UNIVERSIDAD NACIONAL DE COLOMBIA

Introducción a la Investigación Teórica

Análisis entrópico para procesamiento de lenguajes naturales

Santiago Ramirez Vallejo [1], Paula Andrea Plazas Isanoa [2]
Luis Fernando Horta Camacho [3], Juan Sebastian Cachaya Munar [4]

8 de febrero de 2022

1. Introducción

Las palabras en español pueden ser consideradas como parte de una red compleja construida a partir de correlaciones sintácticas entre elementos de diferentes niveles (letras, palabras, oraciones etc.). Siendo así, se pueden generar modelos de probabilidad diferentes con enfoques diversos para el análisis de estas estructuras en el idioma español [19]. Las bases de los modelos de máxima entropía provienen de la entropía de información de Shannon [12], y fueron aplicados al procesamiento del lenguaje natural por Berger [1], así logrando obtener la mayor uniformidad posible anteponiendo aquellas distribuciones que maximicen la entropía del sistema.

El presente trabajo se centra, específicamente, en plantear modelos de densidad de probabilidad de las palabras de cuatro letras que utilizan las 27 letras del alfabeto español (omitiendo dígrafos, guiones o números y sin tener en cuenta diferenciación por mayúsculas), además de las cinco vocales con acento adicionales, dando en total un conjunto de 32 caracteres posibles.

Para ello, es necesario primero seleccionar adecuadamente una base de datos de palabras (al que llamaremos corpus). Éste será seleccionado teniendo en cuenta la naturaleza del texto (principalmente narrativa), y la cantidad de palabras y palabras únicas con las que cuenta (al menos 680 mil palabras).

Posteriormente se debe crear un programa que permita analizar el texto, separando las palabras de cuatro letras, determinando la frecuencia de cada una, y contando los caracteres del alfabeto presentes en cada posición.

Además, se explican y desarrollan cuatro modelos de densidad de probabilidad diferentes: El modelo aleatorio, que no toma ninguna consideración de restricciones en cuenta. El modelo independiente, que evalúa la probabilidad individual de las letras en cada posición en base al corpus. El modelo de correlación por pares, que tiene en cuenta además la probabilidad de combinación de dos letras en diferentes posiciones de la palabra. Y finalmente el modelo completo, que se apega exclusivamente a los patrones y palabras presentes en el corpus, siendo un modelo más empírico, y con el cuál serán comparados los otros tres.

El tercer modelo se construirá maximizando la entropía en las distribuciones de correlación por pares, equivalentes a las distribuciones de Boltzmann con interacciones entre los elementos del sistema [2].

2. Objetivos

2.1. Objetivo General

Construir un modelo de máxima entropía consistentes con la correlación por pares entre letras que aproxime en al menos un 70 % la estadística asociada con las palabras de cuatro letras de un *corpus* en español. Así mismo, se contempla construir otros tres modelos de máxima entropía con el fin de compararlos con el primero: el modelo aleatorio, el modelo independiente, y el modelo completo, tal como se trabajo en [19]. (Encontrando, por ejemplo, palabras no contenidas en el texto base y predichas o creadas por el modelo, diferencia de entropías, probabilidades por palabra, etc.). También se discutirá el aporte de cada modelo a la reducción del vocabulario eficaz hasta llegar al modelo completo. Se discutirá cómo varía el comportamiento de la densidad de probabilidad según el tamaño de la palabra (número de letras que la componen).

Por otro lado, esperamos adquirir las competencias necesarias para desarrollar un programa en Spider-Python el cual sea capaz de analizar los datos que dispondremos para el estudio. Asimismo , esperamos aprender a utilizar R Studio con el fin de realizar las gráficas necesarias para el análisis.

2.2. Objetivos Específicos

- Encontrar un *corpus* en español adecuado para los fines del proyecto, teniendo en cuenta la extensión de los textos utilizados en los trabajos [14] y [19] (680000 a 2160000 palabras).

[1] saramirezva@unal.edu.co
[2] pplazasi@unal.edu.co
[3] lfhortac@unal.edu.co
[4] jscachayam@unal.edu.co

- Construir un programa para la cuantización y análisis de las palabras en el *corpus*.
- Cálculo del modelo aleatorio para la construcción de palabras en español de cuatro letras. Tomando el procedimiento teórico que se realiza en la sección B del marco teórico se evalúa la entropía de este modelo.
- Cálculo del modelo independiente para la construcción de palabras en español de cuatro letras. Se evalúa la entropía la cual considera la probabilidad de aparición de cada letra en cada posición.
- Cálculo del modelo completo derivado del análisis del *corpus* por medio del programa.
- Construcción del modelo de máxima entropía para palabras de cuatro letras en español, teniendo en cuenta distribuciones de probabilidad conjunta de cada letra en cada posición, frecuencia de las palabras en el texto y las seis correlaciones por pares asociadas con las cuatro posiciones. Realizar una visión energética.
- Comparar todos los modelos con la distribución completa y analizar el aporte de cada modelo a la reducción del vocabulario eficaz.
- Analizar la densidad de probabilidad según el tamaño de la palabra (número de letras que la componen).
- Comparación de resultados con trabajos similares basados en textos naturales en inglés.

