Examen General de Conocimientos

Atribución de Autoría
EL MÓNDRIGO

Samuel Ramos Sosa

October 29, 2018

Contenido

Resumen

El presente trabajo se presenta como parte del Exámen General de Conocimientos a realizar para la obtención del grado de **Maestro en Ciencia e Ingeniería de la Computación**, otorgado por el **Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas** (IIMAS) de la **Universidad Nacional Autónoma de México**.

Este trabajo, encargado por el Dr. Gerardo Eugenio Sierra Martínez, consiste en describir y de ser posible poner en práctica, buna metodología que permita la atribución de autoría a la novela *EL MÓNDRIGO*, dados cinco autores candidatos.

Parte 1
Cuantificación de Textos

1 Conceptos Básicos

1.1 *Corpus*

Al conjunto organizado de textos bajo estudio se le conoce como *corpus*.

En el caso de este trabajo el *corpus* consiste en el texto completo de la novela anónima **El Mondrigo** y los textos disponibles de los autores probables de esa novela, para fines de comparación.

1.2 Estilometría

Estilometría, más comunmente referida por su nombre en inglés *Stylometry* se refiere a la medición del estilo de escritura de un autor. Esta medición no es una sóla sino que es la colección de técnicas que se pueden usar para cuantificar ciertos atributos de la escritura, y que pueden usarse en un ejercicio de comparación para detectar si un texto fue escrito por un autor en específico, siempre y cuando el texto en debate sea lo suficientemente largo y la obra del escritor a comparar lo suficientemente extensa.

En una definición más precisa [3] la estilometría es el proceso de cuantificación de estilo usando distintas mediciones para componer un vector, un punto en un espacio multidimensional.

Lo que resta es implementar una medición de distancia entre vectores del mismo tipo que permita establecer un parámetro de cercanía entre dos textos codificados de esta manera.

Las métricas pueden ser de carácter léxico, sintáctico, semántico o gramatical, y todas se obtienen a partir de los textus de un *corpus*, en un proceso denominado **Extracción de Características** o *Feature Extraction*.

1.3 Modelo *Bag of Words*

La primera técnica que revisaremos es para la extracción de características es el modelo llamado *Bag of Words* (BoW), o Bolsa de Palabras.

Esta representación del texto es una simple medida de ocurrencia de palabras. Típicamente se elabora un diccionario común con todas las palabras únicas existentes en todos los documentos a analizar, donde cada una de esas palabras es una ćaracterística.

Por ejemplo, si tenemos las siguientes oraciones:

1. *esta frase es la primera frase*
2. *esta frase es la segunda frase*
3. *esta frase es la tercera frase*

observamos que tenemos las siguientes siete palabras distintas, que conforman el *vocabulario*:

$$vocabulario = [esta, es, la, primera, segunda, tercera, frase]$$

Así, de la primera oración obtenemos la siguiente representación:

$$[1, 1, 1, 1, 0, 0, 2]$$

que significa que, del vocabulario establecido, en esa oración existe una ocurrencia de la primera palabra del vocabulario (*esta*), una ocurrencia de la segunda palabra (*es*), una ocurrencia de la tercera (*la*), una ocurrencia de la cuarta (*primera*), ninguna ocurrencia de la quinta palabra (*segunda*), ninguna ocurrencia de la sexta (*tercera*) y finalmente dos ocurrencias de la séptima palabra (*frase*). En otras palabras, el vector contiene las *frecuencias* de las palabras del vocabulario en la oración.

De manera similar los vectores de las oraciones restantes serán $[1, 1, 1, 0, 1, 0, 2]$ y $[1, 1, 1, 0, 0, 1, 2]$ respectivamente.

Estos vectores, junto con el vocabulario, nos dan la una posible representación numérica de los textos, en este caso las oraciones de ejemplo.

Es importante notar que si bien en estos ejemplos hemos usado la suma de ocurrencias de la palabra del vocabulario en el texto como entrada del vector, esta también podría ser la frecuencia de la palabra en relación con el número de palabras del texto. En lugar de 2 para la suma de ocurrencias de la palabra *frase*, tendríamos $\frac{2}{6}$ porque la oración tiene 6 palabras. La ventaja de esto es que la suma de frecuencias de cada vector siempre será 1. Esta normalización además ayuda a equalizar dos documentos de distintos tamaños que de manera natural tienen ocurrencias distintas (función del tamaño) pero que su frecuencia relativa es comparable. A esta frecuencia relativa se le conoce como *tf*, por las siglas en inglés de *Term Frequency*.

