INSTITUTO POLITECNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO

“Practica 2 – Expresiones regulares y Vectorización de texto”

-Tinoco Videgaray Sergio Ernesto

Grupo: 5BV1

Materia: Tecnologías del Lenguaje Natural

10/10/23

- Introducción.

En esta práctica se van a aplicar las expresiones regulares a diferentes documentos de texto, así como también aplicar la vectorización de documentos.

- Desarrollo.

Primeramente, se van a aplicar las expresiones regulares.

- **Expresiones regulares.**

En esta sección de la práctica se van a aplicar las siguientes expresiones regulares para el documento mostrado en la figura 1.

```
Amount:Category:Date:Description
5.25:supply:20170222:box of staples
79.81:meal:20170222:lunch with ABC Corp. clients Al, Bob, and Cy
43.00:travel:20170222:cab back to office
383.75:travel:20170223:flight to Boston, to visit ABC Corp.
55.00:travel:20170223:cab to ABC Corp. in Cambridge, MA
23.25:meal:20170223:dinner at Logan Airport
318.47:supply:20170224:paper, toner, pens, paperclips, tape
142.12:meal:20170226:host dinner with ABC clients, Al, Bob, Cy, Dave, Ellie
303.94:util:20170227:Peoples Gas
121.07:util:20170227:Verizon Wireless
7.59:supply:20170227:Python book (used)
```

Figura 1. Documento de gastos.

1. A continuación, se aplica una expresión regular para identificar las líneas que contengan una “r” seguida por una “g” no necesariamente en posiciones seguidas (figura 2).

```
Matchs en RegEx 1
ravel:20170223:flig
ravel:20170223:cab to ABC Corp. in Cambridg
r at Log
rive to Big
ravel:20170304:mileage, drive to/from Big
ravel:20170304:Hyatt Hotel, Reston VA, for Big Inc. meeting
ractical G
ravel:20170317:mileage, drive to/from Big
red and Gina, Big
```

Figura 2. Impresión en consola de las líneas que contengan una “r” seguida por una “g”.

2. Se buscan las líneas que describan comidas que cuesten al menos 100 (figura 3 y 4).

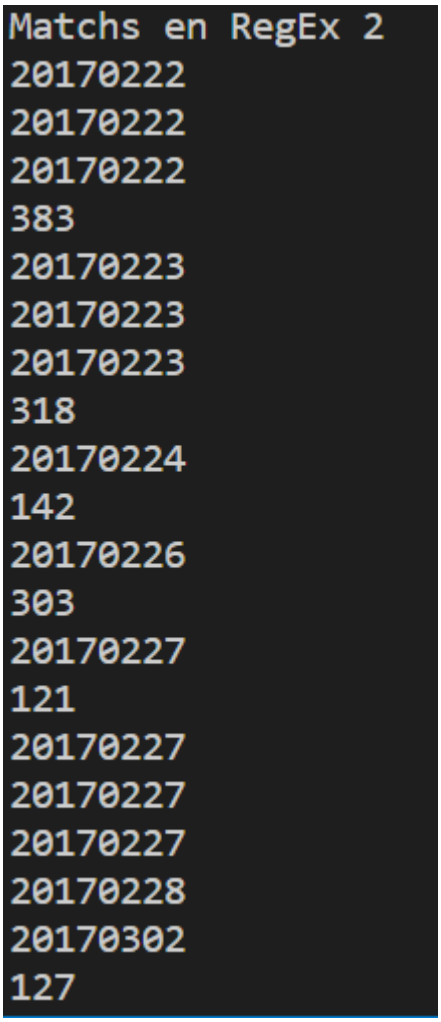

Figura 3. Impresión en consola de las líneas que describan comidas que cuesten al menos 100.

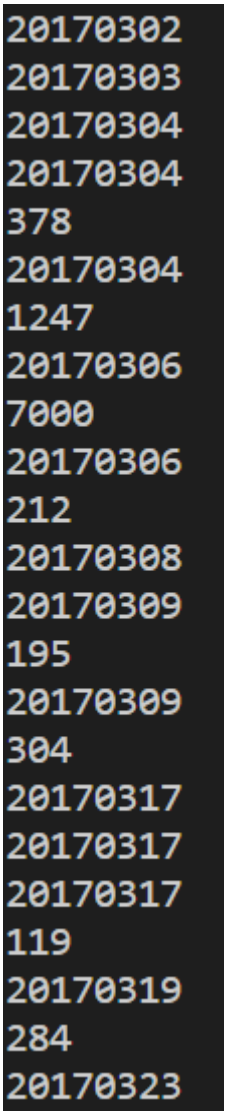

Figura 4. Impresión en consola de las líneas que describan comidas que cuesten al menos 100.