3. Modelos de densidad de probabilidad

3.1. Modelos de máxima Entropía

A mediados del siglo XX, se empezaron a combinar teorías de información y principios físicos para formar un enfoque de máxima entropía como modelo de análisis de información y método de prueba para distribuciones de probabilidad, según la denominada entropía de información de Shannon [12].

$$S_i(p_i) = -K \sum_i p_i \ln p_i. \tag{1}$$

La cuál nos da la entropía en *unidades naturales de información*. Sin embargo, para el presente caso, será preferible en *bits* [6].

$$S_I(p_I) = -K \sum_i p_i \log_2 p_i. \tag{2}$$

Donde S_I corresponde a la entropía de la distribución de probabilidad p_i. K es una constante positiva.

El modelo de máxima entropía, en la búsqueda de medir y encontrar el modelo de mayor *uniformidad*, prioriza, en base al algoritmo de Gibbs, aquellas distribuciones que maximicen dicho valor de entropía, teniendo en cuenta las restricciones que se pueden obtener o conocer del sistema [1]. Así, se llegó a la conclusión de que se podía trabajar a la entropía como principal medida de la cantidad de incertidumbre que presentaba una distribución de probabilidad, incluso siendo más fundamental que la visión puramente energética [12].

Sin embargo, también se puede calcular la denominada entropía conjunta (*joint entropy* $S_I(p_1...p_n)$) de una distribución de probabilidad conjunta (*joint distribution* $P(p_1...p_n)$), cuando se tiene un conjunto de variables [6].

$$S_I(p_1...p_n) = -\sum_{x_1 \in p_1} ... \sum_{x_n \in p_n} P(x_1...x_n) \log_2 P(x_1...x_n). \tag{3}$$

Éste método puede ser ampliamente utilizado en sistemas y redes complejas. En este caso, será necesario un enfoque en aplicaciones para el estudio del lenguaje natural como una gran y compleja red, que cumple las siguientes características [14]:

- Sucesión de unidades simbólicas (algunos modelos, establece la comparación temporal con la mecánica estadística)
- Elementos que siguen relaciones y reglas específicas (gramaticales o sintácticas. Añaden la naturaleza probabilística y aleatoriedad en la construcción de textos).
- Textos completos pueden verse como un sistema jerárquico, en donde los primeros niveles se componen de letras y sílabas, que a su vez construyen palabras, oraciones y párrafos. Por lo tanto, también existen interacciones entre niveles.

3.2. Modelo aleatorio

Como ejemplo, se utilizará la distribución de probabilidad *aleatoria* para la conformación de palabras en inglés con cuatro letras (26 letras posibles). El modelo aleatorio considera absolutamente todas las combinaciones posibles y establece que cada letra tiene una posibilidad de $1/26$ de aparecer en una posición específica. En total, cada combinación (palabra) tendrá la misma probabilidad: $(1/26)^4 = 1/456976$. La distribución de probabilidad en este caso, $P_r(l_1, l_2, l_3, l_4)$ es uniforme, donde cada l_i es una letra, y puede ser representada por una arreglo multidimensional $26 \cdot 26 \cdot 26 \cdot 26$, con probabilidades idénticas para cada configuración. La entropía en este caso del modelo aleatorio S_r será entonces, utilizando la ecuación (3).

$$\begin{gathered}
S_r(l_1, l_2, l_3, l_4) = S_r(l), \\
P_r(l_1, l_2, l_3, l_4) = P_r(l), \\
P_r(x_1, x_2, x_3, x_4) = P_r(x), \\
S_r = -\sum_{x_1 \in l_1} \sum_{x_2 \in l_2} \sum_{x_3 \in l_3} \sum_{x_4 \in l_4} P_r(x) \log_2 P_r(x) \\
S_r = -\sum_{x \in l} (1/26)^4 \log_2 (1/26)^4 \\
S_r = -26^4 (1/26)^4 4 \log_2 (1/26)
\end{gathered}$$

$$S_r = -4 \log_2 (1/26) = -4(\log_2(1) - \log_2(26))$$

$$S_r = -4(-\log_2(26)) = 4 \log_2(26)$$

$$S_r \approx 18.802.$$

Lo cual es consistente con lo enunciado en [19], el paso siguiente será entonces, hallar la distribución de probabilidad cuya entropía asociada sea la adecuada con respecto a los valores hallados en las distribuciones de los corpus, según cada caso.

