INSTITUTO POLITECNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO

“Practica 4- Similitud de palabras, frases y documentos”

-Tinoco Videgaray Sergio Ernesto

Grupo: 5BV1

Materia: Tecnologías del Lenguaje Natural

13/12/23

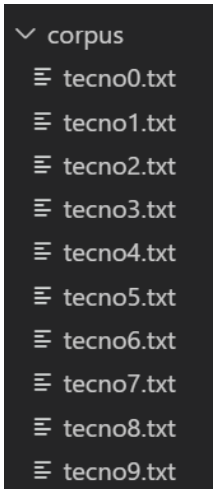
- # Introducción.

En esta práctica se van a implementar varias métricas para medir la similitud entre las palabras de 10 libros de tecnología.

- # Desarrollo

Como primer punto, se va a crear el corpus de esta práctica, el cual va a estar compuesto por 10 libros que abarcan temas de tecnología y desarrollo.

A través de la página https://www.gutenberg.org/ se van a descargar 10 que pertenezcan al tema de tecnología y se va a construir el corpus para esta práctica (Figura 1).



(Figura 1). Corpus compuesto por 10 documentos.

1. **Preparación de texto.**

Una vez creado el corpus para la práctica se va a preprocesar el texto para extraer los términos de interés. Primeramente, se va a obtener el texto correspondiente con el capítulo 1 de cada libro usando una expresión regular (Figura 2).

```
doc=documento[re.search("(Chapter|CHAPTER) (One|one|1|i|I)",documento).end()+1:
              re.search("(Chapter|CHAPTER) (Two|two|2|ii|II)",documento).start()
```

Figura 2. Expresión regular para extraer texto del capítulo 1 de cada libro.

Posteriormente se va a segmentar el texto, primero en oraciones y luego en palabras (tokenizar) para posteriormente generar el diccionario con los términos que no correspondan a stopwords o signos de puntuación. Dicho diccionario va a almacenar las frecuencias de cada palabra dentro del documento (Figura 3).

```
sent_text = nltk.sent_tokenize(doc)
word_freq={}
for sentence in sent_text:
    tokenized_text = nltk.word_tokenize(sentence)
    tagged = nltk.pos_tag(tokenized_text)
    for word,tag in tagged:
        if(tag[0]==category):
            if (word not in list(STOP_WORDS) and word not in punctuation
                if(word not in word_freq.keys()):
                    word_freq[word]=1
                else:
                    word_freq[word]+=1
```

(Figura 3). Código de la función que tokeniza y genera el diccionario con la frecuencia de los términos.

2. **Similitud de palabras con synsets.**

Para realizar la similitud de palabras se va a tomar el termino más común en cada documento que corresponde al primer capítulo de cada libro utilizando las claves del diccionario creado previamente (Figura 4).

```
common_word=max(word_freq,key=word_freq.get)
return common_word,list(word_freq.keys())
```

(Figura 4). Código de la función que devuelve el termino con mayor frecuencia de cada documento.

En este caso se va a calcular la similitud utilizando “Synsets” que como tal describen una estructura jerárquica o taxonomía que relaciona términos a manera de hiperónimo-hipónimo por medio de un árbol de términos. De tal forma que la puntuación entre términos está dada por la distancia que une a cada termino con su hiperónimo más cercano (Figura 5-14).

```
Libro 1
Verbo mas comun:shown
Verbos similares a shown dentro del documento :
Wup similarity:[('flywheel', 0.13333333333333333), ('times', 0.13333333333333333), ('makes', 0.11764705882352941), ('let', 0.11764705
882352941), ('Let', 0.11764705882352941)]
Path similarity:[('flywheel', 0.07142857142857142), ('times', 0.07142857142857142), ('makes', 0.0625), ('let', 0.0625), ('Let', 0.062
5)]

Sustantivo mas comun:pressure dentro del documento
Sustantivos similares a pressure:
Wup similarity:[('represent', 0.1111111111111111), ('times', 0.1111111111111111), ('removal', 0.1111111111111111), ('CF', 0.111111111
1111111), ('Make', 0.1)]
Path similarity:[('times', 0.058823529411764705), ('removal', 0.058823529411764705), ('CF', 0.058823529411764705), ('Make', 0.0526315
7894736842), ('OX', 0.047619047619047616)]
```