3. Que contengan una “a”, seguida por una “b”, seguida por una “c” (puede haber otros caracteres entre la “a” y la “b” y entre la “b” y la ”c”, de ser el caso, estos caracteres no pueden ser a ni b, ni b y c, respectivamente (figura 5).

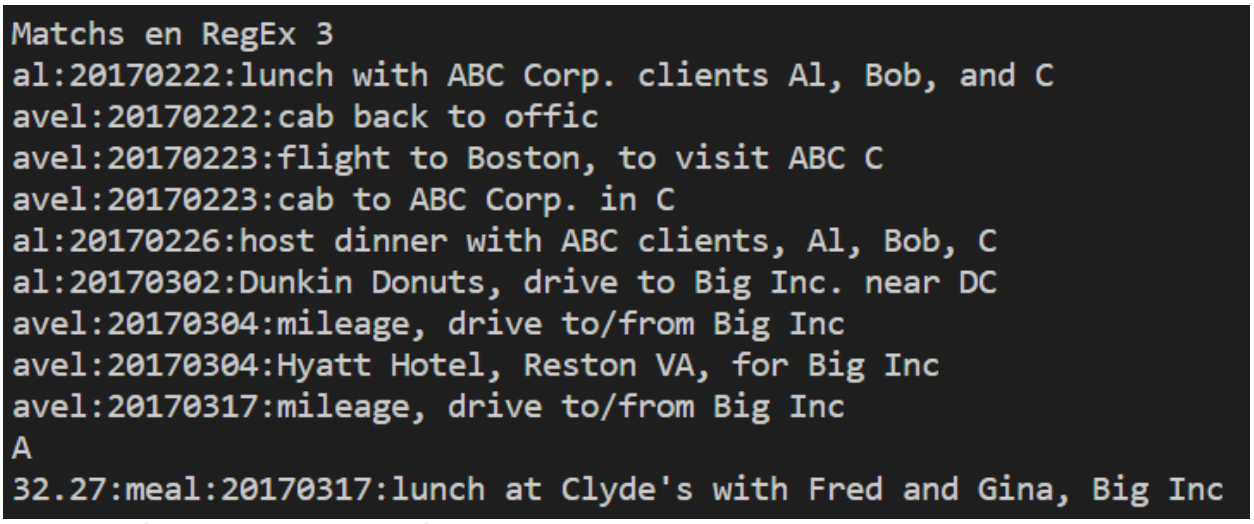

Figura 5. Impresión en consola de las líneas que contengan una “a”, seguida por una “b”, seguida por una “c”.

4. Contengan en la descripción de gastos una “a” minúscula y un digito entre 0 y 9 en cualquier orden. Es decir, el carácter “a” puede aparecer antes o después del digito (figura 6).