3.3. Modelo Independiente

En el modelo independiente, la distribución de probabilidad combinada deja de ser uniforme, pues considera la probabilidad de aparición de cada letra en cada posición, y cada palabra (o combinación) tendrá la probabilidad de aparecer según el producto de cada una de estas probabilidades. La distribución de probabilidad se aproxima, entonces, como la productoria de las posibilidades de cada letra en cada posición. Es de resaltar que para este caso necesitaremos $32 \cdot 4 = 128$ datos de entrada (contando los 32 caracteres de estudio en el idioma español), que representarán la probabilidad de cada letra en cada una de las cuatro posiciones posibles.

$$P(l_1, l_2, l_3, l_4) \approx P_i(l_1, l_2, l_3, l_4) = P_i(l) = \prod_{j=1}^{4} P_j(l). \quad (4)$$

Donde $P_j(l)$ representa la probabilidad de cada caracter en la posición j:

$$P_j(l) = \frac{F_{lj}}{N_{u4}}. \quad (5)$$

Siendo $F_l j$ la frecuencia de aparición de cierta letra en la posición j en las palabras únicas de cuatro letras, y N_{u4} la cantidad de palabras únicas de cuatro letras encontradas en el texto.

Notamos cómo, aunque esta distribución se basa en el análisis del *corpus*, aún no establece correlaciones ni restricciones fonéticas entre las cuatro posiciones posibles.

La entropía de este modelo se notará cómo S_i.

3.4. Modelo Completo

El modelo completo es aquel basado puramente en el análisis computacional del *corpus*. En principio, utiliza las palabras únicas de cuatro letras encontradas en los textos para formar una lista de palabras posibles (las otras combinaciones tendrán, por ende, una probabilidad de 0). Después, a partir de la frecuencia de aparición de cada una de esas palabras, calcula su probabilidad y ese es su valor correspondiente en la matriz de la densidad de probabilidad combinada:

$$P(l_1, l_2, l_3, l_4) \approx P_f(l_1, l_2, l_3, l_4) = \frac{F_l}{N_4}. \quad (6)$$

Siendo F_l la frecuencia de aparición de la combinación l_1, l_2, l_3, l_4 en el corpus y N_4 la cantidad de palabras (totales, no únicas) de cuatro letras encontradas en el texto.

Aquí, al ser un análisis totalmente empírico, no se toman en cuenta tampoco correlaciones entre posiciones ni probabilidades independientes por caracter.

La entropía de este modelo se notará cómo S_f.

3.5. Comparación entre modelos

El modelo completo presentará la menor entropía de todas, y frente a él se compararán los demás modelos. Las diferencias entre las distribuciones de cada uno respecto al modelo completo, son un reflejo de las palabras *predichas* o *inventadas* por cada modelo.

Por otro lado, la diferencia entre las entropías de los modelos completo, independiente y de correlación, frente a la entropía del modelo aleatorio, muestran las restricciones añadidas por cada modelo:

$$S_{r-i} = S_r - S_i, \quad (7)$$

$$S_{r-i} = S_r - S_p, \quad (8)$$

$$S_{r-i} = S_r - S_f. \quad (9)$$

También se puede utilizar una variable llamada la variable de multi-información o la divergencia de Kullback-Leibler, que representa la diferencia entre las entropías de variables interactuantes (modelo correlacionado) y variables independientes (modelo independiente) [17]:

$$I(l) \equiv \sum_j S(l_j) - S_p = S_i - S_p = \sum_l P_p(l) \log \frac{P_p(l)}{\prod_{j=1}^{4} P_j(l)}. \quad (10)$$

A partir de cada elemento de entropía para cada modelo, podemos además calcular otras cantidades que representarán la cantidad de palabras accesibles a través de ese modelo, y al cual llamaremos el vocabulario del modelo. El cálculo general es:

$$N_x = 2^{S_x}. \quad (11)$$

3.6. Distribuciones de Boltzmann para modelo de correlación por pares

Supongamos que existe una densidad de probabilidad, y maximizamos su entropía.Para el caso de relaciones por pares entre los elementos del sistema, la entropía está dada según [2].

$$S_P(p) = -\sum_{x \in p} P(x) \log_2 P(x) - \sum_i \lambda (\sum_{x \in p} P(x) x_i - \mu_i) - \sum_{i<j} \gamma (\sum_{x \in p} P(x) x_i x_j - \nu_i). \quad (12)$$

Donde λ y γ son multiplicadores de Lagrange constantes y uniforme para la distribución. Además:

$$\mu_i = \sum_{x \in p} P(x) x_i, i = 1, ...N. \quad (13)$$

$$\nu_i = \sum_{x \in p} P(x) x_i x_j, i \neq j. \quad (14)$$

A través de la teoría de máxima entropía, llegamos a la mejor distribución de probabilidad, que llamaremos modelo de correlación por pares de máxima entropía:

$$P_P(p) = \frac{1}{Z} exp(-\lambda \sum_i x_i - \gamma \sum_{i<j} x_i x_j). \quad (15)$$

