עקרונות וטכנולוגיות למדע הנתונים

יום לפני תאריך ההגנה יש להעלות למודל (בתיבה שנפתחה לכך) את מחברת הפרויקט.

שימו לב – בקורס זה הוגדר "פרויקט פתוח" – הכוונה שאתם מבצעים את הפרויקט על רעיון שלכם - כאשר אתם נדרשים לממש לפחות את רשימת הדרישות הבאה:

1. לא ניתן להשתמש בדאטאסט קיים מתוך Kaggle או דומיו. חובה להשיג את הנתונים באמצעות טכניקות סקרייפינג
2. הדאטה שצברתם צריך להכיל לפחות 10000 רשומות
3. עליכם לבצע clustering של דאטה מבוסס sequence-ים אשר sequence-ים שונים בו עשויים להיות בעלי אורך שונה. (למשל – דאטה של מסלולי נסיעה של אנשים שונים שדוגמים בכל דקה – יש אנשים שהמסלול שלהם קצר ויש כאלו עם מסלול ארוך – כשהמטרה באמצעות ה clustering לקבל תבניות של מסלולי נסיעה דומים. אפשר לבצע זאת גם על מסלולים של טיסות. דוגמא נוספת - דאטה של בקשות http בין services שונים באפליקציה – יש בקשות שמגיעות לשרת ה web ומשם נשלחות ל db ויש הודעות שנשלחות ל service-ים שונים שקוראים אחד לשני – מה שיוצא sequence-ים. קליסטור של הרצפים האלה שעשויים להיות בגדלים שונים יאפשר לראות את התבניות של הטרנזקציות העיקריות באפליקציה שדגמתם את ההודעות שלה. דוגמא נוספת – על הוראות של מתכונים כאשר sequence מוגדר כפעולות שצריך לבצע על המתכון– להרבה מתכונים יש הנחיות דומות, וקליסטור של הרצפים האלה יגלה תבניות להוראות להכנת מאכלים מסוגים שונים.)
4. עליכם להפעיל על הדאטה קליסטור ב 3 שיטות שונות לפחות – ולהשוות ביצועים עם מדדים (שעליכם להחליט עליהם לבד) בין השיטות השונות. אחת השיטות חייבת להיות בעלת אופי היררכי.
5. עליכם לבצע ניסויים המראים את היתרונות והחסרונות של השיטות שעבדתם עליהן ולגבות את ההחלטות בגרפים/אמצעים ויזואלים
6. עליכם לבחור אחד מהשניים – time series prediction או recommendation על דאטה שונה מהדאטה של בעיית ה clustering. גם כאן עליכם לעשות שימוש בלפחות 3 שיטות שונות, להסביר יתרונות וחסרונות לכל שיטה ולגבות זאת בניסויים וויזואליזציות בהתאם.
7. בכל אחד מהסעיפים הקודמים כאשר אתם מפעילים שיטות, עליכם לנסות כל אלגוריתם עם מגוון של ערכים ב hyperparameters (למשל עם grid search) וב 2 מהאלגוריתמים עליכם להשתמש ב Bayesian optimization על מנת לבחור את הפרמטרים האופטימאלים.

הציון יתייחס למספר מרכיבים:

א. מידת ההשקעה בפרויקט
ב. מידת עיבוד חומר הגלם בשיטות preprocessing
ג. מגוון השיטות שהשתמשתם בהן והיכולת לבחור פרמטרים אופטימאליים ונכונים
ד. ויזואליזצייה ו dashboard-ים שתציגו
ה. הפעלת עקרונות ולידציה נכונים עבור המודלים שהפעלתם ומדידת שגיאות
ו. שליטתכם בקוד של המחברת

בהצלחה,

שי ויוסי