(Figura 5). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 1.

```
Libro 2
Verbo mas comun:brought
Verbos similares a brought:
Wup similarity:[('showing', 0.13333333333333333), ('lies', 0.13333333333333333), ('spin', 0.13333333333333333), ('indigo', 0.125), ('
makes', 0.1111111111111111)]
Path similarity:[('showing', 0.07142857142857142), ('lies', 0.07142857142857142), ('spin', 0.07142857142857142), ('indigo', 0.0666666
6666666667), ('makes', 0.058823529411764705)]

Sustantivo mas comun:cotton
Sustantivos similares a cotton:
Wup similarity:[('Dominate', 0.13333333333333333), ('manufacturing', 0.13333333333333333), ('spinning', 0.13333333333333333), ('grown
', 0.13333333333333333), ('tariff', 0.125)]
Path similarity:[('corn', 0.05263157894736842), ('rice', 0.05263157894736842), ('Slater', 0.05263157894736842), ('monarch', 0.05), ('
Fisher', 0.047619047619047616)]
```

(Figura 6). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 2.

```
Libro 3
Verbo mas comun:sewing
Verbos similares a sewing:
Wup similarity:[('needle', 0.1), ('saw', 0.1), ('pushed', 0.09523809523809523), ('represents', 0.09523809523809523), ('represented',
0.09523809523809523)]
Path similarity:[('needle', 0.05263157894736842), ('saw', 0.05263157894736842), ('pushed', 0.05), ('represents', 0.05), ('represented
', 0.05)]

Sustantivo mas comun:machine
Sustantivos similares a machine:
Wup similarity:[('March', 0.10526315789473684), ('Times', 0.10526315789473684), ('versions', 0.10526315789473684), ('store', 0.1), ('
couching', 0.1)]
Path similarity:[('March', 0.05555555555555555), ('Times', 0.05555555555555555), ('versions', 0.05555555555555555), ('store', 0.05263
157894736842), ('couching', 0.05263157894736842)]
```

(Figura 7). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 3.

```
Libro 4
Verbo mas comun:shown
Verbos similares a shown:
Wup similarity:[('screw', 0.125), ('bent', 0.125), ('screws', 0.125), ('let', 0.11764705882352941), ('makes', 0.11764705882352941)]
Path similarity:[('screw', 0.06666666666666667), ('bent', 0.06666666666666667), ('screws', 0.06666666666666667), ('let', 0.0625), ('m
akes', 0.0625)]

Sustantivo mas comun:water
Sustantivos similares a water:
Wup similarity:[('condenses', 0.13333333333333333), ('allows', 0.13333333333333333), ('diagram', 0.125), ('advertises', 0.125), ('sig
nifies', 0.1)]
Path similarity:[('copper', 0.0625), ('two-thirds', 0.0625), ('store', 0.058823529411764705), ('stores', 0.058823529411764705), ('sig
nifies', 0.05263157894736842)]
```

(Figura 8). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 4.

```
Libro 5
Verbo mas comun:acquire
Verbos similares a acquire:
Wup similarity:[('bottle', 0.2222222222222222), ('involved', 0.2222222222222222), ('travel', 0.2), ('bears', 0.18181818181818182), ('
advance', 0.18181818181818182)]
Path similarity:[('bottle', 0.125), ('involved', 0.125), ('travel', 0.1111111111111111), ('bears', 0.1), ('advance', 0.1)]

Sustantivo mas comun:electricity
Sustantivos similares a electricity:
Wup similarity:[('operation', 0.1111111111111111), ('Modern', 0.1111111111111111), ('electrolysis', 0.10526315789473684), ('store', 0
.10526315789473684), ('magnetism', 0.1)]
Path similarity:[('Modern', 0.058823529411764705), ('Royal', 0.05555555555555555), ('electrolysis', 0.05555555555555555), ('store', 0
.05555555555555555), ('magnetism', 0.05263157894736842)]
```