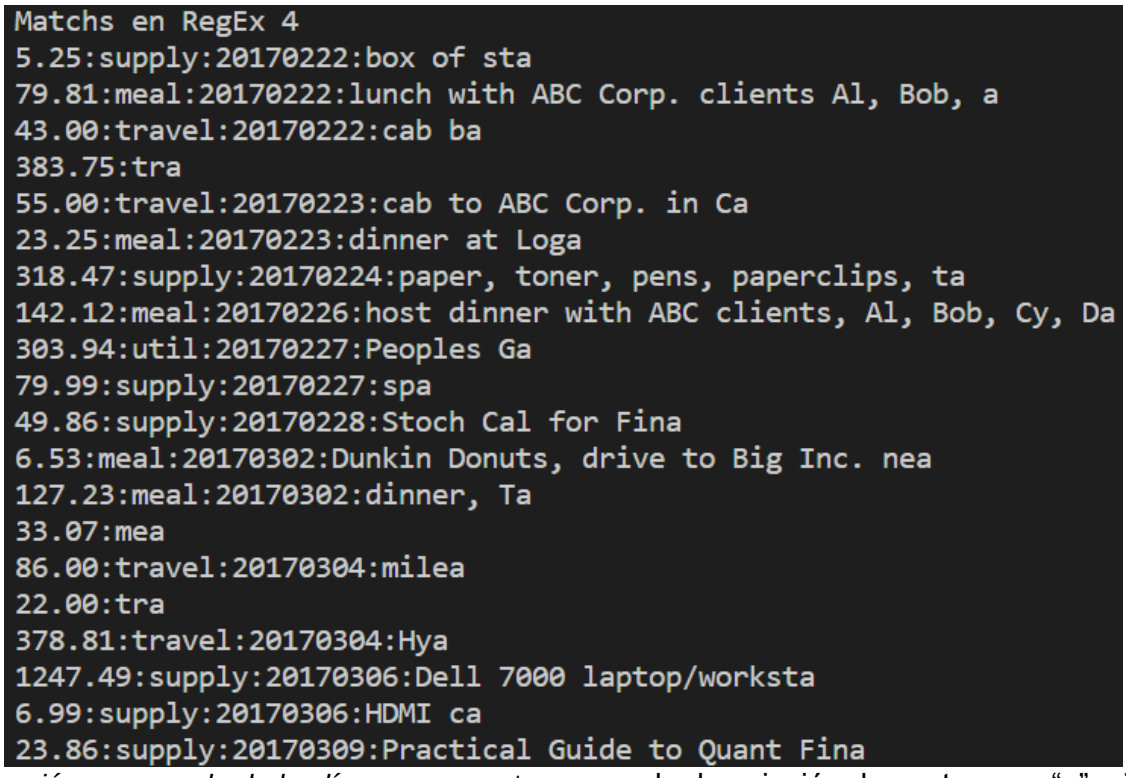

Figura 6. Impresión en consola de las líneas que contengan en la descripción de gastos una “a” minúscula y un digito entre 0 y 9 en cualquier orden.

5. Contengan el carácter “d”, posiblemente seguido de otros caracteres, seguido de una “i”. Coincidencias incluirían palabras tales como: diver, doily, drip, diplomat, etc (figura 7).

```
Matchs en RegEx 5
dinner at Logan Ai
dinner with ABC clients, Al, Bob, Cy, Dave, Elli
drive to Bi
di
dinner, Uncle Juli
drive to/from Bi
de to Quant Finance Intervi
drive to/from Bi
de's with Fred and Gina, Bi
```

Figura 7. Impresión en consola de las líneas que contengan el carácter “d”, posiblemente seguido de otros caracteres, seguido de una “i”.

6. Usando expresiones regulares se van a identificar títulos de películas producidas antes de 2002 de la siguiente lista; (figura 8).

a. The Shawshank Redemption (1994)
b. The Godfather (1972)
c. The Godfather: Part II (1974)
d. 2001: A Space Odyssey (1968)
e. The Good, the Bad and the Ugly (1966)
f. Angry Men (1957)
g. Schindler's List (1993)
h. The Lord of the Rings: The Return of the King (2003)
i. Fight Club (1999)
j. 2010: The Year We Make Contact (1984)
k. 101 Dalmatians (1996)

```
Peliculas estrenadas antes del 2002
The Shawshank Redemption
The Godfather
The Godfather: Part II
2001: A Space Odyssey
The Good, the Bad and the Ugly
Angry Men
Schindler's List
Fight Club
2010: The Year We Make Contact
101 Dalmatians
```

Figura 8. Impresión en consola de los títulos de películas estrenadas antes del 2002.

7. Identificar recetas que no contengan la palabra 'chocolate' de la siguiente lista (figura 9).

a. Cake 1: sugar, flour, cocoa powder, baking powder, baking soda, salt, eggs, milk, vegetable oil, vanilla extract, chocolATE chip.
b. Cake 2: cream cheese, sugar, vanilla extract, crescent rolls, cinnamon, butter, honey.
c. Cake 3: dark chocolate cake mix, instant CHOCOLATE pudding mix, sour cream, eggs, vegetable oil, coffee liqueur.

d. Cake 4: flour, baking powder, salt, cinnamon, butter, sugar, egg, vanilla extract, milk, chopped walnuts.
e. Cake 5: gingersnap cookies, chopped pecans, butter, cream cheese, sugar, vanilla extract, eggs, canned pumpkin, cinnamon, CHOColate.
f. Cake 6: flour, baking soda, sea salt, butter, white sugar, brown sugar, eggs, vanilla extract, Chocolate chips, canola oil.
g. Cake 7: wafers, cream cheese, sugar, eggs, vanilla extract, cherry pie filling.

```
Recetas que no contenan 'chocolate'
Cake 2: cream cheese, sugar, vanilla extract, crescent rolls, cinnamon, butter, honey.
Cake 4: flour, baking powder, salt, cinnamon, butter, sugar, egg, vanilla extract, milk, chopped walnuts.
Cake 7: wafers, cream cheese, sugar, eggs, vanilla extract, cherry pie filling.
```

Figura 9. Impresión en consola de las recetas que no contengan la palabra chocolate.