Cuando los textos a analizar son cuantiosos y por lo tanto el vocabulario de palabras únicas resultado de su unión tiende también a ser muy grande, lo que obtenemos son vectores con muchos 0s. Es decir, de cada texto a analizar hay muchas entradas de vocabulario que no se usan y cuya frecuencia es 0.

Hay muchas maneras de atacar este problema como lo son:

- Ignorando puntuación y capitalización.
- Normalizando palabras arreglando errores ortográficos, aunque para algunos análisis el error ortográfico pueda ser significativo.
- Eliminar vocablos inútiles. Este es el objetivo generalizado, pero va a depender del tipo de análisis que se quiere llevar a cabo. Mientras que para algunos análisis las palabras únicas como nombres propios o palabras inventadas por el autor son eliminables, en otros son estas mismas palabras las que son importantes. En otros casos es útil eliminar elementos gramaticales comunes como preposiciones, conjunciones y artículos, en tanto que en otros si es importante incluirlos.
- Normalización de palabras de acuerdo a su raíz. No nos referimos a una raíz etimológica (aunque para esto la etimología es útil), sino que tratar de normalizar, por ejemplo, todas las conjugaciones de un verbo al verbo en sí o las declinaciones de una palabra para que cada ocurrencia de una conjugación o declinación se refiera al mismo vocablo e.g. $[jugamos, jugando, jugaba] = jugar$, $[silla, sillas, sillon] = silla$. En estos ejemplos vemos que también el uso de esta técnica es situacional, pues claramente una *silla* no es lo mismo que un *sillon*, pero nos deja ver que tampoco es gramatical. No hay ninguna razón para un hacer una equivalencia más fácil de llevar a cabo en un algoritmo, como $[jugando, jugaba] = jug$ que si bien no tiene sentido semántico claro, funciona perfectamente bien como característica del texto en el vector.
- Finalmente, la agrupación de palabras para usar la frecuencia de este agrupamiento como característica del texto. A este agrupamiento se le conoce como *n-grams*. Por ejemplo, *cielo azul* es un *2-gram*, o *bigrama*.

La frecuencia de los términos es, como hemos visto, importante y significativa, pero en ocasiones es importante también restarle importancia a los términos que aparecen con mucha frecuencia en el documento pero cuyo significado no es tan importante como el de aquellos de menor frecuencia pero mayor significado.

Para esto, la frecuencia del término se multiplica por un factor de penalización de frecuencia llamado *Inverse Document Frequency*.

1.4 Modelo *Parts of Speech*

Este modelado se basa en un análisis sintáctico del corpus que clasifica cada término del mismo de acuerdo a su función gramatical: Verbo, Adverbio, Sustantivo, etc.

Esta clasificación se hace con algún algoritmo de clasificación entrenado en un idioma en específico. A este proceso se le llama *POS Tagging*, y el resultado

es, por palabra, una etiqueta, o *tag*, de acuerdo al tipo de palabra que es dentro de la oración.

Para fines de este ejercicio utilizaremos el *Stanford Postagger 3.9.2* disponible en *https://nlp.stanford.edu/software/tagger.shtml*.

Esta librería es un empaquetado *jar* que se puede importar vía la librería **nltk** de *Python*. Usamos el modelo *spanish-ud.tagger*.

Desafortunadamente experimentos sencillos realizados con el siguiente código generaron resultados poco confiables:

```
from nltk.tag.stanford import StanfordPOSTagger

model_path = '/stanford-postagger-full-2018-10-16/models/spanish-ud.tagger'
jar_path = '/stanford-postagger-full-2018-10-16/stanford-postagger-3.9.2.jar'
spanish_postagger = StanfordPOSTagger(model_path, jar_path, encoding='utf8')

sentences = ['esta es una primera oracion de primavera',
        'para destapar el tubo hay que usar drano',
        'la fiesta fue en el zocalo de la ciudad']

for sent in sentences:
        words = sent.split()
        tagged_words = spanish_postagger.tag(words)

        for (word, tag) in tagged_words:
                print(word + ' ' + tag)
```

que dió como resultado el listado:

```
esta VERB
es NOUN
una DET
primera ADJ
oracion NOUN
de ADP
primavera NOUN
para SCONJ
destapar VERB
el DET
tubo NOUN
hay AUX
que CONJ
usar VERB
drano NOUN
la DET
fiesta NOUN
fue AUX
en ADP
el DET
zocalo NOUN
de ADP
la DET
ciudad NOUN
```

