מדעי הנתונים ובינה עסקית, סמסטר ב' תשע"ח

# תרגיל בית 4 – K-Means Clustering

---

**קראו בעיון את כל ההוראות לפני ביצוע העבודה**

---

**<u>הוראות כלליות:</u>**

א. אי עמידה בכל אחת מההוראות יגרור הורדת ציון או פסילת העבודה.

ב. הגשת העבודה ב**זוגות** בלבד .

ג. שפת תכנות – **Python 2.7**, סביבת פיתוח – מומלץ להשתמש ב- JetBrains PyCharm גרסה 2017.1 ומעלה. יש לוודא כי Anaconda מותקן.

ד. יש להגיש את העבודה **לתיקיית ההגשה הרלוונטית באתר הקורס (Moodle).** אחריותכם האישית לבדוק לפני הגשה כי כל הקבצים נפתחים כראוי.

ה. יש להגיש קובץ zip - שם הקובץ יהיה מורכב משני מספרי תעודות הזהות של המגישים באופן הבא: ID_ID.zip

הקובץ יכיל את הקבצים הבאים:

- הפרויקט המלא: קבצי קוד + GUI, חשוב : ללא קבצי הנתונים.
- קובץ readme.txt המכיל את שמות הסטודנטים ותעודות הזהות.
- קובץ PDF של הדוח המתאר את מבנה הפרויקט שיצרתם ותפקיד של כל מחלקה ושיטה בפרויקט.

ו. בנוסף, זוהי עבודה תכנותית ולפיכך יהיה משקל לכך בבדיקה. כלומר: יש לדאוג להערות בקוד, הסבר לפונקציות, חלוקה למחלקות, פונקציות קצרות ועניינים וכדומה.

**ז. תאריך ההגשה: 23:55 23.06.2018**

**הוראות התרגיל:**

בתרגיל זה עליכם להשתמש בספרייה Scikit-learn של Python על מנת לבצע clustering לקובץ נתונים. בנוסף, תטפלו ברשומות עם ערכים חסרים, כחלק מתהליך ניקוי הנתונים ותתרגלו עבודה עם מבני טבלה (Dataframe) שונים של Python. את פלט האלגוריתם תציגו באמצעות כלי הוויזואליזציה של הספרייה Plotly.

תיאור הקבצים שלרשותכם:

1. **Dataset general info** – מידע כללי בנוגע לבסיס הנתונים ממנו לקוחים נתוני התרגיל. קובץ זה הינו לשימושכם בלבד ולא ישמש כנתון שעל תכניתכם לקרוא במהלך הריצה.
2. **data** – קובץ האימון לאלגוריתם ה-clustering, בפורמט xlsx.

תיאור המשימות שעליכם לממש במסגרת התרגיל:

1. ממשק משתמש פשוט שיוצג עם הרצת התכנית. הממשק יכיל:
   1.1. הזנת ה-path לקובץ נתוני התרגיל (יש לממש אפשרות זו בעזרת browser). **על הממשק להכיל תיבת טקסט אחת בלבד, אליה יוכנס הנתיב לקובץ.** במידת הצורך, יש לשמור פלטים בתיקייה זו. **הטקסט אשר יופיע על הכפתור יהיה "Browse".**
   1.2. תיבת טקסט בה ניתן להזין את כמות ה-clusters אליהם יחולקו הנתונים. **שם תיבת הטקסט יהיה "Num of clusters k".**
   1.3. תיבת טקסט בה ניתן להזין את כמות הריצות של האלגוריתם מ-seeds רנדומליים שונים. **שם תיבת הטקסט יהיה "Num of runs".**
   1.4. לחצן לטעינת קובץ הנתונים, הכנתו וניקויו. **הטקסט אשר יופיע על הכפתור יהיה "Pre-process".**
   1.5. לחצן לבניית מודל ה-KMeans והצגת הוויזואליזציה של תוצאותיו. **הטקסט אשר יופיע על הכפתור יהיה "Cluster".**

2. הכנת הנתונים והכנתם (תהליך 1.4):

2.1. עם לחיצה על הלחצן המתאים, התכונית תקרא את קובץ הנתונים הנתון (בפורמט xlsx) ותטען אותו למבנה נתונים מסוג Dataframe.