Donde Z es tomada como una constante de normalización. Las correlaciones se pueden evidenciar a través de las distribuciones marginales de distribución:

$$p_{x_i,x_j}(p,p') = \sum_{x \neq x_i, x \neq x_j} P_p(x_i, x_j, x). \quad (16)$$

A través de la ecuación:

$$p_{x_i,x_j}(p,p') = \sum_p \exp\left[V_{12}(p,x_i) + V_{23}(p,x_i)... - 1\right]. \quad (17)$$

Para la cual será necesario primero calcular cada uno de esos potenciales de interacción (6 en total, para la relación de nuestras 4 posiciones).Debido a que el orden de las letras importa, hay seis potenciales independientes. Cuya solución requiere de métodos de iteración para solucionar los conjuntos de ecuaciones numéricamente [15].

Dado que el calculo de estos potenciales es computacional y debido a que los potenciales son matrices de $32x32$, este calculo con su debido resultado se encuentra en el Anexo 1.

Notamos que esta interacción es dada a través de todos los elemento, lo que contrasta por ejemplo con el modelo Markoviano, el cual solo considera correlaciones entre elementos vecinos [19].

3.7. Ley de Zipf

La ley de Zipf es una ley empírica que establece una relación inversa entre la distribución de frecuencia ($P(x_i)$) con la que un elemento se presenta en la realidad y el rango (x_i) de dicho elemento en la lista de frecuencia [16].

$$P(x_i) = \frac{A}{{x_i}^a}. \quad (18)$$

A es una constante de normalización de la función, y a es un real positivo que toma valores ligeramente superiores a 1.

Un gráfico de Zipf es, entonces, aquel que relaciona el rango del elemento con su frecuencia en un conjunto real de datos. Por ejemplo, puede ser utilizado para comparar el rango y la ocurrencia de las palabras en un texto.

Aunque la ley de Zipf fue originalmente descrita en el estudio de lenguajes, su aplicación ha sido extendida a la física y a estudios demográficos, económicos e incluso musicales, que parten de buscar patrones en elementos individuales que se pueden agrupar en diferentes clases según ciertas restricciones y condiciones [15].

Existen algunas correcciones a la ley de Zipf, siendo la más reconocida la de Mandelbrot. En ella se añade un segundo parámetro C dependiente de los datos que permite un poco más de precisión para los valores de bajo rango:

$$P(x_i) = \frac{A}{(1 + C x_i)^a}. \quad (19)$$

Normalmente, modelos basados en el principio de máxima entropía, o el algoritmo de escalado iterativo mejorado reproducen muy fielmente está ley desde una perspectiva general. Sin embargo, presentan una considerable dispersión en la probabilidad individual de algunas palabras, en comparación con las probabilidades empíricas conocidas [15].

4. Estado del Arte

El principio de máxima entropía fue introducido por primera vez en el año 1957 por E. T. Jaynes en [12]. A partir de la Teoría de la información se obtiene un criterio constructivo, con el que se establecen distribuciones de probabilidad con base a un conocimiento parcial. Estas distribuciones conducen a un tipo de inferencia estadística, el cual llamó estimado de máxima entropía. Este estimado es el menos sesgado que se puede obtener a partir de la información disponible, es decir, es máximamente evasivo con respecto a la información faltan te. Por otro lado, Jaynes determinó que la entropía termodinámica es idéntica a la entropía de la teoría de la información (exceptuando la constante de Boltzmann); por lo tanto, la mecánica estadística debe ser vista como una aplicación particular de la inferencia lógica y la teoría de la información.

En 1991, W. Ebeling y G. Nicoli realizaron un estudio sobre la entropía de secuencias simbólicas y el rol de las correlaciones en la probabilidad a priori de ocurrencia de una secuencia dada [18]. Su análisis se basa en el comportamiento de escalamiento de la entropía en función de la longitud de la secuencia, y se diferencia entre las secuencias caóticas o Markovianas, y periódicas. A partir de esto se concluye que las entropías por letra y las entropías relativas decaen asintóticamente y de manera proporcional a la potencia inversa de $\ln n$.

En 1992, E. Black describió cómo extraer reglas gramaticales de texto anotado automáticamente e incorporar estas reglas en modelos estadísticos de gramática [3], así contribuyendo a la capacidad predictiva de los modelos estadísticos del lenguaje natural.

A. L. Berger realizó en 1996 un acercamiento al procesamiento del lenguaje natural desde el principio de máxima entropía [1]. Concorde a dicho principio se debe elegir el modelo con mayor entropía y que además sea coherente con las restricciones. A partir de esto, Berger observó que este modelo era miembro de una familia exponencial con un parámetro ajustable para cada restricción. Los valores óptimos de estos parámetros se obtienen maximizando la probabilidad de los datos de entrenamiento. De esto concluyó que la máxima entropía y máxima probabilidad dan el mismo resultado: el modelo con la mayor entropía consistente con las restricciones es igual al modelo exponencial que mejor predice la muestra de datos.