Parte 2
Análisis

2 Metodología

2.1 Textos

El *corpora* se compone de textos de cinco autores y el archivo correspondiente al texto anónimo en cuestión. El documento anónimo es la novela ***El Móndrigo***.

Los autores candidatos para su autoría son:

- Emilio Uranga (detalle de los textos en la tabla 1)
- Ortega Molina (detalle de los textos en la tabla 2)
- Jorge Joseph (detalle de los textos en la tabla 3)
- Gregorio Ortega Hernández (detalle de los textos en la tabla 4)
- Roberto Blanco Moheno (detalle de los textos en la tabla 5)

	TEXTO
1	*Artículo* ***Universidad Popular****. La Prensa, 20 de julio de 1968, pp. 3 y 35*
2	*Artículo* ***La Represión Extremada****. La Prensa, 1968, pp. 3 y 38*
3	***La Ambigedad Universitaria****. La Prensa, 1968, pp. 3 y 38*
4	***Ornato y Orden****. La Prensa, 1968, pp. 3 y 47*
5	***Viaje del Canciller****. La Prensa, 1968, pp. 3 y 32*
6	***Prólogo de Astucias literarias****, Mxico, Federacin Editorial Mexicana, 1971, p. 9*
7	***De Astucias literarias****, México, Federación Editorial Mexicana, 1971, p. 31-35 (8-I-70)*
8	***De Astucias literarias, México****, Federación Editorial Mexicana, 1971, p. 228-233 (11-II-70)*
9	***Lectura de galeras de De Astucias literarias****, México, Federación Editorial Mexicana, 1971, p. 299-303*

Tabla 1: Textos Emilio Uranga

2.2 Proceso

Para poder analizar y cuantificar el corpora, fue necesario convertir los textos de .DOC/.DOCX a texto plano codificado UTF-8.

	TEXTO
1	***Hacia la extrema derecha?****, Revista de Amrica, México, 1968, 21 diciembre 1968.*
2	***La guerra de los zombies****, Revista de América, México, 1968, octubre 1968*
3	***La sociedad del átomo****, Revista de América, México, 1968, julio 1968.*
4	***Hacia la extrema derecha?****, Revista de Amrica, México, 1968, 21 diciembre 1968.*

Tabla 2: Textos Ortega Molina

	TEXTO
1	***Picaluga Vende a los Rojo Gomistas por una Curul Senatorial****, de El ministro del odio, México, Edición del autor, 1961, pp. 35-40*
2	***Pitoloco en Palacio****, de El ministro del odio, México, Edición del autor, 1961, pp. 119-124*
3	***Aviso****, de El ministro del odio, México, Edición del autor, 1961, p. 125*
4	***Pruebas de la existencia de la Atlántida****, de México: cuna de la civilización universal, México, Ramírez Editores, 1965, pp. 87-96*
5	***Esoterismo en Anahuac****, de México: cuna de la civilización universal, México, Ramírez Editores, 1965, pp. 263-277*

Tabla 3: Textos Jorge Joseph

Adicionalmente, el texto en disputa, *El Móndrigo*, que se encontraba ya seccionado, se unió en un sólo corpus.

Todos los análisis se realizaron con scripts *Python 3* utilizando las librerías **NLTK**, **NUMPY**, **MATPLOTLIB**, **PANDAS** y **SKLEARN**, incluidos en la entrega de este trabajo.

Además de la bibliografía anotada al final de este documento, fueron de gran utilidad los tutoriales [4] y [5].

El proceso para el análisis del corpora fué el siguiente:

1. Se crearon varios diccionario de datos donde las llaves son los autores. Estos diccionarios varían dependiendo del análisis pero sirven para mantener el orden de los autores.