2.2. לאחר קריאת הקובץ, יתבצע תהליך **ניקוי הנתונים**:

א. יש **להשלים** ערכים נומריים חסרים בערך הממוצע של כל ערכי התכונה.

ב. יש **לנרמל** את כל ערכי קובץ הנתונים לערך הסטנדרטי שלהם (חיסור הממוצע וחלוקה בסטיית התקן). פעולה זו מכונה Standardization.

ג. **קיבוץ הנתונים** לפי התוכנה "country". כחלק מתהליך זה, יש ליצור רשומה אחת עבור כל מדינה, כך **שתמצע** את ערכי התכונות על פני השנים (תכונה year).

2.3. בשלב זה יופיע dialog אשר יכיל את הודעה ***" Preprocessing completed successfully!"*** המודיעה על סיום הכנת הנתונים ויאפשר למשתמש ללחוץ על "OK" להמשך.

✓ ניתן להשתמש בספרייה הייעודית של Scikit-learn עבור ניקוי הנתונים.

3. חלוקת הנתונים לאשכולות (תהליך 1.5):

3.1. קובץ הנתונים ישמש לבניית מודל k-means באמצעות הספרייה הייעודית של Scikit-learn. קראו את התיעוד של המחלקה `sklearn.cluster.KMeans` באתר הרשמי של הספרייה: http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

3.2. יש להפעיל את האלגוריתם KMeans על סט האימון עם הפרמטרים הבאים:

א. מספר האשכולות (`n_clusters`) שהוכנס ע"י המשתמש בתיבת הטקסט המתאימה (תהליך 1.2 בממשק המשתמש).

ב. מספר הריצות (`n_init`) עם ערכי centroid רנדומליים שהוכנס ע"י המשתמש בתיבת הטקסט המתאימה (תהליך 1.3 בממשק המשתמש).

3.3. יש להצמיד את פלט האלגוריתם לכל רשומה (מדינה) בקובץ הנתונים.

3.4. יש ליצור שני פלטים המסכמים את תוצאות הריצה של האלגוריתם:

א. יש ליצור תרשים פיזור (scatter) של ערכי התכונה Generosity כתלות בערכי התכונה social_support. יש לצבוע את הנקודות על פי ערכי הפלט המתאים מהאלגוריתם.

❖ יש להקפיד להוסיף כותרות מתאימות לצירים ולתרשים כולו.

❖ יש להשתמש בפונקציה `scatter` של הספרייה `matplotlib`.

ב. יש ליצור horopleth map (מפת מדינות) המדגימה את פלט האלגוריתם (חלוקה של המדינות לאשכולות) עבור המדינות בקובץ הנתונים.

- יש להשתמש בספרייה Plotly ע"י התקנתה (בהנחה שפלטפורמת ה- Anaconda מותקנת על גבי python):
  - יש להתקין את הספרייה באמצעות הפקודה `pip install plotly` בחלון ה-Terminal.
  - יש להירשם (חד פעמי עם אי-מייל כלשהו) באתר https://plot.ly/accounts/login/?action=login
  - יש להיכנס למסך ה-Settings אחרי login לאתר Plotly וללחוץ על תת-תפריט ה-API Keys ואז על Regenerate Key כדי לקבל את ה-APIKEY.
  - שם המשתמש בתוספת המפתח יאפשר לכם לשמור את פלט מפת המדינות בחינם (עד 25 תרשימים) כקובץ על המחשב.
- יש להיעזר בדוגמא הנתונה בקישור הבא: https://plot.ly/python/choropleth-maps/.
- יש לשמור את תרשים המדינות כתמונה סטטית תוך שימוש בפונקציה הבאה:

```
import plotly.plotly as py
py.sign_in(username, API key)
py.image.save_as(choromap,filename='name.png')
```

3.5. יש להציג את שני הפלטים זה לצד זה במסך ה-GUI.

3.6. יש להציג dialog נוסף שיעדכן שתהליך ה-clustering הסתיים. לחיצה על "OK" תסיים את ריצת התכנית.