En 2005, T. Cover definió la entropía conjunta como una medida de la incertidumbre asociada a un conjunto de variables [5]. Luego en 2010, G. J. Stephens planteó el modelo aleatorio como la entropía máxima posible que se puede obtener del modelo completo de distribución conjunta [19], el cual parte de la entropía conjunta. En este se consideran todas las posibles combinaciones del sistema. También planteó que el modelo independiente toma en consideración la aparición de cada letra en cada posición, implicando así una distribución de probabilidad combinada. Por último, Stephens planteó el modelo completo, el cual realiza un análisis empírico donde se construye la matriz de densidad de probabilidad combinada basándose solamente en el análisis del corpus.

En 2010, C. Papadimitriou realizó un análisis de entropía de textos de lenguaje natural en inglés y griego, en donde clasificaba los textos con un sistema jerárquico: las letras y palabras estarían en el nivel más bajo, luego ascendía a frases, oraciones, y finalmente significados [14]. En esta investigación se estudio algunos aspectos de la comunicación entre los niveles de la jerarquía; para esto se calculó la entropía de Shannon y Kolmogorov para los textos y se determinó que ambas mostraban el mismo patrón de dependencia del lenguaje y de la categoría de los textos.

5. Resultados

Se realizó una compilación de libros literarios en formato .txt, siendo así el corpus la suma de todos los textos. A continuación se especifica el nombre de cada obra, el autor y la base de datos de donde esta se obtuvo.

- Cien años de soledad, G. García. Github Gist [13].
- El crimen y el castigo, F. Dostoyevsky. Project Gutengerg [10].
- El retrato de Dorian Gray, O. Wilde. Google Drive [8].
- Don Quijote de la Mancha, M. de Cervantes. Github Gist [7].
- La isla del tesoro, R. L. Stevenson. Project Gutengerg [11].
- Sentido y sensibilidad, J. Austen. Opus Corpus [4].

Para seleccionar estos textos se tuvo en cuenta que la gran mayoría del texto estuviera exclusivamente en español; la Divina Comedia fue considerada al escoger el corpus pero fue descartada por su amplio uso de palabras en italiano. Por otro lado, se implementó el diccionario de la RAE como complemento para análisis finales [9].

5.1. Programa para cuantización y análisis del corpus

En lenguaje Python, haciendo uso de los recursos nombrados con anterioridad, se generó un programa para realizar el análisis del corpus. Los pasos pueden verse mejor resumidos en la sección 0.010 del Anexo 1 a este documento. Los resultados principales se muestran en el Cuadro 1.

Además, se añadieron los algoritmos necesarios para hallar las densidades de probabilidad y entropías de los modelos, que se revisarán a continuación.

5.2. Modelo aleatorio

A partir de la ecuación 3 se determina la entropía del modelo aleatorio:

$$S_r = -\sum_{x_1\in l_1}\sum_{x_2\in l_2}\sum_{x_3\in l_3}\sum_{x_4\in l_4} P_r(x)\log_2 P_r(x).$$

Donde $P_r(x) = 1/32$ debido a que se toman en cuenta 32 caracteres: las 27 letras del alfabeto español y las cinco vocales con tilde. Este valor es igual para cada uno de los textos, ya que todos son textos en español.

	total	únicas	total de 4 letras	únicas de 4 letras
Cien años de soledad	139345	15698	11197	551
Crimen y castigo	160469	27409	14395	707
El retrato de Dorian Gray	70075	15642	6536	634
Quijote de la Mancha	184834	15799	17765	642
La isla del tesoro	81864	15968	7891	624
Sentido y sensibilidad	1938316	140224	169936	2161
Corpus total	2574898	175859	227717	2596

Tabla 1: Resultados principales para cada texto y para el corpus total del programa de cuantización.

$$S_r = -\sum_{x\in l}(1/32)^4\log_2(1/32)^4,$$

$$S_r = -32^4(1/32)^4 4\log_2(1/32),$$

$$S_r = -4\log_2(1/32) = -4(\log_2(1) - \log_2(32)),$$

$$S_r = -4(-\log_2(32)) = 4\log_2(32),$$

$$S_r = 20.$$

5.3. Modelo independiente

A partir de las ecuaciones 4 y 5, podemos hallar la distribución de probabilidad para el modelo independiente. Notamos entonces, que es preciso acudir a los valores F_{lj} y N_{u4} obtenidos previamente:

$$N_{u4} = 2594.$$

Y F_{lj} es una matriz de $32\cdot 4$. Con esos valores, al calcular la productoria, obtenemos una probabilidad para todas las combinaciones posibles entre caracteres. A lo que llamamos $P_i(l)$. Por ejemplo, la palabra *hola*, presenta una probabilidad de $8.395\cdot 10^{-5}$. Mientras la combinación *zzzz*, presenta una probabilidad de $2.571\cdot 10^{-9}$.

