הצגת פרוייקט במדעי הרוח הדיגיטליים

"סטטיסטיקת שילובי מקצועות מאז ועד היום"

שחר אזולאי

<u>מטרת הפרוייקט:</u>

יצירת כלי אינטראקטיבי לסינון והצגה נוחה של מידע מספרי אודות שילובי מקצועות לפי חתך זמן בהתבסס על מאגר Wikidata.

<u>רקע:</u>

לפני כעשור יצא לי להיחשף לחיים שפירא, ד"ר למתמטיקה, חוקר ומרצה בנושא תורת המשחקים שגם חיבר רב מכר בתחום זה("שיחות על תורת המשחקים") ובכך הצליח לסחוף רבים להתעניין בתחום. הרצאותיו בנושא זכו אף הן לאהדה כולל הרצאה שנתן במסגרת TED ושפירא נחשב לשם דבר בענף מתמטי זה.
לפני כשנתיים זכיתי לקבל הזמנה פתוחה להרצאה של הדוקטור וכמובן שלא נדרשו שכנועים נוספים. רק שכותרת ההרצאה מאוד לא התאימה למה שציפיתי: "משמעות החיים על פי קהלת" היה שמה. מה למתמטיקאי שמגדיר עצמו כחילוני ולספר קהלת? התשובה היא שהקשר של שפירא לספר קהלת לא נובע מהיותו מתמטיקאי וגם לא מאמונתו הדתית, שפירא הוא גם פילוסוף והוא דואג לציין זאת גם בהרצאתו.
הופתעתי מהאמירה של הדוקטור למתמטיקה לגבי היותו פילוסוף. "מה הקשר?" חשבתי, אך במחשבה שנייה נזכרתי בכל אותם מתמטיקאים גדולים שלמדתי אודות עבודתם(אפילו בתיכון) וגם הוגדרו גם כפילוסופים. פיתגורס(המאה ה6 לפנה"ס), תלמי(המאה ה2), תאלס(המאה ה7), הרמב"ם(המאה ה13) ורנה דקארט (המאה ה17) הם רק חלק מהשמות שהופכים את השילוב הזה ליותר טבעי לאוזן ורובם אם לא כולם אף הצליחו להגיע לתובונות בתחוממאחד בזכות עבודתם בתחום השני, כך שאפשר לומר שלמתמטיקה ולפילוסופיה יש נקודת השקה ואף כמה.
אך בכל זאת משהו גרם לי להסתכל על שילוב המקצועות כצירוף לא טריויאלי במחשבה ראשונה וניסיתי להבין מהו אותו הדבר. מעט מחשבה על זה הביאה אותי לשתי סברות, הראשונה היא שהשילוב הזה פשוט נשחק לאורך ההיסטוריה והשנייה יותר כוללנית והיא שפילוסופיים בני ימינו כבר לא מקבל את אור הזרקורים.

השאלה שהניעה אותי לפרוייקט הזה היא - האם אכן ניתן לאמוד ירידה בקבוצת החיתוך שבין המקצועות לאורך השנים?
על השאלה הזו אני אוכל לענות בתוך כמה דקות של עבודה על מאגר Wikidata אך תהיה לי תשובה על שאלה מאוד ספציפית ונראה שניתן להכליל את השאלה כך שיתקבל כלי שיעזור לענות על שאלות כדוגמתה ועל שאלות נוספות אודות סטטיסטיקת תחומי העיסוק לאורך ההיסטורייה האנושית. מציאת חתכים בין מקצועות נוספים, מציאת 2 התחומים שמשולבים הכי הרבה וויזואליזציה של היכחדות מקצועות כמו גם בריאתם הם חלק מהדברים שהכלי יאפשר לענות עליהם בלחיצת כפתור כמו גם על השאלה "לאן הפילוסופים הולכים כשהאגם קפוא?"

תכנית העבודה:

