar x y y juntas (la probabilidad de unión) con las probabilidades de observarlas independientemente.

Las probabilidades $P(x)$ y $P(y)$ son estimadas contando el número de observaciones de x y y en un corpus, $f(x)$ y $f(y)$, y normalizando por N, el tamaño del corpus. Las probabilidades de unión, $P(x,y)$, son estimadas contando el número de veces que x es seguida por y en una ventana de w palabras, $f_w(x,y)$, y normalizando por N.

En este sentido, podemos concluir que la medida de información mutua es equivalente a la columna "size" ya que el total de datos es casi de 10,000.

A partir de este ejercicio, me doy cuenta de que es muy importante para un estudiante de Ciencia de Datos desarrollar habilidades para limpiar y preprocesar textos, ya que estas tareas son las que más tiempo consumen.

Una encuesta de CrowdFlower, proveedor de una plataforma de "enriquecimiento de datos" mostró que los científicos de datos pasan la mayor parte de su tiempo masajeando en lugar de extrayendo o modelando datos (Press, 2016).

Por ello, considero imperativo que me dedique los próximos meses, a la par de mis actividades laborales y escolares, a practicar mucho el preprocesamiento de datos.

Referencias Bibliográficas

Church, K. W. and Hanks, P. (1989) Word Association Norms, Mutual Information, and Lexicography. 27th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 1989

Press, G. (2016) Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Mar 23, 2016