Además, añadimos el cálculo de la entropía resultante, según la ecuación 3:

$$S_i = 16.462.$$

El programa elaborado para el análisis de este modelo se encuentra en la sección 0.1.1 del Anexo 1 adjunto.

5.4. Modelo completo

El modelo completo podemos obtenerlo a partir de la ecuación 6. Revisando los resultados obtenidos en el programa de cuantización, sabemos la cantidad de palabras de cuatro letras en el texto N_4.

$$N_4 = 227717.$$

Y para cada combinación, revisamos su frecuencia en el texto. Al resolver la ecuación obtenemos una nueva densidad de probabilidad. En esta ocasión, por ejemplo, la palabra *hola*, presenta una probabilidad de $5.709 \cdot 10^{-5}$. Mientras la combinación zzzz, presenta una probabilidad de 0, pues en ningún texto apareció en ningún momento esa combinación.

La entropía resultante para este caso viene a ser:

$$S_f = 7.527.$$

El programa elaborado para el análisis de este modelo se encuentra en la sección 0.1.2 del Anexo 1 adjunto.

5.5. Modelo de correlación por pares

Cómo se puede ver en el programa, primero se calcularon los valores marginales p_{ij}. Después, a partir de ellos, se construyeron los marginales totales para cada par de posiciones.

Además, se pudo obtener el vocabulario de correlaciones, que establece las palabras cuyas probabilidades de todas sus combinaciones por pares son diferentes de 0.

Corrigiendo y alterando un poco el pseudocódigo presentado en [15], se realizó una iteración numérica para definir los potenciales de interacción, en base a los marginales de probabilidad y la ecuación 15. Una vez establecidos los potenciales, la función nos despeja también la densidad de probabilidades para todas las palabras. Se muestra que, en este modelo, la palabra *hola* tiene una probabilidad un poco superior a la del modelo completo; y la palabra zzzz tiene probabilidad nula. Como estos valores son producto de una iteración, vemos cómo existe una reducción en la condición de normalidad, inferior al $0.001\,\%$.

La entropía resultante en este caso es:

$$S_p = 9.034.$$

El programa elaborado para el análisis de este modelo se encuentra en la sección 0.1.3 del Anexo 1 adjunto.

5.6. Comparación de modelos

Para tener una idea de las correlaciones añadidas por cada modelo frente al aleatorio, utilizamos las ecuaciones 9, 7 y 8:

$$S_{r-i} = S_r - S_i = 20 - 16.462 = 3.538,$$

$$S_{r-p} = S_r - S_p = 10.966,$$

$$S_{r-f} = S_r - S_f = 20 - 7.527 = 12.473,$$

Además, el valor de la multi-información añadida es:

$$I(l) \equiv \sum_j S(l_j) - S_p = S_i - S_p = 7.428.$$

Finalmente, los vocabularios de cada modelo (recordando que son palabras de 4 letras) son:

$$N_r = 2^{S_r} = 1048576,$$

$$N_i = 2^{S_i} \approx 90273,$$

$$N_f = 2^{S_f} \approx 184,$$

$$N_p = 2^{S_p} \approx 524.$$

5.7. Densidad de probabilidad de palabras según longitud

Se hizo un conteo, además, del peso que tienen los tipos de palabras (según que tan extensas son) en cada uno de los textos del corpus y del texto completo. Se diferenció si eran palabras totales o palabras completas. Los resultados generales pueden observarse en las figuras 1 y 2. La información para cada texto está extendida en la sección 0.0.10 del Anexo 1.

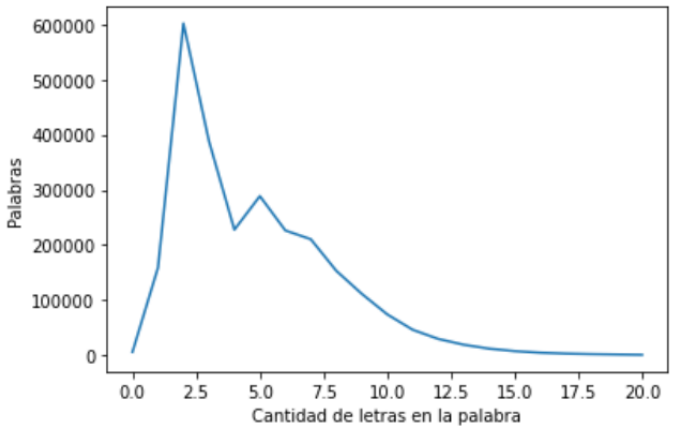

Figura 1: Frecuencia en el *corpus* de tipo de palabras según cantidad de letras. Elaboración propia