8. Comas entre grupos de cada tres dígitos para las siguientes poblaciones por país (figura 10).

a. China 1361220000
b. India 1236800000
c. United States 317121000
d. Indonesia 237641326
e. Brazil 201032714
f. Pakistan 184872000
g. Nigeria 173615000
h. Bangladesh 152518015
i. Russia 143600000

```
Poblaciones con formato de comas
('China', '1,361,220,000')
('India', '1,236,800,000')
('United States', '317,121,000')
('Indonesia', '237,641,326')
('Brazil', '201,032,714')
('Pakistan', '184,872,000')
('Nigeria', '173,615,000')
('Bangladesh', '152,518,015')
('Russia', '143,600,000')
```

Figura 10. Impresión en consola de las poblaciones de cada país separando con comas cada grupo de 3 digitos.

9. Simplificar direcciones IPv6.

Una dirección IP está compuesta de 8 bloques de números hexadecimales. Los bloques son separados por el signo “:” y contienen cuatro dígitos. Existen además ciertas reglas para simplificar y reducir el tamaño de una dirección IPv6. Dichas reglas son:

a) los bloques compuestos de puros pueden ser omitidos.
b) Los ceros al principio de un bloque pueden ser omitidos.

Se van a abreviar las siguientes direcciones IPV6 como se muestra en la figura 11.

2001:0db8:0000:0000:0000:ff00:0042:8329
2607:f0d0:1002:0051:0000:0000:0000:0004
2001:0db8:3c4d:0015:0000:0000:1a2f:1a2b

```
IPV6 abreviadas
2001:0db8:0000:0000:0000:ff00:0042:8329      |      2001:db8::ff00:42:8329
2607:f0d0:1002:0051:0000:0000:0000:0004      |      2607:f0d0:1002:51::4
2001:0db8:3c4d:0015:0000:0000:1a2f:1a2b      |      2001:db8:3c4d:15::1a2f:1a2b
```

Figura 11. Impresión en consola de las IPV6 con sus respectivas abreviaciones.

PARTE 2. VECTORIZACIÓN DE DOCUMENTOS

En esta sección se van a generar las representaciones numéricas de cada uno de los siguientes documentos mostrados en la tabla 1:

Doc. ID	Clinical Statement (Before pre-processing)
1	Pancreatic cancer with metastasis. Jaundice with transaminitis, evaluate for obstruction process.
2	Pancreatitis. Breast cancer. No output from enteric tube. Assess tube.
3	Metastasis pancreatic cancer. Acute renal failure, evaluate for hydronephrosis or obstructive uropathy.

Tabla 1. Documentos para analizar.

Primeramente, se van a normalizar los documentos para generar una representación numérica de cada termino adecuada.

Normalización del documento 1 (figura 12).

```
Normalizacion del documento 1

Documento inicial:
 Pancreatic cancer with metastasis. Jaundice with transaminitis, evaluate for obstruction process.
```

Figura 12. Impresión en consola del documento 1 antes de realizar la normalización.

a) Se convierten las mayúsculas a minúsculas, se obtienen los tokens iniciales y se remueven tanto stop-words como puntos y comas antes de realizar alguna técnica de stemming o lematización (figuras 13 a 16).

```
Documento en minusculas:
 pancreatic cancer with metastasis. jaundice with transaminitis, evaluate for obstruction process.
```

Figura 13. Impresión en consola del documento 1 al cambiar las mayúsculas a minúsculas.

```
Tokens iniciales:
 ['pancreatic', 'cancer', 'with', 'metastasis', '.', 'jaundice', 'with', 'transaminitis', ',', 'evaluate', 'for', 'obstruction', 'pro
cess', '.']
```

Figura 14. Impresión en consola de los tokens iniciales del documento 1.

```
Tokens despues de remover stop-words:
 ['pancreatic', 'cancer', 'metastasis', '.', 'jaundice', 'transaminitis', ',', 'evaluate', 'obstruction', 'process', '.']
```

Figura 15. Impresión en consola de los tokens del documento 1 después de remover las stop-words.

```
Tokens despues de remover puntos y comas:
 ['pancreatic', 'cancer', 'metastasis', 'jaundice', 'transaminitis', 'evaluate', 'obstruction', 'process']
```

Figura 16. Impresión en consola de los tokens del documento 1 después de remover puntos y comas.

b) Se aplica la técnica de stemming a los tokens obtenidos en el punto anterior (figuras 17 y 18).