2. Se desarrollaron las rutinas para la extracción de características.

3. Se ejecutaron los scripts y se anotaron los resultados en tablas y/o en gráficas.

	TEXTO
1	***Mxico resistió a la C.I.A. y a Fidel Castro**, Revista de América, 2 de noviembre de 1968.*
2	***Un chamaco de Tepito da una lección de civismo**, Revista de América, 27 de abril de 1968.*
3	***Un injusto artículo de Blanco Moheno**, Revista de América, julio de 1968*

Tabla 4: Textos Ortega Hernández

	TEXTO
1	***Tlatelolco. Historia de una infamia**, México, 1969, Editorial Diana, pp. 193-206*
2	***La noticia detrás de la noticia**, México, 1966, Edición de autor, pp. 103-105*
3	***La noticia detrás de la noticia**, México, 1966, Edición de autor, pp. 180-190*
4	***La noticia detrás de la noticia**, México, 1966, Edición de autor, pp. 208-211*

Tabla 5: Textos Roberto Blanco Moheno

3 Mediciones

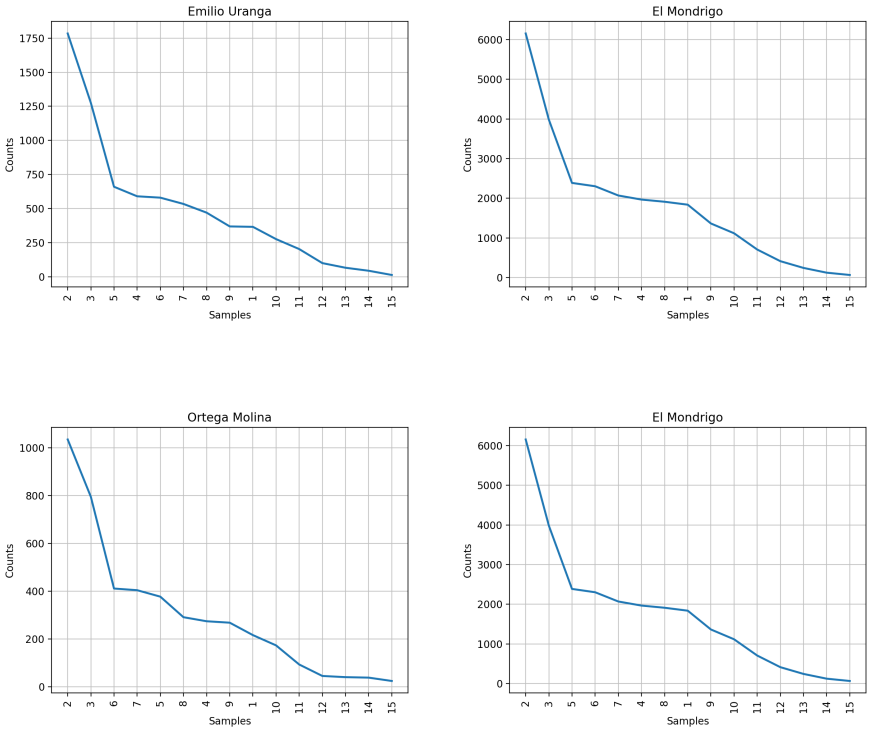
3.1 Método de Mendenhall

Esta técnica es la más antigua y la menos confiable, incluída aqui por su facilidad de implementación y con el espíritu de darle al trabajo una connotación histórica.

Mendenhall en 1887 [6] propuso un método de identificación de autoría basado en la longitud media de las palabras usadas por el autor en sus escritos, bajo la creencia de que un autor tenía lo que el llamó *word spectrum* o espectro de palabras i.e. distribución de longitud de palabra a lo largo de un texto.

Mendenhall hizo experimentos tomando distintos trabajos de un autor (Dickens) midiendo la curva de los trabajos y curvas existentes en un mismo trabajo dividido en secciones. Comparó con otros autores como Thackeray o Mills.

Concluyó que era poco probable que dos autores presentaran la misma curva. Así, usaremos este método para generar curvas de todos los autores y del texto anónimo y luego las compararemos, para ver que autor presenta una curva más parecida a la curva generada con el texto bajo estudio.