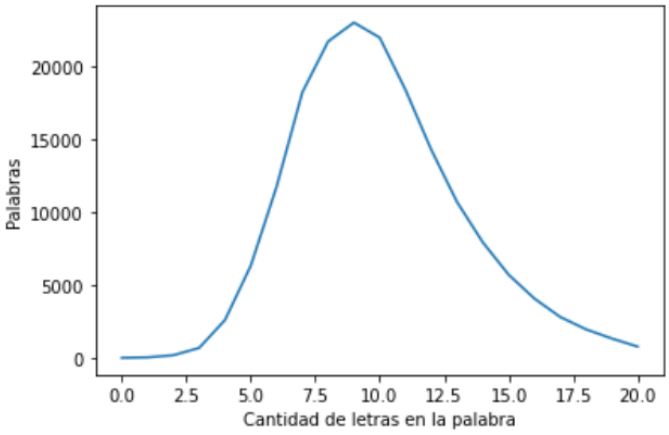

Figura 2: Frecuencia en el *corpus* de tipo de palabras únicas según cantidad de letras. Elaboración propia

5.8. Comparación con otros trabajos

Podemos hacer una comparación entre este estudio y uno similar ejecutado para el idioma inglés, por ejemplo, a partir de las entropías de cada modelo en este trabajo, y las entropía obtenidas en [19]. A continuación se muestra la diferencia porcentual que hay entre los dos estudios:

$$e\,\%_s r = 6\,\%,$$

$$e\,\%_s i = 14.5\,\%,$$

$$e\,\%_s f = 8.1\,\%.$$

Y se pueden calcular, además, las diferencias del vocabulario añadido por cada estudio.

$$N_r - N_r^* = 591600,$$

$$N_i - N_i^* = 72918,$$

$$N_f - N_f^* = 63,$$

$$N_p - N_p^* = 347.$$

6. Análisis de resultados

6.1. Corpus

El *corpus* utilizado provee una colección de 2574898 palabras totales, de las cuales 227717 son de cuatro letras. Estas cantidades son comparables con algunos trabajos consultados (420 mil palabras totales en inglés y 250 mil palabras en griego [14], o 676302 palabras totales y 135441 de cuatro letras en inglés [19]). Además de compartir la naturaleza de los texto, siendo estos principalmente trabajos literarios no técnicos (literatura universal, por ejemplo).

6.2. Modelo Aleatorio y Modelo Independiente

Ambos casos son una primera aproximación al ejercicio propuesto y son puntos de comparación para los otros modelos. Sin embargo no se puede decir que son fieles a las estructuras de nuestro idioma, pues presentan probabilidades relativamente altas para combinaciones de letras totalmente contrarias a las reglas ortográficas y fonéticas del idioma (Por ejemplo, *aaaa* presenta una probabilidad del orden de $2 \cdot 10^-5$ para el modelo independiente y de 10^-6 para el aleatorio, cuando, obviamente, en el modelo completo no aparece en ningún momento)

6.3. Modelo Completo

El modelo completo se caracteriza por llevar a 0 gran parte de las probabilidades, al no contar las palabras que no existen en el *corpus*. Para este caso, los valores diferentes a 0 son apenas un $3.75\,\%$ de todas las probabilidades.

6.4. Modelo de correlación por pares

El modelo de correlación tiene una entropía superior a la del modelo completo (lo cual era de esperarse), e inferior a los demás modelos. Corresponde al modelo de máxima entropía con las restricciones dadas (potenciales de interacción).

El programa ayudó a revisar cuántas palabras coincidían con las del modelo completo, siendo esta cantidad 278 palabras. De las 246 restantes, a través de una comparación con el diccionario, solamente 34 existen realmente, lo cual nos deja con 212 palabras creadas por el modelo, que, aunque fonéticamente corresponden con la cotidianidad del idioma español, no existen realmente.

Entre las palabras predichas, vemos que encabezan en probabilidad *piro* y *pira*, que, aunque no son muy comunes, hacen parte del idioma español. El dato contrasta, por ejemplo, con las palabras *sapa*, o *moto*, que, aunque más comunes, no aparecen en los textos, y el modelo les da una baja probabilidad.

6.5. Comparación con otros trabajos

La entropía y los vocabularios de cada modelo son mayores para el idioma español. Esto puede darse, principalmente, debido a que se decidió hacer distinciones de acentos para no perder información semántica (diferenciar *papa* de *papá*, por ejemplo), mientras en inglés se restringieron a los 26 caracteres del alfabeto. Ninguno de los dos mecanismos ahonda en el problema de las palabras homónimas, por ejemplo, lo cual, aunque de mayor complejidad, podría dar una visión ampliada de las palabras de cuatro letras en cada idioma.

Mientras en los trabajos de [19], el modelo final logra capturar el $92\,\%$ de la multi información, en nuestro caso es de $83.32\,\%$, lo cual sigue siendo cercano.

Sin embargo, la diferencia no es abismal, lo cual se debe, en parte, a que ambos lenguajes comparten similitudes en sus modos fonéticos y en la cantidad de caracteres, en comparación, por ejemplo, a idiomas como el Taa, que maneja alrededor de 164 consonantes.

