היום נממש את אלגוריתם k-means.

את האלגוריתם נריץ על iris data set שניתן לגשת אליו על ידי הפונקציה load_iris מספריית scikit-learn בלינק הבא:
http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris
בשדה של ה- classes של הפרחים אין צורך להשתמש, כיוון ש k-means הינו אלגוריתם unsupervised.

לשם כך נכתוב ארבעה פונקציות:

1. פונקציה שמקבלת כ- input
   a. וקטור של נקודות מה - data set
   b. ווקטור של ממוצעים (k ערכים כאלו)

   ומוצאת לכל אחד מהנקודות את הממוצע הקרוב אליו מבין האופציות ומחזירה dict שמכיל כמפתח מספר cluster וכערך list עם הנקודות שנמצאות בו.
2. פונקציה שמקבלת את ה-dict מסעיף 1 ומחשבת מחדש את וקטור הממוצעים של כל אחד מהקלסטרים, על ידי mean.
3. פונקציה שבודקת האם כבר הייתה התכנסות של האלגוריתם, לפי סף על גודל השינוי בערכים של הממוצעים. כלומר אם הממוצע לא משתנה הרבה, כנראה שהאלגוריתם התכנס.
4. פונקציה שקוראת לסעיפים 1-3 עבור המקרה של שלושה קלאסטרים על ה- dataset של iris ומחשבת את הקלאסטרים שיש ב- dataset

5. למתקדמים: אפשר לכתוב פונקציה נוספת שמקבלת כ-input את ה-dict מסעיף 1 ומשרטטת כל קלסטר בצבע אחר. רמז: ניתן לקחת את הצבעים מתוך list בעזרת index. פונקציה זאת נריץ במהלך הפיתוח, אחרי כל איטרציה.

הסברים מפורטים לפיתרון יש בלינק הבא:

https://datasciencelab.wordpress.com/2013/12/12/clustering-with-k-means-in-python/