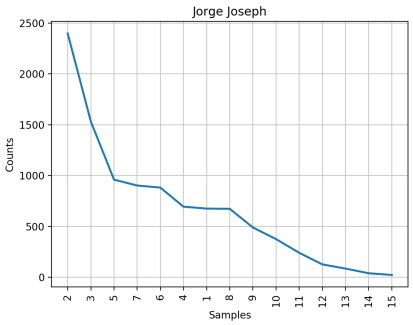


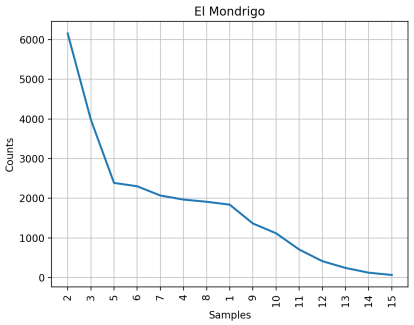
La implementación consiste en los siguientes pasos:

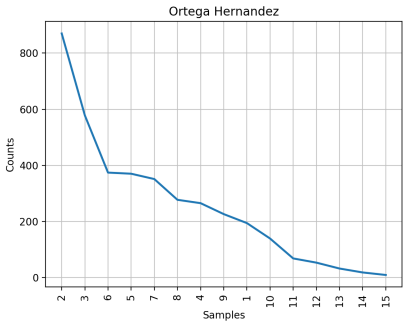
1. Agrupar todos los textos del autor en un sólo texto.
2. Convertir ese texto (string) en una lista de términos (list), que incluyen palabras y puntuación.
3. Eliminar la puntuación para tener una lista de palabras.
4. Crear una lista de la misma dimensin de la anterior, pero que contenga la longitud de cada palabra.
5. Obtener la distribución de estas longitudes.

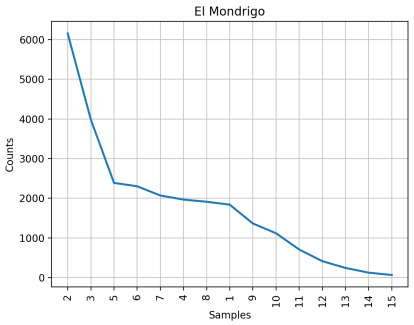
Para cada autor, los resultados se pueden apreciar en las correspondientes gráficas donde se muestra la gráfica resultante de cada autor al lado de la gráfica resultante del texto anónimo, para fines de comparación.

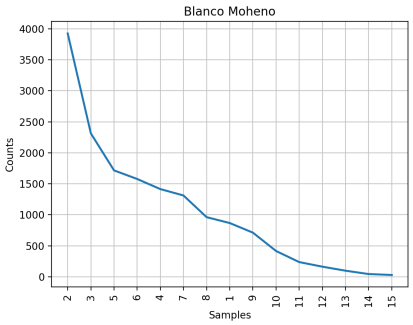
Al comparar las curvas correspondientes a cada uno de los autores con la curva del texto anónimo, podemos apreciar que, sin ser idénticas, las curvas correspondientes a Jorge Joseph y Emilio Uranga son las que presentan mayor grado de semejanza.

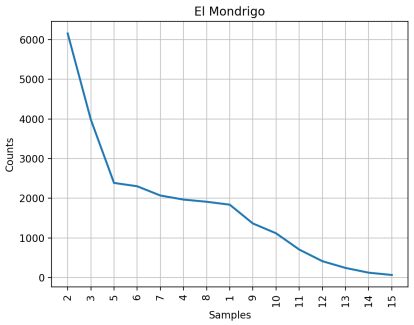

Jorge Joseph
Counts
Samples


El Mondrigo
Counts
Samples


Ortega Hernandez
Counts
Samples


El Mondrigo
Counts
Samples


Blanco Moheno
Counts
Samples


El Mondrigo
Counts
Samples

3.2 Método de Kilgariff

Este método, propuesto por Adam Kilgariff en [7] introduce en este trabajo la idea de *distancia* entre textos que supone que si un autor escribe dos textos, y ambos se comparan, la *distancia* entre ellos debe ser muy pequeña o suficientemente pequeña para llegar a la conclusión de que ambos textos provienen del mismo autor.