Los mecanismos de análisis por correlaciones también se pueden extender, de distintas maneras. Por ejemplo, añadiendo las correlaciones de tríos de letras, o, estableciendo el caracter vacío, que hubiera permitido analizar también palabras de una, dos y tres letras.

7. Conclusiones

- Se halla por medio del programa de análisis de palabras que el corpus está compuesto por 2574898 palabras totales, de las cuales 227717 son palabras únicas; Comparándolo a los corpus implementados en los artículos base, se tiene que el tamaño de la colección es adecuado para realizar el estudio propuesto.
- Los modelos aleatorio e independiente presentan probabilidades relativamente altas para palabras que no existen en el español; luego, se considera que no son modelos fieles a las estructuras del idioma.
- El modelo completo presenta resultados considerablemente confiables, pero este podría ser más exacto al español si se incluyeran corpus que contenga un lenguaje técnico o científico.
- Semejanzas fonéticas y de estructura del lenguaje permiten pocas diferencias de entropía entre los modelos del inglés y el español para palabras de cuatro letras, siendo las diferencias debidas principalmente a los caracteres adicionales utilizados para el español.
- El modelo por correlaciones alcanza a capturar un porcentaje aceptable de la multiinformación tomando en cuenta los artículos y estudios de referencia.
- Una principal diferencia respecto a los resultados de otros estudios, es la cantidad de nuevas palabras que considera el modelo de correlaciones que no están en el corpus, siendo casi la mitad de las palabras con mayor probabilidad.

Referencias

[1] Stephen A. Della Pietra y Vincents J. Della Pietra Adam L. Berger. «A Maximum Entropy Approach to Natural Language Processing». En: *Computational Linguistics* 22 (2002).

[2] Jascha Sohl-Dickstein y Michael R. DeWeese Badr F. Albanna Christopher Hillar. «Minimum and Maximum Entropy Distributions for Binary Systems with Known Means and Pairwise Correlations». En: *Entropy* 427.19 (2017), págs. 1-33. DOI: 10.3390/e19080427.

[3] Ezra Black y col. «Towards History-based Grammars: Using Richer Models for Probabilistic Parsing». En: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. 1992. URL: https://www.aclweb.org/anthology/H92-1026.

[4] Opus corpus. *Sentido y sensibilidad.txt*. https://opus.nlpl.eu/Books.php.

[5] Thomas M. Cover y Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006. ISBN: 0471241954.

[6] Gabriela M. Marinescu Dan C. Marinescu. *Classical and Quantum Information*. Academic Press, 2011.

[7] J. Dario. *Don Quijote de la Mancha.txt*. https://gist.github.com/jsdario/6d6c69398cb0c73111e49f1218960f79.

[8] *El retrato de Dorian Gray.txt*. https://drive.google.com/file/d/0B2r-leeVzSZdOWExZ\DFlNjAtYmFlYS\MGYxMDk1/view?resourcekey=0-6VGNafB1MyQI6p4nx8iiVQ.

[9] Giusseppe. *Diccionario de la RAE.txt*. https://www.giusseppe.net/blog/archivo/2015/10/29/diccionario-de-la-rae-en-modo-texto-plano/.

[10] Project Gutenberg. *El crimen y el castigo.txt*. https://www.gutenberg.org/ebooks/61851.

[11] Project Gutenberg. *La isla del tesoro.txt*. https://www.gutenberg.org/ebooks/45438.

[12] E. T. Jaynes. «Information Theory and Statistical Mechanics». En: *The Physical Review* 106.04 (1957), págs. 620-630.

[13] I. Jimenez. *Cien años de soledad.txt*. https://gist.github.com/ismaproco/6781d297ee65c6a707cd3c901e87ec56. 2018.

[14] C. Papadimitrioua y K. Karamanos y F.K. Diakonos y V. Constantoudis y H. Papageorgiou. «Entropy analysis of natural language written texts». En: *Physica A* 389.16 (2010). DOI: 10.1016/j.physa.2010.03.038.

[15] A. Corral y M. García. «From Boltzmann to Zipf through Shannon and Jaynes». En: *Entropy* 22.2 (2020). DOI: 10.3390/e22020179.

[16] Marcelo A. Montemurro. «Beyond the Zipf–Mandelbrot law in quantitative linguistics». En: *Physica A* 300 (2001).

[17] Elad Schneidman y col. «Network Information and Connected Correlations». En: *Physical review letters* 91 (ene. de 2004), pág. 238701. DOI: 10.1103/PhysRevLett.91.238701.

[18] G. Nicoli W. Ebeling. «Entropy of Symbolic Sequences: the Role of Correlations». En: *Europhysics Letters* 14 (1991), pág. 191.

[19] Greg J. Stephens y William Bialek. «Statistical mechanics of letters in words». En: *PHYSICAL REVIEW E* 81.06 (2010). DOI: 10.1103/PhysRevE.81.066119.