התחלתי פרוייקט דאטה קטן

ראשית, הבנתי את הצורך - המטה לשמה אני עושה את הפרויקט. עניינה אותי השאלה "איזה קורסי בחירה בפקולטה שלי יניבו את התוצאות הכי כדאיות?" כלומר, לאיזה קורסים יש את הציונים הטובים ביותר? ככה סטודנטים צעירים יוכלו לבחור מגמת בחירה בצורה נתמכת-נתונים.

פירקתי את הקורסים ל-3 המגמות, כך כל סטודנט יוכל לבחור את ה-2 המתאימות לו בהתאם לממוצע שלהם.

הבעיה הראשונה שהייתה לי היא מקור המידע - מאיפה ניתן למצוא את הנתונים שלי?

למזלי, קיים אתר בשם Cheesefork שנוצר בידי מיכאל מלצב, שיצר מאגר של היסטוגרמות מסטודנטים שהסיכמו לסנכרן את המידע שלהם על המערכת והוא חילץ מתוכו נתונים כמו ממוצע ותוחלת.

לקח לי קצת זמן לעבור על הקוד שלו ב-Github, ולמרות שאני לא יודע JS או PHP אני הצלחתי להבין בעזרת קצת ChatGPT את הקוד, והתחלתי לכתוב קוד פייתון משלי שידוג מתוך הקבצים שלו את המידע שדרוש לי.

את הקוד ההתחלתי כתבתי באמצעות ChatGPT ואחרי כמה שינויים ווידוא שאני מבין אותו הצלחתי להריץ אותו ולקבל בפעם הראשונה את המידע שאני צריך לגבי הקורס הראשון!

החלטתי להסתמך בעיקר על החציון, משום שממוצע יהיה רגיש יותר לoutliers במקרים בהם יש מעט סטודנטים בקורס או מידע לא מייצג.

הגיע הזמן לעשות Scale-up לכל קורסי הבחירה. התחלתי עם מגמת אותות, שאת הקורסים שלה הוצאתי מקטלוג הלימודים העדכני ביותר וסידרתי את הרשימה בצורה ידנית. יצרתי קוד פייתון שיקח את הרשימה ויחלץ את הדאטה הרלוונטי מהמאגר של Cheesefork, וישים לי אותו בקובץ csv מסודר ויפה.

אחרי קצת שיפצורים לקוד זה עבד! היו 2 קורסים חדשים שהקוד לא מצא בספריות של cheesefork כי עדיין אין להם היסטוגרמות, אבל חוץ מזה - מושלם!

צריך עכשיו רק להשיג ממוצע של כל החציונים בקורס, ונוסיף גם progress bar קטן כדי שאדע מתי הקוד נתקע ו.... אוף- נהרס🤦‍ נגמרו לי ההרשאות לגשת לתקיות ב Github וצריך להבין איך לתקן את זה

אחרי הפסקה קטנה החלטתי פשוט להוריד את כל התקיות מgit למחשב שלי, שזו גם דרך

אבל לא! החלטתי שזה חייב להיות פרויקט web-scraping וגיליתי איך לפתור את הבעיה בעזרת access token שיוצרים בגיטהאב.

קצת דיבוג לקורסים בינאריים, ולהוספת פונקציה שתחשב את הממוצע של כל ציוני החציון של הקורס, במקום לחשב את הכל באקסל בצורה נפרדת לכל רשימה, ועוד קצת פונקציות שיעשו את העבודה יותר אוטומטית.

**מבט לעתיד**

יש עוד מקומות לשיפור הפרויקט בעתיד, למשל הוצאת המידע על הקורסים בעזרת OCR לקריאת ה-PDF ישירות ולעשות את התהליך אוטומטי יותר. במצב כזה לא אצטרך להוציא את נקודות הזכות של הקורסים בצורה ידנית.

אפשר להוציא קורסים ספציפיים למחזור מסויים ולחשב את הציונים המייצגים של מחזור ספציפי (לא לפי שנה, אלא לפי מחזור)

ובהמשך יש לי כוונה לכתוב קוד שישתמש במידע מהפרויקט הזה כדי לחשב מה הדרישות שסטודנט צריך למלא עד סוף התואר שלו, בתנאי שהוא מזין את גליון הציונים העדכני שלו (למשל, סטודנט בשנה ב' מכניס טופס ויוצאת לו רשימה של קורסי הבחירה, ספורט/מל"ג, חובה שהוא צריך למלא וכמה נק"ז ומאיזה מגמה או שרשרת)