Kilgariff hace una comparación entre distintos tipos de medidas para semejanza de corpora utilizando conteo de palabras, y lleva a cabo sus experimentos dividiendo el texto de un autor en secciones y mezclandolas en dos corpora, para medir las diferencias estadísticamente. Su conclusión es que la mejor medida es la llamada χ^2 (fórmula 1) que es un índice típicamente utilizado para medir independencia entre variables aleatorias categóricas (es decir, no numéricas).

$$\chi^2 = \sum_i \frac{(C_i - E_i)^2}{E_i} \tag{1}$$

El detalle es que en los experimentos de Kilgariff se ocupa un sólo texto dividido, y para lograr un efecto similar lo que procede es que se toman los n términos más comunes de cada corpus y se calcula el valor esperado de las ocurrencias de cada una de estas n palabras como si las dos corpora fueran muestras aleatorias de la misma población. Tomando el ejemplo de [7], si los tamaños de las corporas 1 y 2 fueran respectivamente N_1 y N_2, y la palabra w ha sido observada respectivamente con frecuencias o_{w1} y o_{w2}, entonces los valores esperados de ocurrencia de la palabra w en cada una de las corpora serán los mostrados en las ecuaciones 2 y 3.

$$e_{w1} = \frac{N_1 \times (o_{w1} + o_{w2})}{N_1 + N_2} \tag{2}$$

$$e_{w2} = \frac{N_2 \times (o_{w1} + o_{w2})}{N_1 + N_2} \tag{3}$$

Con estos valores esperados, ya es cuestión de sistemáticamente obtener los sumandos de la fórmula 1 para obtener el valor final de χ^2

La implementación consiste en los siguientes pasos:

1. Agrupar todos los textos del autor en un sólo texto.
2. Convertir ese texto (string) en una lista de términos (list), que incluyen palabras y puntuación.
3. Eliminar la puntuación para tener una lista de palabras.
4. Para cada autor, juntar su corpus y el corpus del text anónimo en una sóla lista.

5. Encontrar la ocurrencia de cada término y elegir los n más frecuentes.

6. Calcular el valor esperado de cada palabra de las n más frecuentes tomando la contribución de cada autor a este corpus conjunto que se calcula con el número de términos del corpus de cada uno entre el número de términos del corpus conjunto y multiplicandolo por el número de ocurrencias de esa palabra en el corpus de cada autor.

7. Se va calculando asimismo para cada palabra un término cuyo valor es el cuadrado de la diferencia entre ese valor esperado y el número de ocurrencias en el corpus de cada autor, dividido entre el valor esperado.

8. χ^2 es la suma de todos estos términos de acuerdo a la fórmula 1.

El resultado final de comparar los textos de cada autor con el texto en disputa de acuerdo al método de Kilgariff tomando las 500 palabras más comunes se muestran en la tabla 6 donde concluimos que de acuerdo a la menor distancia (χ^2) es más probable que el autor del texto anónimo haya sido Jorge Joseph, seguido de cerca por Emilio Uranga, en tanto que el menos probable es Roberto Blanco Moheno.

AUTOR	DISTANCIA χ^2
Emilio Uranga	1771.0364889378193
Gregorio Ortega Molina	2002.4749684727008
Jorge Joseph	1643.598487830984
Gregorio Ortega Hernández	1986.895626691162
Roberto Blanco Moheno	3221.980751878392

Tabla 6: Resultados método de Kilgariff

3.3 Método DELTA de John Burrows

Sofisticando el concepto de distancia, nos encontramos con [8]. Burrows propone un método que, dado un documento de prueba y una serie de documentos cuya autoría esta plenamente definida, arroja una medida de distancia con todos los autores, identificando una *estilo* particular de cada autor y midiendo la distancia entre ese estilo y el texto en disputa, el de prueba. Esta medida se llama Δ y entre más pequeñ mas certeza habrá del origen de ese texto.

En resumen, el método consiste en:

1. Crear, con todos los textos del corpora disponibles, un gran corpus.

2. Buscar las palabras más frecuentes y usarlas como atributos o características (dimensiones).

3. Encontrar la proporció con la que contribuye cada autor a esa característica.

4. Calcular el promedio y la desviación estandar y usarlas como atributos de el corpus conjunto.

5. Para cada uno de los atributos (palabras) y usando los valores calculados en el punto anterior, calcular el valor Z donde C_i es el promedio de los promedios como la frecuencia de esa palabra en el corpus, y aplicando la fórmula 4.