1. איסוף כלל הנתונים ממאגר Wikidata. מדובר בקובץ השוקל כ-107GB.
2. חילוץ המידע הרלוונטי בלבד(קרי, כל העמודים הקשורים במקצועות/אישים)
3. ניקיון וקישור הנתונים - מבדיקה מדגמית של הדאטה התברר כי הערכים המתקבלים עבור שדה "תחום עיסוק"(occupation) אינם מתאימים מסיבות שונות:
   - מה שנקרא באנגלית occupation (תחום עיסוק) לאו דווקא מתייחס למקצוע או משהו שהייתי רוצה להחשיב כמקצוע. לצורך העניין ניתן למצוא כתחום עיסוק ערכים כמו סטודנט, נוסע(passenger), שורד וציוני ואף ערכים פחות ידידותיים כמו טרוריסט, אנס ורוצח.
   - את חלק מתחומי העיסוק המצויינים במאגר הייתי רוצה להציג כתחום עיסוק כוללני יותר. לדוגמה: amateur mathematician ,applied mathematician ו-mathematician כולם יחשבו מתמטיקאיים.
   - חלק מהמקצועות הם כל כך נישתיים ועל כן אין מספיק מקורות אודות אנשים שעסקו בהם. לדוגמה:
     Büdner - Historical class of small landowners in northern Germany
     בקטגוריה זו קיים ערך עבור אדם אחד בשם Friedrich Schuchardt וגם על כך אין מידע האם השתכר מהיותו בעלים של חלקה כזו או אחרת.
4. בניית מבנה נתונים יעיל בשליפת המידע אותו הייתי רוצה להנגיש בעזרת הכלי.
5. יצירת API נוח בצורת פונקציות לתשאול מאגר הנתונים שבניתי.
6. בניית UI לכלי והוספת ויזואליזציות מתאימות למידע שחוזר מהשאילתות השונות.
7. שיעל תביא לי 100.

היעד:

בחזון המינימלי הייתי רוצה שהמשתמש יוכל לבצע את הדברים הבאים:

1. לבחור באמצעות ממשק גרפי שילוב כלשהו של מקצועות מתוך רשימה(גם מקצוע אחד) ולראות גרף המציג את אחוז האנשים (מתוך כלל האוכלוסייה המתועדת) שעסקו בשילוב זה כפונקציה של זמן בשנים.
2. בהינתן אינטרוול של שנים(לדוגמה: 1850-1948) לראות גרף קודקודים המציג את היחסים בין המקצועות השונים. כלומר, קודקודי הגרף יהיו המקצועות באותן שנים וגודל הקודקוד יקבע על פי כמות האנשים שתועדו כעוסקים במקצוע זה. 2 קודקודים יחוברו בקשת במידה והיה אדם אשר עסק באותן שנים בשילוב מקצועות זה. כמו כן גם עובי הקשת יקבע על פי מספר האנשים אשר עונים לתיאור זה.
3. אין לי עוד רעיון איך להציג זאת באופן גרפי אך אשמח להוסיף אופציה שתשלב מידע אודות המקום בעולם בו פעלו אותם עוסקים בשילובי מקצועות שונים.

מקורות ביבליוגרפיים:

מקורות מידע:

1. מאגר Wikidata יהווה את הבסיס לכלל המידע בנושא.

כלים וספריות:

1. אנסה לעבוד עם OpenRefine בתקווה שהוא יתאים לעבודת הניקיון שמצפה לי.
2. ספריית qwikidata לפייתון המאפשרת תשאול sparql של מאגר Wikidata ובנוסף עבודה על קובץ המאגר(107GB) לאחר שיהיה שמור אצלי באופן לוקאלי.
3. ספריות פייתון השונות ובעיקרן pandas.
4. ספריות וכלים שונים לצורכי UI וויזואליזציה. יש לי מספר רפרנסים לכלים כאלו ואחרים(כולל אלו ששלחת לי) אך אני טרם יודע להגיד מה יתאים לי באופן מיטבי.

כמה מילות סיכום:

מאוד מעניין אותי להתחיל את הפרוייקט שכן הוא הגיע מתוך רעיון שהעליתי בעצמי ובעזרתו אצליח לקבל אישוש או הפרכה לאותן סברות. כמו כן אני חושב שהוא יניב מספר ממצאים מעניינים שיכולים להוות בסיס למחקרים עתידיים בנושא. בנוסף הוא הולך לכלול בתוכו כמה אספקטים טכנולוגיים שמעניינים אותי החל מעבודה עם דאטה גולמי ביותר ועד לויזאליזציה של מידע נקי וUI.
כרגע אני צופה שניקוי וסידור הדאטה הולך להיות הקושי כמו גם גזלן הזמן הגדול ביותר אך זהו גם השלב החשוב ביותר שכן הוא מזקק את כל המידע הנחוץ לשם הסקת תובנות. במילים אחרות, המאגר הנקי והמסודר ללא ממשק גרפי הוא מספיק לצורך הסקת תובנות משמעותיות אך לא להפך.

בעמוד הגיטהאב ניתן למצוא קובץ מחברת פייתון(.ipynb) המציג אינטרקציה קלה עם הספרייה שתשמש אותי בדליית המידע והצגה גרפית של החיתוך בין כל המתמטיקאים והפילוסופים המתועדים. מעניין יהיה לראות איך הנתונים האלו נראים על גבי ציר הזמן.