U2c_A1_GSV

February 23, 2023

Gustavo Santana Velázquez

1 Introducción

En lingüistica, la entropía es una medida que calcula la complejidad y la impredicibilidad de un texto. Se basa en la probabilidad de ocurrencia de cada carácter o palabra en un texto y se utiliza para evaluar cuánta información se puede obtener del texto.

Cuanto mayor sea la entropía, mayor será la complejidad e impredicibilidad del texto, es decir, hay una mayor variedad de palabras o caracteres utilizados y su distribución es más uniforme. Por el otro lado, una entroía baja indica que el texto es más predecible y hay una menor variedad de palabras o caracteres utilizadas con una distribución más concentrada.

En resumen, la entropía se puede interpretar como una medida de la riqueza léxica de un texto, y puede utilizarse para comparar la complejidad y el estilo de diferentes textos.

En este trabajo se evaluará la entropía de 5 textos cortos: Cantata a Satanás, el hereje rebelde, mimí sin bikini, los locos somos otro cosmos y un gurú vudú; cada uno tiene la particularidad de utilizar únicamente una vocal, por lo que se espera observar una entropía baja. Posteriormente se calculará la entropía de 2 libros: Niebla, de Miguel Unamono; y Marianela, por Benito Pérez Galdós. Para estos últimos se calculará la entropía con y sin *stopwords* para determinar qué tanto impacto tienen estas en el resultado.

2 Desarrollo

La entropía $H(X)$ de un texto se puede obtener a través de la siguiente fórmula:

$$H(X) = -\sum p(x) \log_2 p(x)$$

donde $p(x)$ la función de probabilidad de una variable aleatoria X sobre un conjunto discreto de símbolos (o alfabeto).

Los textos y libros fueron procesados antes de calcular la entropía, removiendo acentos, símbolos y números (Apéndice 5.2). Se definió una función en *Python* para calcular la entropía a nivel carácter o palabra para se utilziada en este ejercicio (Apéndice 4.3).

Para los textos 1-5 se obtuvieron los siguientes resultados: (Apéndice 5.4)

Texto	Entropía	Número de caracteres	Repetición de vocal	% de repetición
text_1	3.080115476	4,267	1,863	43.66%
text_2	3.087512876	3,227	1,360	42.14%
text_3	3.265331452	1,326	485	36.58%
text_4	3.116194809	2,946	1,179	40.02%
text_5	3.171419752	1,340	557	41.57%

Se puede observar que la entropía es inversamente proporcional al % de repetición de la vocal en el texto. Esto es debido a que es la misma vocal la que se repite, la variedad de carácteres es menor y el texto es predecible, es por esto que en los textos en los que se repite la vocal más veces, la entropía disminuye.

En los libros 1 y 2 con stopwords (Apéndice 5.5.1)

Libro	Entropía	Número de palabras
libro_1	9.228294320935682	56,810
libro_2	9.611731124522716	50,536

Entropía de los libros 1 y 2 sin stopwords (Apéndice 5.5.2)

Libro	Entropía	Número de palabras
libro_1	11.353860554400953	26,307
libro_2	11.723176507369077	25,170

La entropía en los libros resulta mucho mayor que en los 5 textos pasados, esto se debe a la incorporación de más variaciones (palabras vs carácteres del alfabeto) además de menos repeticiones.

La entropía de los libros es muy similar, aunque la del libro 2 es un poco más grande, los que indica que hay una probabilidad de que sea más complejo.

Tras remover las stopwords (que cabe mencionar que usualmente son las palabras más repetidas en un texto), la entropía aumenta en apróximadamente 2.1 para ambos libros. Por lo que se puede apreciar la influencia que estas tienen en el cálculo de esta medida. También se observa a partir de las tablas que las stopwords componen aproximadamente la mitad del total de palabras para cada uno de los libros.

3 Conclusiones

En conclusión, se realizó un análisis de entropía para un conjunto de textos cortos y dos libros. Se encontró que los libros tenían una entropía más alta que los textos cortos, lo que sugiere una mayor variedad léxica en el lenguaje utilizado en los libros. Además, se descubrió que al eliminar las stopwords, la entropía aumentó en ambos libros, lo que indica que las stopwords contribuyen significativamente a la repetición de palabras en los textos. En general, la entropía es una medida útil para cuantificar la diversidad léxica de un texto y puede ser útil para comprender las características del lenguaje utilizado en diferentes contextos.

4 Bibliografía

Jurafsky, D. & Martin, J.H. (2008). Speech and Language Processing. Prentice Hall, Segunda Edición.

Manning, C. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA.

Chen, R., Haitao L. & Altmann, G. (2016). Entropy in different text types. Digital Scholarship in the Humanities, vol. 32, Issue 3, pp. 528–542.

