תיעוד פרוייקט בית המשפט העליון

קישורים

בגט, להעתקה של אתר בית המשפט העליון: https://github.com/andyil/baguette
נמלה, או הנמלה, אפליקציית וב מבוססת ג'נגו לקידוד תיקים: https://github.com/andyil/nemala
בסיס הנתונים האחרון של הנמלה, מסוג sqlite3 ניתן להורדה פה:
https://my-versioned-bucket324.s3.eu-central-1.amazonaws.com/coding-backup/db.sqlite3

תיאור כללי של הפרוייקט

החלקים העיקריים של הפרויקט

העתקת נתונים מאתר בית המשפט העליון למחשב

הפרויקט הזה חי בתוך הפרויקט בגט בתוך תיקיית scraper.
החלק הזה הוא סקריפט בפייתון שמוריד מבית המשפט העליון לדיסק את כל המידע בטווח תאריכים. למשל מ-1/1/2010 עד 31/12/2019.
הטווח מחולק לימים, וכל יום מורד בחוט (Thread) נפרד, על מנת לזרז את ההליך זה יורד למיטב זכרוני בכ-20 חוטים במקביל.
עבור כל יום מתקבלת רשימת תיקים, ועבור כל תיק מבצעים שאילתה של כל המסמכים שיש. בנוסף באחד מהשלבים מתקבלת מטה דטה עבור התיק.
לשם פשטות הכל מועתק לדיסק לקבצים ולא לבסיס נתונים.
כל הקריאות לשרת של בית המשפט העליון מבוצעות באמצעות ספריית requests.
מומלץ לבלות כמה זמן בגלישה באתר בית המשפט העליון, לבצע שאילתות שונות לצפות כיצד מתורגמות לבקשות http באמצעות חלונית המפתחים בדפדפן כרום.

גיבוש הנתונים לכדי קובץ שורות

פשוט ונח להוריד את הנתונים לקבצים רבים אבל מבחינת ביצועים זה בעייתי. לכן יש חלק שקוראים לו parser שמגבש את כל הנתונים, כולל ניתוח מסויים מתוך מסמכי החלטות html ובונה קבצי csv.
נעשה שימוש בספריית BeautifulSoup לפיענוח קבצי html.

הנמלה

הנמלה היא אפליקציית ג'נגו שמאפשרת למספר אנשים לקודד תיקים ידנית. קטגוריות הקידוד והערכים האפשריים קבועים מראש, כמו שמות השופטים (אם יש שופטים חדשים יש להוסיף אותם).
בקישורית לעיל רשמתי כתובת להורדת בסיס הנתונים מסוג sqlite3. מדובר בבסיס נתונים קטן ופשוט שלא דורש התקנה ורץ בתוך ג'נגו.

כל מצב העניינים (כלומר ה-state) בנמלה שמור בקובץ של בסיס הנתונים.
על מנת שמספר של מקודדים יוכלו לקודד מכל מקום יש צורך למצוא לזה שרת (הדבר הכי קטן וזול שאפשר).

מיזוג קידוד הנמלה עם הנתונים שהורדו

זה נמצא בספריית merger בבגט.

חלקים קלים וחלקים מאתגרים

החלקים הפשוטים וקלים יחסית הם: העתקת נתונים מבית המשפט העליון, והנמלה.
החלקים הקשים והמאתגרים הם: ניתוח מסמכי החלטות ב-html. אפילו דברים יחסית טריויאליים כמו לחלץ שמות של שופטים הם מסובכים כי הפורמט הוא לא אחיד ויש כמות גדולה מאוד של פינות ומקרי יוצאי דופן.
עוד חלק מאוד קשה ומאתגר זה מיזוג התיקים של הנמלה עם הנתונים שגורדו אוטומטית. יותר מדי מקרי קצה.

המשך שאלות ותיעוד