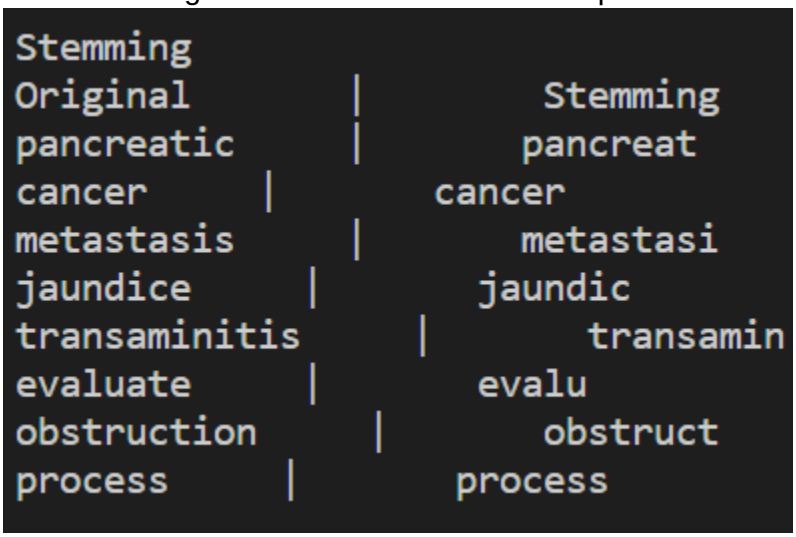

Figura 17. Impresión en consola de los tokens en su forma original y stemming.

```
Tokens despues de aplciar el stemming:
 ['pancreat', 'cancer', 'metastasi', 'jaundic', 'transamin', 'evalu', 'obstruct', 'process']
```

Figura 18. Impresión en consola de los tokens del documento 1 después de aplicar el stemming.

c) Se aplica la técnica de POS-tagging para obtener las etiquetas correspondientes de cada token previo a la normalización (figura 19).

```
Tokens con el POS-tagging:
 [('pancreat', 'n'), ('cancer', 'n'), ('metastasi', 'n'), ('jaundic', 'a'), ('transamin', 'n'), ('evalu', 'a'), ('obstruct', 'n'),
process', 'n')]
```

Figura 19. Impresión en consola de los tokens junto con su tag del documento 1 al aplicar el POS-tagging.

d) Se aplica la lematización utilizando las etiquetas asignadas a cada token (figura 20).

```
Tokens despues de lematizar:
 ['pancreat', 'cancer', 'metastasi', 'jaundic', 'transamin', 'evalu', 'obstruct', 'process']
```

Figura 20. Impresión en consola de los tokens del documento 1 después de aplicar la lematización.

Normalización del documento 2 (figura 21).

```
Normalizacion del documento 2

Documento inicial:
 Pancreatitis. Breast cancer. No output from enteric tube. Assess tube.
```

Figura 21. Impresión en consola del documento 2 antes de realizar la normalización.

e) Se convierten las mayúsculas a minúsculas, se obtienen los tokens iniciales y se remueven tanto stop-words como puntos y comas antes de realizar alguna técnica de stemming o lematización (figuras 22 a 25).

```
Documento en minusculas:
 pancreatitis. breast cancer. no output from enteric tube. assess tube.
```

Figura 22. Impresión en consola del documento 2 al cambiar las mayúsculas a minúsculas.

```
Tokens iniciales:
 ['pancreatitis', '.', 'breast', 'cancer', '.', 'no', 'output', 'from', 'enteric', 'tube', '.', 'assess', 'tube', '.']
```

Figura 23. Impresión en consola de los tokens iniciales del documento 2.

```
Tokens despues de remover stop-words:
 ['pancreatitis', '.', 'breast', 'cancer', '.', 'output', 'enteric', 'tube', '.', 'assess', 'tube', '.']
```

Figura 24. Impresión en consola de los tokens del documento 2 después de remover las stop-words.

```
Tokens despues de remover puntos y comas:
 ['pancreatitis', 'breast', 'cancer', 'output', 'enteric', 'tube', 'assess', 'tube']
```