(Figura 9). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 5.

```
Libro 6
Verbo mas comun:know
Verbos similares a know dentro del documento :
Wup similarity:[('work', 0.125), ('press', 0.11764705882352941), ('prints', 0.11764705882352941), ('marked', 0.11764705882352941), ('
mean', 0.1111111111111111)]
Path similarity:[('press', 0.0625), ('prints', 0.0625), ('marked', 0.0625), ('mean', 0.058823529411764705), ('store', 0.0588235294117
64705)]

Sustantivo mas comun:brain dentro del documento
Sustantivos similares a brain:
Wup similarity:[('guesses', 0.10526315789473684), ('astronomy', 0.10526315789473684), ('store', 0.1), ('stores', 0.1), ('Make', 0.095
23809523809523)]
Path similarity:[('guesses', 0.05555555555555555), ('astronomy', 0.05555555555555555), ('store', 0.05263157894736842), ('stores', 0.0
5263157894736842), ('Make', 0.05)]
```

(Figura 10). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 6.

```
Libro 7
Verbo mas comun:known
Verbos similares a known:
Wup similarity:[('leak', 0.125), ('makes', 0.125), ('Let', 0.125), ('let', 0.125), ('fix', 0.125)]
Path similarity:[('leak', 0.06666666666666667), ('makes', 0.06666666666666667), ('Let', 0.06666666666666667), ('let', 0.0666666666666
6667), ('fix', 0.06666666666666667)]

Sustantivo mas comun:wire
Sustantivos similares a wire:
Wup similarity:[('leaks', 0.08695652173913043), ('magnetism', 0.08695652173913043), ('Let', 0.08695652173913043), ('commerce', 0.0869
5652173913043), ('consumption', 0.08333333333333333)]
Path similarity:[('leaks', 0.045454545454545456), ('magnetism', 0.045454545454545456), ('Let', 0.045454545454545456), ('commerce', 0.
045454545454545456), ('consumption', 0.043478260869565216)]
```

(Figura 11). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 7.

```
Libro 8
Verbo mas comun:referred
Verbos similares a referred:
Wup similarity:[('meaning', 0.15384615384615385), ('finding', 0.14285714285714285), ('standard', 0.14285714285714285), ('cause', 0.14
285714285714285), ('linseed', 0.1111111111111111)]
Path similarity:[('meaning', 0.08333333333333333), ('finding', 0.07692307692307693), ('standard', 0.07692307692307693), ('cause', 0.0
7692307692307693), ('linseed', 0.058823529411764705)]

Sustantivo mas comun:oil
Sustantivos similares a oil:
Wup similarity:[('lesser', 0.14285714285714285), ('working', 0.14285714285714285), ('XIV', 0.14285714285714285), ('Drying', 0.1333333
3333333333), ('drying', 0.13333333333333333)]
Path similarity:[('terms', 0.06666666666666667), ('industry', 0.0625), ('analysis', 0.0625), ('attempt', 0.058823529411764705), ('val
ues', 0.058823529411764705)]
```

(Figura 12). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 8.

```
Libro 9
Verbo mas comun:found
Verbos similares a found:
Wup similarity:[('cross', 0.1), ('lie', 0.1), ('required', 0.1), ('sweep', 0.09523809523809523), ('pushed', 0.09090909090909091)]
Path similarity:[('required', 0.05263157894736842), ('sweep', 0.05), ('steal', 0.047619047619047616), ('pushed', 0.047619047619047616
), ('let', 0.047619047619047616)]

Sustantivo mas comun:trenches
Sustantivos similares a trenches:
Wup similarity:[('quarters', 0.10526315789473684), ('operation', 0.10526315789473684), ('assault', 0.10526315789473684), ('reserve',
0.1), ('drainage', 0.1)]
Path similarity:[('operation', 0.05555555555555555), ('cattle', 0.05555555555555555), ('assault', 0.05555555555555555), ('reserve', 0
.05263157894736842), ('drainage', 0.05263157894736842)]
```

