Objetivo
Determinar si hay un grupo de características que definen el estilo de un autor, con respecto al estilo de otros autores.

Hipótesis
Utilizando técnicas de clustering, podemos obtener un cluster de características que definen a un autor ‘x’ con respecto a otro grupo de autores ‘y’.

Metodología del experimento
Se tiene un grupo de características (trigramas de caracteres, trigramas de palabras funcionales, ‘func3g’, ‘lexsl’, ‘lexwl’, trigramas de partes de la oración, inicio de partes de la oración, fin de partes de la oración, ‘possh3, ‘possh1nvpa’, signos de puntuación, ‘vocab’). Cada característica contiene 25 renglones que corresponden a libros de los cuales conocemos su autoría. El objetivo es encontrar el grupo de características que definen a un autor ‘x’, por lo tanto, si el conjunto de características [trigramas de caracteres’, ‘lexsl’, ‘signos de puntuación’] definen a un autor ‘x’, cualquier libro de dicho autor ‘x’ debería tener ese grupo de características presentes con valores similares.

Para el sistema, se implementaron dos algoritmos de clustering sin supervisión, K-means y Agglomerative clustering. Hice uso de la implementación en la biblioteca sklearn de python. A ambos métodos los alimentamos con una concatenación de todas las características disponibles. En otras palabras, entrenamos el sistema con una matriz de dimensión 25xn donde n es la suma de las longitudes de los vectores de características.

Cómo es muy difícil determinar el número de características que determinan a un autor, implementamos un algoritmo iterativo, que agrupa las características en cantidades variables entre 3 y 12 (el máximo número de características), el algoritmo genera todas las combinaciones posibles

de características y las toma como candidatos a ser el grupo de características que definen al autor ‘x’. Cada grupo candidato es probado con los algoritmos de agrupamiento.

Para cada uno de los algoritmos de agrupamiento, se realizaron diferentes pruebas:

- K-MEANS: primero se hicieron las pruebas con 2 clusters (el primero perteneciente al autor ‘x’ y el segundo perteneciente al resto de autores). La segunda prueba fue realizada con 3 clusters (asumimos que alguno de los tres clusters contendrá el grupo de características que definen al autor ‘x’).
- Agglomerative Clustering: las pruebas con este algoritmo fueron hechas con 2 clusters (el primero perteneciente al autor ‘x’ y el segundo perteneciente al resto de autores). Primero se probó utilizando una afinidad euclidiana. La segunda prueba fue hecha con una afinidad manhattan.

Obtuvé en total 40 resultados, donde cada uno corresponde a un experimento. Por ejemplo, 1 resultado fue obtenido al haber usado el algoritmo K-MEANS con 2 clusters y con agrupaciones de 4 características.

Finalmente, hice una búsqueda en todos los resultados para encontrar un cluster que haya encontrado una agrupación de características que definen a un solo autor. No encontré ningún cluster que satisficiera el criterio.

Conclusiones

En conclusión, con el método propuesto y los parámetros probados, no fuí capaz de encontrar un resultado que cumpla con lo requerido. Sin embargo, hay otros algoritmos de agrupamiento que faltan por probar. Uno de los retos de esta tarea, es que al tener que generar combinaciones de forma iterativa, el algoritmo se vuelve muy complicado computacionalmente hablando, por lo que los experimentos consumen

mucho tiempo para completarse. En el futuro se podrían explorar formas para hacer obtener combinaciones candidatas de forma más eficiente sin tener que recurrir a cálculos combinacionales iterativos.