**M2.855 – Modelos avanzados de mineria de datos – aula 1**

**PEC 2 – Métodos No Supervisados**

Inicio: 31/10/22

Entrega: 20/11/22

Dedicación: 25 h

**Descripción del enunciado**

En esta actividad se aplican algoritmos de clustering y de reducción de dimensionalidad a conjuntos de datos no etiquetados. También conocido como aprendizaje no supervisado.

El documento "20221\_M2.855\_PEC2-Enunciado.ipynb" presenta el enunciado de la segunda actividad en formato ipython notebook, que contiene celdas específicas para la implementación de la solución y la respuesta a las preguntas planteadas en la PEC.

En la entrega debe figurar el Notebook (.ipynb) proporcionado y el HTML del notebook (se puede generar desde el propio notebook en el menú "File" > "Download as" > "HTML") tras ser ejecutado el notebook completo de forma secuencial comenzando por el principio. Por lo tanto, en la entrega deben figurar 2 ficheros.

También se adjunta el fichero environment\_uoc20221pec2.yml que contiene las dependencias del entorno para realizar la PEC.

Y el fichero kruskal.py que debe estar disponible en el mismo directorio que el notebook. Ya que contiene una definición de clase en la que se apoya la PEC.

20221\_M2.855\_PEC2-Enunciado.ipynb

environment\_uoc20221pec2.yml

kruskal.py

uber\_sample.csv

**Objetivos y competencias**

* Aplicar algoritmos no supervisados a conjuntos de datos utilizando librerías específicas de python.
* Conocer aplicaciones reales en las que se utilizan métodos no supervisados sobre conjuntos de datos no etiquetados.

**Continguts i recursos**

Introducción a scikit-learn: <https://www.oreilly.com/ideas/intro-to-scikit-learn>

Introducción a matplotlib: <https://matplotlib.org/tutorials/introductory/pyplot.html>

Múltiples figuras en matplotlib: <https://jakevdp.github.io/PythonDataScienceHandbook/04.08-multiple-subplots.html>

Clustering jerárquico en python: <https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/>

Introducción a t-SNE: <https://github.com/oreillymedia/t-SNE-tutorial>

Visualización animada para entender los hiperparámetros de t-SNE: <https://distill.pub/2016/misread-tsne/>

Tutorial de manejo de imágenes en numpy: <https://scikit-image.org/docs/dev/user_guide/numpy_images.html>

**Recursos d'aprenentatge**

Libro: Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. O'Reilly Media, Inc., 2019. ISBN 9781492032649

Cichosz, Pawel. Data Mining Algorithms: explained using R

Libro: Bonaccorso, Giuseppe. Hands-On Unsupervised Learning with Python. Packt Publishing, 2019. ISBN 9781789348279

Libro manual: Andreas C. Müller; Sarah Guido. Introduction to Machine Learning with Python. O'Reilly Media, Inc, 2016. 394 p. ISBN 9781449369415.

Llibre: Chollet, François. Deep Learning with Python. Manning Publications, 2017. ISBN: 9781617294433

Raschka, Sebastian. Python machine learning

¿Cómo se lee un libro digital de ProQuest Ebook Central?

Edwin de Jonge, Mark van der Loo. An introduction to data cleaning with R

Libro manual: Hackeling, Gavin. Mastering Machine Learning with scikit-learn. 2nd edition. Packt Publishing, 2014. 238 p. ISBN 9781783988372.

Libro manual: Hearty, John. Advanced Machine Learning with Python. Birmingham, GB: Packt Publishing, 2016. ISBN 9781784393830.

Espacio de recursos de ciencia de datos

Interfaz de usuario gráfica, Aplicación, Sitio web

Descripción generada automáticamente

Para los edges, usas el la funcion plot, no?

Si plot con las coordenadas

Hola Vanesa, síi, eso ya lo he hecho y ya tengo todos las parejas de nodos i los pesos. Yo lo que quiero saber es que pide en la pregunta siguiente que pide aplicar kruskal ejecutandolo variando el n° de ejes en funcion de nuestro criterio

Yo entiendo que los ejes son las uniones entre nodos de las que se han encontrado los pesos. Pero que se hace ahí?

La pregunta siguiente de visualizar el n°de clusters indicando la importancis variando el diametro o el color en funcion del n° de puntos, tampoco veo claro que hay que hacer.

Alguna idea para estos 2 apartados? Alguien tiene claro que hsy que hacer?