(Figura 13). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 9.

```
Libro 10
Verbo mas comun:asked
Verbos similares a asked:
Wup similarity:[('spoke', 0.11764705882352941), ('shot', 0.11764705882352941), ('shooting', 0.1111111111111111), ('Let', 0.1052631578
9473684), ('let', 0.10526315789473684)]
Path similarity:[('spoke', 0.0625), ('shot', 0.0625), ('shooting', 0.058823529411764705), ('Let', 0.05555555555555555), ('let', 0.055
55555555555555)]

Sustantivo mas comun:Dick
Sustantivos similares a Dick:
Wup similarity:[('turn', 0.1), ('None', 0.1), ('cent', 0.1), ('trial', 0.09523809523809523), ('performances', 0.09523809523809523)]
Path similarity:[('turn', 0.05263157894736842), ('None', 0.05263157894736842), ('cent', 0.05263157894736842), ('trial', 0.05), ('performances', 0.05)]
```

(Figura 14). Impresión en consola de los términos más comunes junto con los términos más semejantes dentro del documento 10.

3. **Similitud de documentos con synsets.**

En esta sección se va a obtener la similitud entre los documentos, más específicamente entre el libro 1 y los demás, para esto se va a obtener la frase clave por medio del modelo BERT y posteriormente se va a calcular la similitud utilizando nuevamente los Synsets (Figura 15-25).

```
Libro 1
Frase clave del libro 1:heat steam
```

(Figura 15). Impresión en consola de la frase más representativa del libro 1.

```
Libro 2
Frase clave del libro 2:inventor cotton
Similitud con heat steam
Synsets:0.09545454545454546
```

(Figura 16). Impresión en consola de la frase clave del documento 2 y su similitud con la frase del libro 1.

```
Libro 3
Frase clave del libro 3:machine sewing
Similitud con heat steam
Synsets:0.07692307692307693
```

(Figura 17). Impresión en consola de la frase clave del documento 3 y su similitud con la frase del libro 1.

```
Libro 4
Frase clave del libro 4:boiler steam
Similitud con heat steam
Synsets:0.5370879120879121
```

(Figura 18). Impresión en consola de la frase clave del documento 4 y su similitud con la frase del libro 1.

```
Libro 5
Frase clave del libro 5:study electricity
Similitud con heat steam
Synsets:0.16145833333333331
```

(Figura 19). Impresión en consola de la frase clave del documento 5 y su similitud con la frase del libro 1.

```
Libro 6
Frase clave del libro 6:brains mechanical
Similitud con heat steam
Synsets:0.09423076923076923
```

(Figura 20). Impresión en consola de la frase clave del documento 6 y su similitud con la frase del libro 1.

```
Libro 7
Frase clave del libro 7:electricity important
Similitud con heat steam
Synsets:0.175
```

(Figura 21). Impresión en consola de la frase clave del documento 7 y su similitud con la frase del libro 1.

```
Libro 8
Frase clave del libro 8:paint oil
Similitud con heat steam
Synsets:0.10416666666666666
```

(Figura 22). Impresión en consola de la frase clave del documento 8 y su similitud con la frase del libro 1.

```
Libro 9
Frase clave del libro 9:germans trenches
Similitud con heat steam
Synsets:0.08173076923076922
```

(Figura 23). Impresión en consola de la frase clave del documento 9 y su similitud con la frase del libro 1.

```
Libro 10
Frase clave del libro 10:gasoline driven
Similitud con heat steam
Synsets:0.10606060606060605
```

(Figura 24). Impresión en consola de la frase clave del documento 10 y su similitud con la frase del libro 1.

4. **Similitud de palabras con embedding.**

En esta sección se van a obtener los términos similares a la palabra más común de cada documento utilizando “embeddings”, que corresponden a los vectores n-dimensionales que representan cada palabra en un vocabulario global proporcionado por el paquete de GloVe (Figura 25-35).