---

הטקסט על הכפתורים לא ניתן לשינוי, וחשוב שיהיה זהה למוגדר לעיל.
ניתן להניח שהפעולות יבוצעו בסדר הנכון – הכנת הנתונים ואז חלוקה לאשכולות.
כותרת **כל** החלונות (כולל הדיאלוגים שפורטו לעיל) צריכה להיות "K Means Clustering"

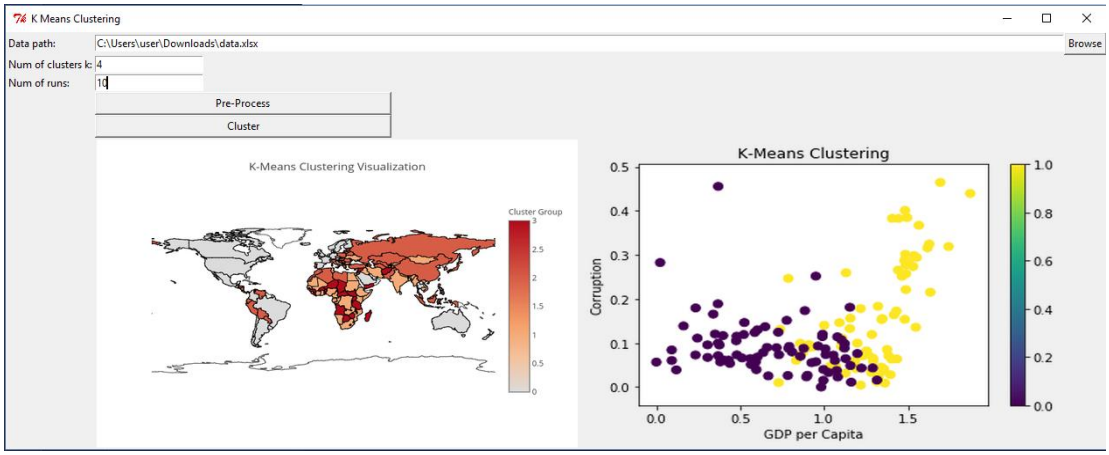
---

4. דוח:

יש לצרף דוח המתאר את המחלקות והשיטות שכתבתם בקוד.

<u>הערות חשובות נוספות</u>:

- מצורף תצלום חלון ה-GUI הנדרש לתרגיל זה. מומלץ להשתמש בממשק פשוט ביותר של Tkinter (from Tkinter import *). התמונה להמחשה בלבד. ממשק המשתמש לא חייב להראות כבתמונה, אך צריך להיות אינטואיטיבי, נוח ופשוט.




- ההנחה היחידה בעבודה היא כי סדר הפעולות הנדרשות יתבצע בסדר הנכון, כלומר קודם הכנת הנתונים ("Pre-process") ורק אז החלוקה לאשכולות ("Cluster").
- על התכנית לדעת להתמודד עם שגיאות כמו למשל קובץ נתונים ריק, מספר לא תקין בתיבות הטקסט השונות. במקרה של נתון לא תקין, יש להציג הודעת שגיאה מתאימה (המעידה על סוג השגיאה) ולא לאפשר לחיצה על כפתור ה-"Cluster".
- יש לבצע בדיקות קלט מלאות לכל השדות בממשק המשתמש ולכל הפרמטרים של האלגוריתם. היעזרו בתיעוד כדי לדעת מה טווח הערכים שמקבל כל פרמטר.
- הקוד ייבדק על קובץ נתונים דומה במבנה (אותם עמודות) אך עם ערכים שונים מהקובץ לדוגמא שניתן לכם.
- אינם נדרשים להתקין חבילות תוכנה נוספות חוץ מ-plotly (בהנחה ש-Anaconda מותקנת). <u>אין</u> להשתמש בחבילות שאינן קיימות בפלטפורמת ה-Anaconda וב-Plotly (כי הן מחייבות התקנה נפרדת).
- **תתבצע בדיקה לאיתור עבודות מועתקות (גם אם חלקית). הקפידו לא לשתף קטעי קוד!**
- שאלות בנוגע לתרגיל יש לשאול **<u>אך ורק</u>** בפורום השאלות הרלוונטי המופיע ב-moodle (ולא במייל - שאלות במייל לא יענו).

***בהצלחה!***