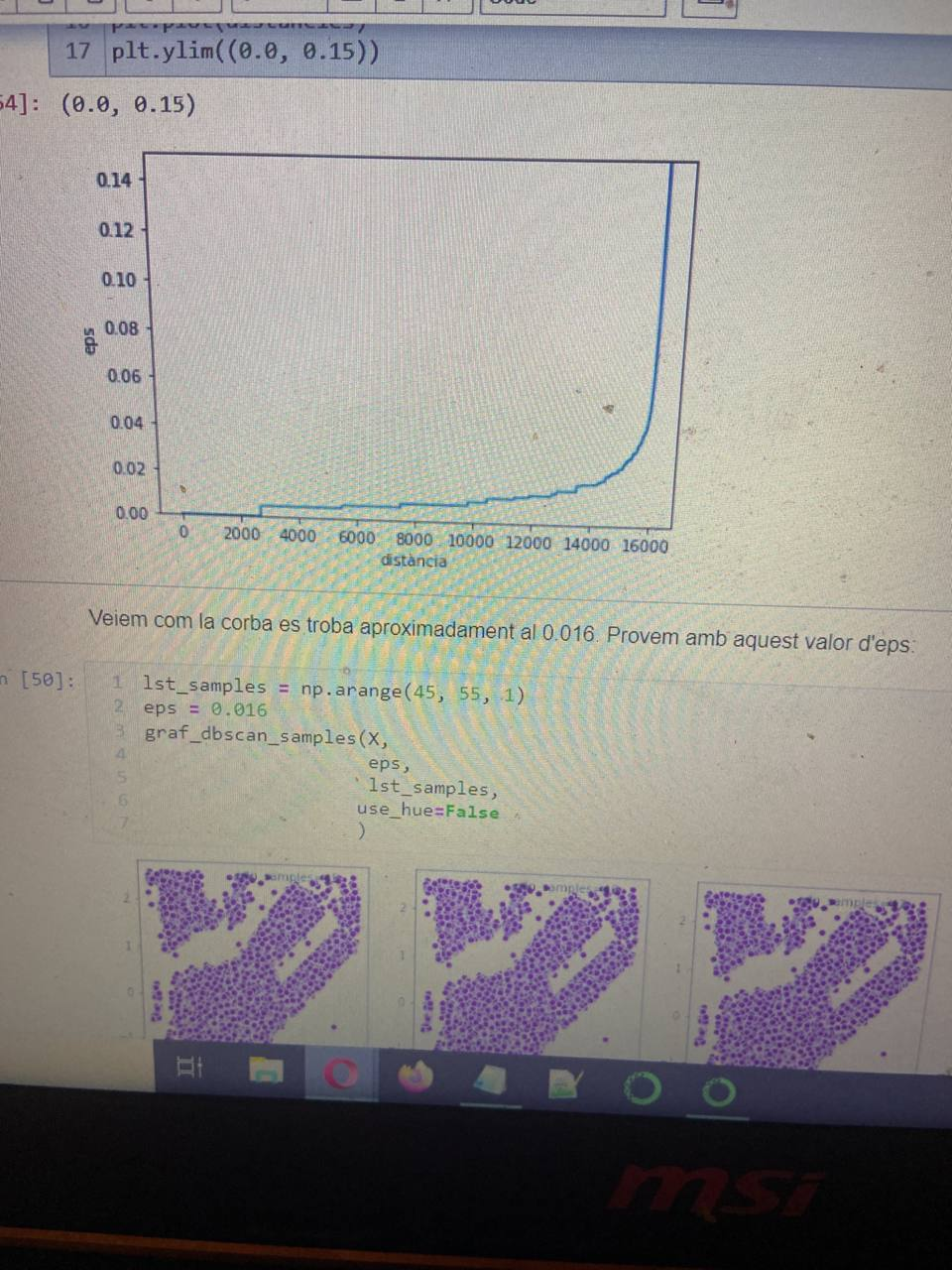
Yo con ejes entiendo lo mismo: la linea que une las estaciones

Donde El algoritmo de ejemplo dice g.add\_edge(

Perdón, quería decir que tienes que sustituir los g.add\_edge del algoritmo de ejemplo con tus resultados

Dentro del bucle cuando generas los nodos con el peso, en la última línea añades g.add\_edge(tus nodos, tus pesos) para que vaya añadiendo a g tus vértices

Y cuando ejecutes g.kruskal() la g contiene todos esos vértices que has calculado



Alguna idea para la última implementacion del 2c de la pac2?

La de visualizacion de los centros de los clusters indicando su importancia variando el diametro y el color en funcion del n° de puntos?

No veo por donde cogerla!

Alguien la ha hecho y puede dar alguna pista?

Cuando realices el scatter plot de cada centroide, en el tamaño le pones la cantidad de gente de ese punto

Es decir, el número de puntos de cada cluster que has calculado en el 2b

O sea. Se hace un scatter plot con la latitud i longitud del centroide de cada cluster, no?

Con el parametro 's'

Vanesa, s=nPuntos?

Si, s= puntos

Una cosa Vanesa, para hacer los enlaces mas o menos gordos...como lo haces en el scatter plot? Hay algun parametro adicional?

Linewidth

A la salida del kruskal, el valor que te da

El kruskal te da una columna con los pesos y el par de centroides. Ese peso pones en el linewidth de la línea entre esos centroides