6. Calcular Δ comparando el corpus de cada autor con el corpus en disputa de acuerdo a la fórmula 5 donde los subíndices $c(i)$ y $t(i)$ se refieren al atributo del autor y de prueba respectivamente.

$$Z_i = \frac{C_i - \mu_i}{\sigma_i} \tag{4}$$

$$\Delta_c = \sum_i \frac{|Z_c(i) - Z_t(i)|}{n} \tag{5}$$

El resultado final de comparar los textos de cada autor con el texto en disputa de acuerdo al método DELTA de Burrows es el que se muestra en la tabla 7. Podemos observar que el autor que tiene el menor DELTA y por lo tanto el que con mayor probabilidad es el autor del texto anónimo es Jorge Joseph, que es consistente con el resultado del método de Kilgariff. Sin embargo, es importante notar que en este método Emilio Uranga es el que tiene la menor probabilidad.

AUTOR	DELTA
Emilio Uranga	1.5134148360747048
Gregorio Ortega Molina	1.0422030585665234
Jorge Joseph	0.5190645312503361
Gregorio Ortega Hernández	1.2069074912312774
Roberto Blanco Moheno	1.2278603090115858

Tabla 7: Resultados método DELTA de Burrows

3.4 Aprendizaje no supervisado

Los dos últimos métodos que utilizamos (Kilgariff y Burrows) utilizan una medición estadística donde la característica del texto que se toma más en cuenta es la frecuencia de ocurrencia de palabras.

Esta frecuencia es solamente una de las posibles métricas que se le pueden asociar a un texto para poder compararlo con otro.

En esta sección trataremos de clasificar la corpora en *clusters* o agrupaciones con el objetivo de verificar dos cosas: Si los cinco autores en efecto pertenecen

a cinco agrupaciones distintas y si el texto en disputa queda dentro uno de esas cinco agrupaciones, lo que establecería de acuerdo a este método su autoría.

Esta clasificación la haremos con un algoritmo de aprendizaje no supervisado. Al igual que en el aprendizeje supervisado para poder llevar a cabo cualquier algoritmo de aprendizaje de máquina es necesario asociar a cada texto un vector de características que pueden ser:

- Promedio de palabras por oración
- Variación en la longitud de oraciones
- Diversidad léxica, que representa la riqueza de vocabulario del texto
- Ocurrencia de puntuación dentro de las oraciones

Así, el procedimiento general es:

1. Preparación del corpora. Esto ya lo hicimos para los métodos anteriores.
2. Extracción de características. Este importante proceso se le conoce por su nombre en inglés *feature extraction* y es la colección de métodos para obtener las características ejemplificadas antes.
3. Clasificación. Este es el resultado del proceso, es decir, la clasificación de cada texto de acuerdo a los agrupamientos encontrados.

Asi, utilizaremos *feature vectors* o vectores de características, para codificar los textos y poderlos clasificar.

Aqui haremos entonces es crear vectores de características, uno por autor, y utilizaremos el método de *k-means* con 5 clusters para clasificar esos seis vectores. El resultado entonces es que por característica obtendremos una lista de 6 etiquetas con 5 posibles etiquetas. El objetivo es que es que haya una etiqueta repetida, correspondiente a los dos textos que fueron escritos por el mismo autor.

Las características que vamos a extraer en forma de vector son:

- **VECTOR LEXICO**. Palabras por oración de cada autor, omitiendo la puntuación. Obtendremos un vector por autor que se componga de el promedio de palabras por oración de entre todas las oraciones del corpus del autor, la desviación estandar y la diversidad léxica del autor en un dato obtenido dividiendo las palabras únicas que se presentan en el corpus entre el total de palabras.
- **VECTOR PUNTUACION**. Puntuación por oración. Se contarán las ocurrencias de comas (,), punto y coma (;) y dos puntos (:) en las oraciones del corpus del autor. Se obtendrá un vector por autor que tenga también tres componentes, cada uno correspondiente a la cuenta de los signos de puntuación antes mencionados.