5 Apéndice

5.1 Librerías

[ ]:
```
# Librerías
import io
import unicodedata
import pandas as pd
from nltk.corpus import stopwords
import math
from collections import Counter
```

5.2 Preprocesamiento de datos

[ ]:
```
# Stopwords y símbolos
spanish_stopwords = stopwords.words('spanish')
spanish_stopwords.extend(['si', 'mas'])
symbols = list(set("«-;:,.\\-\"'/()[]¿?¡!{}~<>|\r_'\ufeff'"))
```

[ ]:
```
def preprocess(input_str):
    '''
    Converts input string to lower case and removes accents and symbols.
    '''
    input_str = input_str.lower()
    # input_str = input_str.replace('\n', '')
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return u''.join([c for c in nfkd_form if not unicodedata.combining(c) and\
        c not in symbols and not c.isnumeric()])
```

5.3 Función para determinar entropía

[ ]:
```
def entropia(input_str):
    '''
    Calcula la entropía de los caracteres o palabras en un texto
    '''
    # Contar la frecuencia de cada carácter en el texto
    freqs = Counter(input_str)
```

```
    # Calcular la probabilidad de cada carácter
    probs = [float(freqs[c]) / len(input_str) for c in freqs]
    # Calcular la entropía
    entropy = - sum(p * math.log2(p) for p in probs)
    return entropy
```

5.4 Determinar la entropía global de los textos 1-5

Se leen los 5 textos y se guardan en un diccionario (para poder iterarlos), eliminando símbolos, acentos y saltos de línea en el proceso.

Posteriormente se calcula la entropía para cada uno de los textos.

[ ]:
```
texts = dict(text_1 = [], text_2 = [], text_3 = [], text_4 = [], text_5 = [])

for text in texts:
    # Read lines from text files changing to lower and removing accents
    raw = ''.join([preprocess(line) for line in io.open\
        (f'./text_files/{text}.txt', 'r', encoding = 'UTF-8').readlines()])

    # Remove whitespaces and line breaks
    texts[text] = raw.replace(' ', '').replace('\n', '')
```

[ ]:
```
for key, value in texts.items():
    print(f'Entropía de {key}: {entropia(value):.10}' +
        f' \tNúmero de caracteres: {len(value):,}' +
        f' \tVocales: {sum(1 for c in value if c in "aeiou"):,}')
    #print(f'{100*sum(1 for c in value if c in "aeiou")/len(value):.4}%')
```

```
Entropía de text_1: 3.080115476         Número de caracteres: 4,267     Vocales:
1,863
Entropía de text_2: 3.087512876         Número de caracteres: 3,227     Vocales:
1,360
Entropía de text_3: 3.265331452         Número de caracteres: 1,326     Vocales:
485
Entropía de text_4: 3.116194809         Número de caracteres: 2,946     Vocales:
1,179
Entropía de text_5: 3.171419752         Número de caracteres: 1,340     Vocales:
557
```

5.5 Entropía global de los libros 1 y 2

Se comienza por importar los textos, removiendo acentos, símbolos y saltos de línea en el proceso.

Posteriormente calcula la entropía para ambos textos (incluyendo stopwords).

Finalmente, se remueven las stopwords en español y se repite el ejercicio.

[ ]:
```
# Import books
books = dict(libro_1 = [], libro_2 = [])

for book in books:
    books[book] = ''.join([preprocess(line.replace('\n', ' ')) for line in io.
 ↪open\
        (f'./text_files/{book}.txt', 'r', encoding = 'UTF-8').readlines()])
```

5.5.1 Entropía global a nivel palabra considerando stopwords

[ ]:
```
for key, value in books.items():
    print(f'Entropía del {key}: {entropia(value.split())}'+
        f' \tNúmero de palabras: {len(value.split()):,}')
```

```
Entropía del libro_1: 9.228294320935682         Número de palabras: 56,810
Entropía del libro_2: 9.611731124522716         Número de palabras: 50,536
```

5.5.2 Entropía global a nivel palabra sin considerar stopwords

[ ]:
```
# Remover stopwords
books_no_sw = dict()
for key, value in books.items():
    books_no_sw[key] = ' '.join([w for w in value.split() if w not in␣
 ↪spanish_stopwords])
```

[ ]:
```
for key, value in books_no_sw.items():
    print(f'Entropía del {key} sin stopwords: {entropia(value.split())}' +
        f' \tNúmero de palabras: {len(value.split()):,}')
```

```
Entropía del libro_1 sin stopwords: 11.353860554400953  Número de palabras:
26,307
Entropía del libro_2 sin stopwords: 11.723176507369077  Número de palabras:
25,170
```