Después de ver regresión lineal y regresión logística como métodos de aprendizaje supervisado, pasamos al aprendizaje no supervisado. El método muy eficaz y comúnmente utilizado de este tipo se llama K promedios o en inglés K means. Lo que realiza este método es agrupamiento de datos (clustering en inglés). Se espera que datos en cada grupo (cluster) pertenezca a una categoría (por ejemplo, spam). Entonces este método de manera no supervisada realiza clasificación. No supervisado quiere decir que el método no ve las etiquetas de clases en los datos. En casos reales estas etiquetas no hay, pero para evaluar el método no supervisado se usa el conjunto etiquetado para ver si los datos en los clusters realmente pertenecen a una cierta clase, también se puede evaluar el error de clustering viendo cuantos datos en un cluster son de la clase esperada y cuantos pertenecen a otra clase o clases.

Por favor estudien de Lecture 13.1 a Lecture 13.5. Vamos a usar el mismo conjunto de spam-ham para programar el método de K means desde cero (NO USEN este método implementado en scikit learn) dándole a K el valor de 2, es que tenemos 2 clases: spam y ham. Vamos también evaluar este método, haciendo análisis de error: para cada de dos clusters hay que ver cuantos datos son de spam y cuantos de ham. Se espera que en el primer cluster la mayoría de datos serán de spam y en el segundo cluster, de ham.

Envíenme por favor el programa y la captura de pantalla con el análisis de error: por ejemplo, en la forma de esta tabla que ustedes van a llenar con números correspondientes:

	Número de mensajes spam	Número de mensajes ham
Cluster 1		
Cluster 2		