- **VECTOR BOW**. Se utilizará el modelo *Bolsa de Palabras* como lo explicamos en la sección 1.3 pero simplificando el modelo. Lo que vamos a obtener es obtener es un vector que utiliza las 10 palabras más comunes encontradas en todos los corpora y usarlas como características del vector, donde cada componente del vector será el número de ocurrencias de cada palabra. Asi, obtendremos seis vectores de dimensión 10.
- **VECTOR POS**. Se utilizará el modelo *Parts of Speech* visto en la sección 1.4 utilizando las cinco etiquetas más frecuentes arrojadas por el *tagger* y usando esas como características de un vector por autor, de dimensión 5.

Entonces, cada vector supondrá un entrenamiento independiente y una obtención de agrupamientos distinta.

Lo que se mantiene constante es el orden de los autores dentro de los vectores obtenidos. Es decir, el primer vector corresponde al primer autor, que es Emilio Uranga, y el último vector corresponde al texto anónimo.

El entrenamiento se realiza con el siguiente código, donde v es el vector, que puede ser cualquiera de los cuatro definidos.

```
k_means = KMeans(n_clusters=5, init='k-means++', n_init=10, verbose=0)
k_means.fit(v)
k_means_labels = k_means.labels_
print(k_means_labels_)
```

Los resultados obtenidos se muestran en la tabla 8. Se muestran los autores en orden como columnas, en tanto que los renglones son cada una de las pruebas realizadas con este método. El número que aparece en las celdas es la etiqueta que el algoritmo asoció al vector de entrada. El número en sí no significa nada, pero si un número aparece en un mismo renglón en más de una celda correspondientes a múltiples corpora, eso quiere decir que los textos de esos corpora fueron escritos por el mismo autor.

El orden de los autores es Emilio Uranga(1), Gregorio Ortega Molina(2), Jorge Joseph(3), Gregorio Ortega Hernández(4) y Roberto Blanco Moheno(5). El texto anónimo se encuentra en la última celda (6).

	1	**2**	**3**	**4**	**5**	**6**
LEXICO	2	1	0	3	4	0
PUNTUACION	4	2	3	3	0	1
BOW	3	0	1	2	4	1
POS	*	*	*	*	*	*

Tabla 8: Resultados Aprendizaje no supervisado vía *k-means*

Como podemos ver, las pruebas LEXICO y BOW arrojan resultados consistentes con nuestras otras mediciones, pues ponen en la misma agrupación al

autor 3 y 6, es decir, al corpus de Jorge Joseph y al corpus anónimo.

La medición PUNTUACION presenta probablemente errores en su diseño pues junta en una misma agrupación al corpus de Jorge Joseph y Gregorio Ortega Hernández.

La medición POS no se pudo llevar a cabo porque la interacción con el *tagger* llevaba mucho tiempo. Quizá con un poco de más tiempo se podría poner un servidor con suficiente RAM en esta tarea.

Parte 3
Conclusión

Observamos una consistencia en las conclusiones de cada uno de los metodos señalando a **Jorge Joseph** como el más probable autor de *El Mondrigo* pero sin haber hecho un trabajo expermiental exhaustivo que nos permita reproducir el resultado alterando las variables de los modelos.

Referencias

[1] *Helena Gómez-Adorno, Juan-Pablo Posadas-Duran, Germán Ríos-Toledo, Grigori Sidorov, Gerardo Sierra,* ***Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts****, Computación y Sistemas, Vol. 22, No. 1, 2018, pp. 47-53*

[2] *Moshe Koppel, Jonathan Schler, Dror Mughaz,* ***Text Categorization for Authorship Verification***

[3] *Fernanda López-Escobedo, Julián Solorzano-Soto, Gerardo Sierra Martínez (2016):* ***Analysis of Intertextual Distances Using Multidimensional Scaling in the Context of Authorship Attribution****, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2016.1142324*

[4] *François Dominic Laramée,* ***Introduction to stylometry with Python****, The Programming Historian 7 (2018), https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python.*

[5] *Neil Yager,* ***Authorship Attribution with Python****, http://www.aicbt.com/authorship-attribution/*

[6] *T. C. Mendenhall,* ***The Characteristic Curves of Composition****, Science, vol. 9, no. 214 (Mar. 11, 1887), pp. 237-249.*

[7] *Adam Kilgarriff,* ***Comparing Corpora****, International Journal of Corpus Linguistics, vol. 6, no. 1 (2001), pp. 97-133*

[8] *John Burrows,* ***Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship****, Literary and Linguistic Computing, vol. 17, no. 3 (2002), pp. 267-287.*