‏8 דצמבר, 2024

**פגישת פרויקט גמר - קוד המס בארה"ב**

משתתפים: פרופ' דורון נרוצקי, ד"ר תמיר שאנן, ד"ר אור אנידג'ר, אורי בראל, תומר סורוז'ון ויפעת גסטפריינד

* חידוד המטרה הכללית של הפרויקט – השוואת קוד המס בארה"ב לפני ואחרי הרפורמה משנת 2017 (קרי, יש להתייחס לקוד המס של שנת 2017 וכן, לקוד המס של 2018, בהתאם לקישור ששלחו תמיר ודורון)
* שלב 1: Pre-processing לרידוד ולפירוק הקוד ל- topics / ראשי פרקים (בכל אחת מהשנים), קרי לרדד את המילים מ- 10 מיליון לסדר גודל סביר יותר (למשל, כאלף)

ניתן לאתר כותרות / פרקים ראשיים ע"י מודל שפה כגון GPT.

* שלב 2: Encoding לטקסט – שימוש במודל [Agnostic BERT](https://drive.google.com/drive/folders/1AmWfiCUhSGgjB_v2Bxcldu466Q2RG8o7?usp=sharing), לרבות:
  + איתור המקומות שבהם יש מקסימום דמיון (לירי וסמנטי) בין הקוד לפני/אחרי הרפורמה, קרי באיזה ראשי פרקים הקוד "מועתק" לפחות ב- 90% ביחס לקוד לפני הרפורמה
  + תשומת לב לדמיון/השוואה בין הקוד לפני/אחרי הרפורמה, בכל החיתוכים, קרי איפה יש דמיון ברמת פרקים זהים, איפה יש פרקים שהופיעו בקוד לפני הרפורמה ולא קיימים בקוד אחרי (פרקים שהוסרו) ואיפה יש פרקים שקיימים בקוד אחרי הרפורמה ולא לפני (פרקים שנוספו).
  + במקומות שבהם קיים דמיון ברמת ראשי הפרקים, יש לבחון את הדמיון בין הפרקים.
  + חישוב הסף (Threshold) שמתקבל ברמת clusters – לצורך השוואה ראשונית, נקבע ערכים כדלקמן:

Threshold עבור דמיון לירי – 90% חפיפה בין המילים

Threshold עבור דמיון סמנטי – 80% דמיון

* + התוצר צריך להיות מעין דו"ח של מה שאותר ע"ב האמור לעיל, ברמת ראשי הפרקים (למשל בפרק דיני תאגידים היתה התאמה/דמיון של 80% ברמה הלירית וכיו"ב).
  + הצגת התוצר בעוד שבועיים, במפגש הזום הבא
* שלב 3 (מאוחר יותר): שימוש במודלי שפה נוספים, כגון: LLaMA 3.2 , Mistral AI,, Nova (Amazon)
* הערה: ככל שעושים שימוש ב chat GPT, יש למלא הצהרת שימוש ב- AI