Figura 25. Impresión en consola de los tokens del documento 2 después de remover puntos y comas.

f) Se aplica la técnica de stemming a los tokens obtenidos en el punto anterior (figuras 26 y 27).

```
Stemming
Original      |       Stemming
pancreatitis     |       pancreat
breast     |       breast
cancer     |       cancer
output     |       output
enteric     |       enter
tube     |       tube
assess     |       assess
tube     |       tube
```

Figura 26. Impresión en consola de los tokens en su forma original y stemming.

```
Tokens despues de aplciar el stemming:
 ['pancreat', 'breast', 'cancer', 'output', 'enter', 'tube', 'assess', 'tube']
```

Figura 27. Impresión en consola de los tokens del documento 2 después de aplicar el stemming.

g) Se aplica la técnica de POS-tagging para obtener las etiquetas correspondientes de cada token previo a la normalización (figura 28).

```
Tokens con el POS-tagging:
 [('pancreat', 'n'), ('breast', 'n'), ('cancer', 'n'), ('output', 'n'), ('enter', 'v'), ('tube', 'n'), ('assess', 'v'), ('tube', 'n')
]
```

Figura 28. Impresión en consola de los tokens junto con su tag del documento 2 al aplicar el POS-tagging.

h) Se aplica la lematización utilizando las etiquetas asignadas a cada token (figura 29).

```
Tokens despues de lematizar:
 ['pancreat', 'breast', 'cancer', 'output', 'enter', 'tube', 'assess', 'tube']
```

Figura 29. Impresión en consola de los tokens del documento 2 después de aplicar la lematización.

Normalización del documento 3 (figura 30).

```
Normalizacion del documento 3

Documento inicial:
 Metastasis pancreatic cancer. Acute renal failure, evaluate for hydronephrosis or obstructive uropathy.
```

Figura 30. Impresión en consola del documento 3 antes de realizar la normalización.

i) Se convierten las mayúsculas a minúsculas, se obtienen los tokens iniciales y se remueven tanto stop-words como puntos y comas antes de realizar alguna técnica de stemming o lematización (figuras 31 a 34).

```
Documento en minusculas:
 metastasis pancreatic cancer. acute renal failure, evaluate for hydronephrosis or obstructive uropathy.
```

Figura 31. Impresión en consola del documento 3 al cambiar las mayúsculas a minúsculas.

```
Tokens iniciales:
 ['metastasis', 'pancreatic', 'cancer', '.', 'acute', 'renal', 'failure', ',', 'evaluate', 'for', 'hydronephrosis', 'or', 'obstructiv
e', 'uropathy', '.']
```

Figura 32. Impresión en consola de los tokens iniciales del documento 3.

```
Tokens despues de remover stop-words:
 ['metastasis', 'pancreatic', 'cancer', '.', 'acute', 'renal', 'failure', ',', 'evaluate', 'hydronephrosis', 'obstructive', 'uropathy
', '.']
```

Figura 33. Impresión en consola de los tokens del documento 3 después de remover las stop-words.

```
Tokens despues de remover puntos y comas:
 ['metastasis', 'pancreatic', 'cancer', 'acute', 'renal', 'failure', 'evaluate', 'hydronephrosis', 'obstructive', 'uropathy']
```

Figura 34. Impresión en consola de los tokens del documento 3 después de remover puntos y comas.

j) Se aplica la técnica de stemming a los tokens obtenidos en el punto anterior (figuras 35 y 36).

```
Stemming
Original        |        Stemming
metastasis      |       metastasi
pancreatic      |       pancreat
cancer      |       cancer
acute       |       acut
renal       |       renal
failure     |       failur
evaluate        |       evalu
hydronephrosis      |       hydronephrosi
obstructive     |       obstruct
uropathy        |       uropathi
```

Figura 35. Impresión en consola de los tokens en su forma original y stemming.

```
Tokens despues de aplciar el stemming:
 ['metastasi', 'pancreat', 'cancer', 'acut', 'renal', 'failur', 'evalu', 'hydronephrosi', 'obstruct', 'uropathi']
```

Figura 36. Impresión en consola de los tokens del documento 3 después de aplicar el stemming.

k) Se aplica la técnica de POS-tagging para obtener las etiquetas correspondientes de cada token previo a la normalización (figura 37).

```
Tokens con el POS-tagging:
 [('metastasi', 'r'), ('pancreat', 'n'), ('cancer', 'n'), ('acut', 'v'), ('renal', 'a'), ('failur', 'n'), ('evalu', 'n'), ('hydronephros
i', 'n'), ('obstruct', 'n'), ('uropathi', 'n')]
```

Figura 37. Impresión en consola de los tokens junto con su tag del documento 3 al aplicar el POS-tagging.

l) Se aplica la lematización utilizando las etiquetas asignadas a cada token (figura 38).

```
Tokens despues de lematizar:
 ['metastasi', 'pancreat', 'cancer', 'acut', 'renal', 'failur', 'evalu', 'hydronephrosi', 'obstruct', 'uropathi']
```

Figura 38. Impresión en consola de los tokens del documento 3 después de aplicar la lematización.