Analizando los archivos de GloVe se puede observar la estructura de cada archivo, la cual consiste en un archivo de texto que se divide en varias columnas separadas por espacios, en la primera columna se encuentra cada palabra del diccionario inglés, algunas de ellas son incluso son stopwords, y en las columnas de la derecha se encuentran los valores de cada una de sus dimensiones como se observa en la figura 25.

```
the -0.038194 -0.24487 0.72812 -0.39961 0.083172 0.043953 -0.39141 0.3344 -
, -0.10767 0.11053 0.59812 -0.54361 0.67396 0.10663 0.038867 0.35481 0.0635
. -0.33979 0.20941 0.46348 -0.64792 -0.38377 0.038034 0.17127 0.15978 0.466
of -0.1529 -0.24279 0.89837 0.16996 0.53516 0.48784 -0.58826 -0.17982 -1.35
to -0.1897 0.050024 0.19084 -0.049184 -0.089737 0.21006 -0.54952 0.098377 -
and -0.071953 0.23127 0.023731 -0.50638 0.33923 0.1959 -0.32943 0.18364 -0.
in 0.085703 -0.22201 0.16569 0.13373 0.38239 0.35401 0.01287 0.22461 -0.438
a -0.27086 0.044006 -0.02026 -0.17395 0.6444 0.71213 0.3551 0.47138 -0.2963
" -0.30457 -0.23645 0.17576 -0.72854 -0.28343 -0.2564 0.26587 0.025309 -0.0
's 0.58854 -0.2025 0.73479 -0.68338 -0.19675 -0.1802 -0.39177 0.34172 -0.60
for -0.14401 0.32554 0.14257 -0.099227 0.72536 0.19321 -0.24188 0.20223 -0.
- -1.2557 0.61036 0.56793 -0.96596 -0.45249 -0.071696 0.57122 -0.31292 -0.4
that -0.093337 0.19043 0.68457 -0.41548 -0.22777 -0.11803 -0.095434 0.19613
```

(Figura 25). Archivo de GloVe con 50 dimensiones.

Para fines prácticos se va a utilizar el archivo de 200 dimensiones ya que, aunque pueda requerir más costo computacional para realizar las operaciones, los resultados tienden a ser mas precisos.
Por lo que a partir de estos valores se van a generar los vectores o “embeddings” de cada palabra dentro de cada documento para calcular su similitud utilizando la similitud de coseno (figura 26-36).

```
Verbos similares a shown dentro del documento (embeddings):[('showing', 3.2296640872955322), ('shows', 3.4920928478240967), ('taken',
 4.75035285949707), ('having', 4.833104610443115), ('given', 4.899876594543457)]
```

(Figura 26). Palabras similares al termino común del documento 1 utilizando embeddings.

```
Verbos similares a brought dentro del documento (embeddings):[('turned', 3.617495059967041), ('come', 3.860267162322998), ('taken', 3
.973478317260742), ('followed', 4.019801616668701), ('having', 4.078278064727783)]
```

(Figura 27). Palabras similares al termino común del documento 2 utilizando embeddings.

```
Verbos similares a sewing dentro del documento (embeddings):[('stitching', 5.447918891906738), ('sew', 5.489567756652832), ('weaving'
, 5.861759662628174), ('stitched', 6.000957012176514), ('quilting', 6.186944961547852)]
```

(Figura 28). Palabras similares al termino común del documento 3 utilizando embeddings.

```
Verbos similares a shown dentro del documento (embeddings):[('showing', 3.2296640872955322), ('shows', 3.4920928478240967), ('seen', 4.192713737487793), ('taken', 4.75035285949707), ('indicate', 4.781532287597656)]
```

(Figura 29). Palabras similares al termino común del documento 4 utilizando embeddings.

```
Verbos similares a acquire dentro del documento (embeddings):[('develop', 5.543446063995361), ('offered', 5.816229820251465), ('enter', 5.825678825378418), ('employ', 5.921205520629883), ('trying', 6.098968029022217)]
```

