Aprendizaje no supervisado: Clustering

1. Objetivos

En este práctico se explorarán diferentes soluciones de clustering, para desarrollar las capacidades de análisis de soluciones de clustering. Es preferible que los conjuntos de datos con los que trabajar sean propios, ya que de esta forma podrán aplicar su conocimiento del dominio en la interpretación de las diferentes soluciones. Alternativamente, pueden usar conjuntos de datos de los ejemplos de la materia.

En los mismos, hacer una breve discusión del problema y explicar cómo puede ser útil usar técnicas de *clustering*.

2. Consignas

Para cumplir los objetivos, realizar las siguientes actividades:

- Explorar soluciones con diferentes parámetros y compararlas. Por ejemplo, variar el número de clusters, las métricas de distancia, el número de iteraciones o el número de veces que se inicializan las semillas. Describir brevemente: número de clusters, población de cada cluster, algunas características distintivas de cada cluster, algunos elementos que se puedan encontrar en cada cluster.
- Incorporar un embedding como preproceso a los datos, aplicar los algoritmos de clustering después de ese preproceso y describir la solución o soluciones resultantes, discutiendo las ventajas que resultan. Se pueden usar:
  - Principal Component Analysis `http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`
  - para texto, embeddings neuronales Gensim `https://pypi.org/project/gensim/`
  - para texto, embeddings neuronales Fastext `https://pypi.org/project/fasttext/`
- Proponer (y en lo posible, implementar) métricas de evaluación de soluciones de clustering basadas en testigos. Los testigos son pares de objetos que un experto de dominio etiqueta como “deberían estar en el mismo cluster” o “deberían estar en distintos clusters”.
- El método k-means de scikit-learn no provee una forma sencilla de obtneer los objetos más cercanos al centroide de un cluster. Proponga alguna forma de obtener una muestra de los elementos de un cluster que sean cercanos al centroide, por ejemplo, usando clasificadores, usando distancia coseno, etc. En lo posible, implementarlos y mostrar esos elementos, discutir la representatividad de los elementos encontrados.

3. Entrega

Entregar un notebook documentado a través de slack. Idealmente, enviar una primera versión del notebook el viernes 17 de agosto para obtener feedback y poder hacer una iteración antes de la entrega final, que será el jueves 23 de agosto para poder discutir los resultados en la clase del 24 de agosto.