1. Se genera el diccionario de términos extraídos de los documentos normalizados (figura 39).

```
['pancreat', 'cancer', 'metastasi', 'jaundic', 'transamin', 'evalu', 'obstruct', 'process', 'breast', 'output', 'enter', 'tube', 'assess
', 'acut', 'renal', 'failur', 'hydronephrosi', 'uropathi']
```

Figura 39. Impresión en consola del diccionario generado a partir de los documentos.

Se generan los vectores para representar numéricamente cada documento utilizando cada una de las siguientes técnicas:

One Hot Encoding o “Term Presence”.
Se genera un vector de valores binarios, se asigna un 1 si el termino se encuentra en el documento y 0 en caso contrario (figura 40).

```
Vectores de Term Presence
Documento 1
 [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Documento 2
 [1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]
Documento 3
 [1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
```

Figura 40. Impresión en consola de los vectores term presence de cada documento.

Cantidad de términos o “Term Count”.
Se genera un vector con el numero de veces que se encontró el termino en el documento (figura 41).

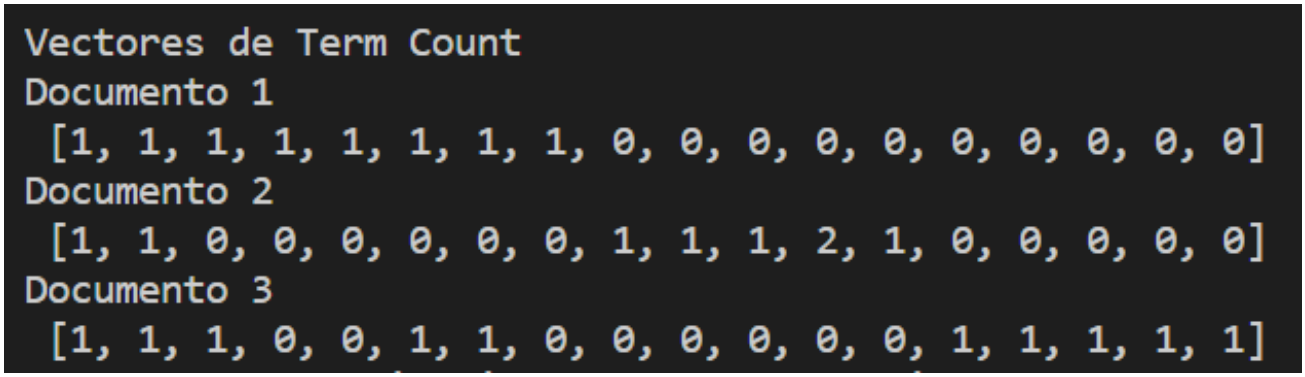

Figura 41. Impresión en consola de los vectores term count de cada documento.

Probabilidad de término.
Se genera un vector de probabilidades de cada termino de acuerdo con la ecuación de la ecuación 1 (figura 42).

$$P(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ corpus}$$

Ecuación 1. Probabilidad de términos.

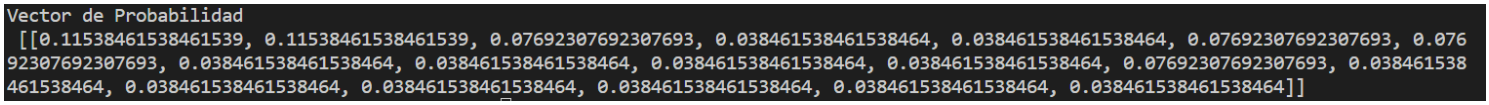

Figura 42. Impresión en consola del vector de probabilidad de cada termino.

Frecuencia de términos o “Term Frequency (TF)”.
Se genera un vector de la frecuencia de términos por cada documento con base en la ecuación 2 (figura 43).

$$TF = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

Ecuación 2. Frecuencia de Término.