(Figura 30). Palabras similares al termino común del documento 5 utilizando embeddings.

```
Verbos similares a know dentro del documento (embeddings):[('sure', 2.566188097000122), ('think', 2.705873727798462), ('tell', 2.706615686416626), ('knows', 3.014061450958252), ('want', 3.2368240356445312)]
```

(Figura 31). Palabras similares al termino común del documento 6 utilizando embeddings.

```
Verbos similares a known dentro del documento (embeddings):[('called', 3.75176739692688), ('referred', 3.7618305683135986), ('named', 4.326873779296875), ('regarded', 4.507499694824219), ('described', 4.523844242095947)]
```

(Figura 32). Palabras similares al termino común del documento 7 utilizando embeddings.

```
Verbos similares a referred dentro del documento (embeddings):[('known', 3.7618305683135986), ('given', 4.655069828033447), ('meaning', 4.655766010284424), ('taken', 4.969942092895508), ('differ', 5.457228660583496)]
```

(Figura 33). Palabras similares al termino común del documento 8 utilizando embeddings.

```
Verbos similares a found dentro del documento (embeddings):[('taken', 4.514556884765625), ('find', 4.55593729019165), ('having', 4.845678329467773), ('seen', 4.913298606872559), ('looking', 5.140047550201416)]
```

(Figura 34). Palabras similares al termino común del documento 9 utilizando embeddings.

```
Verbos similares a asked dentro del documento (embeddings):[('wanted', 3.540018081665039), ('decided', 3.8039300441741943), ('explained', 3.9224679470062256), ('tell', 4.014706134796143), ('invited', 4.3715715408325195)]
```

(Figura 35). Palabras similares al termino común del documento 10 utilizando embeddings.

Similitud de documentos con embeddings. Análogamente a lo que se hizo en el punto 3, se va a calcular la similitud entre documentos, pero ahora utilizando embeddings y el modelo “largo” pre-entrenado BERT para obtener los embeddings.

```
Libro 1
Frase clave del libro 1:steam mechanical
```

(Figura 36). Impresión en consola de la frase clave del libro 1.

```
Libro 2
Frase clave del libro 2:invention cotton
Similitud con steam mechanical
Embeddings:0.5475450158119202
```

(Figura 37). Frase clave del documento 2 y su similitud con la frase clave del libro 1.

```
Libro 3
Frase clave del libro 3:sewing machines
Similitud con steam mechanical
Embeddings:0.5155001878738403
```

(Figura 38). Frase clave del documento 3 y su similitud con la frase clave del libro 1.

```
Libro 4
Frase clave del libro 4:steam molecules
Similitud con steam mechanical
Embeddings:0.6671524047851562
```

(Figura 39). Frase clave del documento 4 y su similitud con la frase clave del libro 1.

```
Libro 5
Frase clave del libro 5:electricity historical
Similitud con steam mechanical
Embeddings:0.6146590113639832
```

(Figura 40). Frase clave del documento 5 y su similitud con la frase clave del libro 1.

```
Libro 6
Frase clave del libro 6:mechanical brains
Similitud con steam mechanical
Embeddings:0.5895987153053284
```

(Figura 41). Frase clave del documento 6 y su similitud con la frase clave del libro 1.

```
Libro 7
Frase clave del libro 7:practical electricity
Similitud con steam mechanical
Embeddings:0.5841083526611328
```

(Figura 42). Frase clave del documento 7 y su similitud con la frase clave del libro 1.

```
Libro 8
Frase clave del libro 8:paint oils
Similitud con steam mechanical
Embeddings:0.5514983534812927
```

(Figura 43). Frase clave del documento 8 y su similitud con la frase clave del libro 1.

```
Libro 9
Frase clave del libro 9:germans preparing
Similitud con steam mechanical
Embeddings:0.5186160206794739
```

(Figura 44). Frase clave del documento 9 y su similitud con la frase clave del libro 1.

```
Libro 10
Frase clave del libro 10:car glided
Similitud con steam mechanical
Embeddings:0.47184228897094727
```

(Figura 45). Frase clave del documento 10 y su similitud con la frase clave del libro 1.

Análisis de resultados.

Haciendo un análisis se pudo observar que la similitud entre palabras dentro de cada documento fue más precisa utilizando embeddings que utilizando synsets (Figura 46 y 47).

```
Verbos similares a shown dentro del documento :
Wup similarity:[('flywheel', 0.13333333333333333), ('times', 0.13333333333333333), ('makes', 0.11764705882352941), ('let', 0.1176470588235
2941), ('Let', 0.11764705882352941)]
Path similarity:[('flywheel', 0.07142857142857142), ('times', 0.07142857142857142), ('makes', 0.0625), ('let', 0.0625), ('Let', 0.0625)]

Embeddings:[('showing', 3.2296640872955322), ('shows', 3.4920928478240967), ('taken', 4.75035285949707), ('having', 4.833104610443115), ('
given', 4.899876594543457)]
```

(Figura 46). Comparación entre la similitud de términos con diferentes métricas para el documento 1.

```
Libro 2
Verbo mas comun:brought
Verbos similares a brought dentro del documento :
Wup similarity:[('showing', 0.13333333333333333), ('lies', 0.13333333333333333), ('spin', 0.13333333333333333), ('indigo', 0.125), ('makes
', 0.1111111111111111)]
Path similarity:[('showing', 0.07142857142857142), ('lies', 0.07142857142857142), ('spin', 0.07142857142857142), ('indigo', 0.066666666666
66667), ('makes', 0.05882352941176705)]
Embeddings:[('turned', 3.617495059967041), ('come', 3.860267162322998), ('taken', 3.973478317260742), ('followed', 4.019801616668701), ('h
aving', 4.078278064727783)]
```

(Figura 47). Comparación entre la similitud de términos con diferentes métricas para el documento 2.

Donde se puede observar que en la figura 46, se muestran los terminos “showing” y “shows”, los cuales no aparecen en los Synsets, lo que implica que el archivo de GloVe esta “más completo” en comparación con los Synsets.

Si bien, esto puede variar dependiendo los términos y el número de dimensiones de los embeddings.

De igual forma la similitud entre documentos fue bastante parecida utilizando ambas métricas como se observa en la figura 48 y 49:

```
Frase clave del libro 4:steam molecules
Similitud con steam mechanical
Synsets:0.775
Embeddings:0.6671524047851562
```

(Figura 48). Similitudes entre el documento 4 y el documento 1 utilizando synsets y embeddings.

```
Frase clave del libro 10:car glided
Similitud con steam mechanical
Synsets:0.08380681818181818
Embeddings:0.47184228897094727
```

(Figura 49). Similitudes entre el documento 10 y el documento 1 utilizando synsets y embeddings.

Aquí se observa como los valores de similitud del libro 4 presentan una frase clave bastante similar a la del libro 1, en comparación con la similitud del libro 10, que contiene una frase diferente de acuerdo con el contexto del libro.

Conclusiones

En esta práctica se realizaron varias técnicas de similitud entre palabras y documentos que permiten comparar textos y documentos basado en términos o frases representativas, de tal forma que se compara el texto desde un punto de vista semántico ya que se tiene conocimiento acerca de los términos y su significado y esto nos permite contextualizar más el procesamiento del lenguaje natural.

Por un lado, los synsets permiten medir la similitud entre términos basándose en la distancia entre su término hiperónimo, es decir el termino común que los clasifica dentro de la taxonomía proporcionada por WordNet.

Por otro lado, los embeddings sirven como una representación numérica, más específicamente una representación vectorial de los términos dentro de un espacio n-dimensional que engloba todos, o al menos, la mayoría de los términos de un idioma, y de esta forma se pueden realizar operaciones como la distancia entre términos para conocer su similitud dentro de este espacio vectorial o incluso la suma o resta aritmética para obtener otros términos.

Por lo anterior puedo concluir que ambas métricas son de utilidad y seria adecuado aplicar ambas en la mayoría de los modelos de procesamiento de lenguaje natural dándole prioridad a la que se adapte mejor al contexto del modelo.