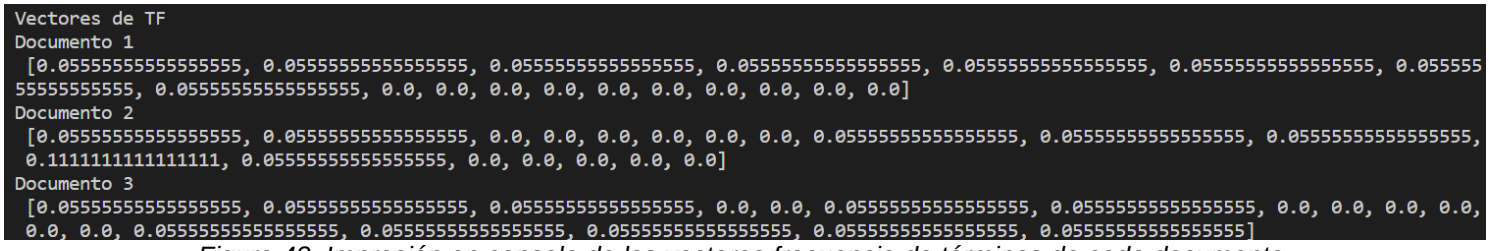

Figura 43. Impresión en consola de los vectores frecuencia de términos de cada documento.

Frecuencia inversa de documentos “Inverse Document Frequency (IDF)”.
Se genera un vector de la frecuencia inversa de cada documento siguiendo la ecuación 3 (figura 44).

$$IDF = log \frac{Number\ of\ documents\ in\ corpus}{Number\ of\ documentos\ where\ term\ appers}$$

Ecuación 3. Frecuencia Inversa del documento.

```
Vector de IDF
 [0.0, 0.0, 0.17609125905568124, 0.47712125471966244, 0.47712125471966244, 0.17609125905568124, 0.17609125905568124, 0.47712125471966244
, 0.47712125471966244, 0.47712125471966244, 0.47712125471966244, 0.47712125471966244, 0.47712125471966244, 0.47712125471966244, 0.477121
25471966244, 0.47712125471966244, 0.47712125471966244, 0.47712125471966244]
```

. *Figura 44. Impresión en consola de los vectores de la frecuencia inversa de cada termino.*

TDF-IDF.

Se genera un vector haciendo el producto entre la frecuencia de términos y la frecuencia inversa de cada documento (figura 45).

```
Vectores de TF-IDF
Documento 1
 [0.0, 0.0, 0.009782847725315624, 0.026506736373314577, 0.026506736373314577, 0.009782847725315624, 0.009782847725315624, 0.026506736373
314577, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
Documento 2
 [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.026506736373314577, 0.026506736373314577, 0.026506736373314577, 0.05301347274662915, 0.02650
6736373314577, 0.0, 0.0, 0.0, 0.0, 0.0]
Documento 3
 [0.0, 0.0, 0.009782847725315624, 0.0, 0.0, 0.009782847725315624, 0.009782847725315624, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0265067363733145
77, 0.026506736373314577, 0.026506736373314577, 0.026506736373314577, 0.026506736373314577]
```

. *Figura 45. Impresión en consola de los vectores TF-IDF de cada documento.*

Análisis y conclusiones

En esta práctica se abordaron temas como las expresiones regulares que nos ayudan a realizar búsquedas en una cadena de texto por medio de símbolos con el fin de realizar búsquedas filtradas, añadir subcadenas como el caso de las comas que separan grupos de 3 dígitos en el caso de las poblaciones de cada país o bien, remplazar subcadenas de texto con respecto el caso de las direcciones IPV6 donde se tienen bloques de ceros adyacentes que se reemplazan por dos puntos.

También se abordó el tema de la vectorización de documentos en la cual se aplicó la normalización de texto en cada documento y se generó un diccionario de términos al cual se le aplicaron técnicas de vectorización las cuales aplican formulas estadísticas a cada termino dentro de cada documento o bien a cada termino como el caso del IDF y de la probabilidad que se mide con respecto a cada termino dentro del corpus. Y sin importar la técnica de vectorización que se utilice, el tamaño del vector que representa a cada documento conserva su tamaño o en este caso, conserva su dimensión de acuerdo con el número de características representado por cada termino en el diccionario que se generó.

Ambas herramientas son de gran utilidad para darle un formato adecuado a los datos que se necesiten procesar en un modelo de procesamiento de lenguaje natural.