‏22 דצמבר, 2024

**פגישה 2 - פרויקט גמר - קוד המס בארה"ב**

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

* הוצגו תוצאות ראשוניות של הניתוח שבוצע, בהתאם לפגישה שהתקיימה בתאריך 8.12.24 - תוצאות ניתוח סמנטי וניתוח לירי של הכותרות וראשי הפרקים של קודי המס משנים 2017 ו- 2018, בחיתוכים שונים.
* מהתוצאות ניתן ללמוד, כי במרבית הכותרות / ראשי הפרקים (למעלה מ- 80%), קיים דמיון סמנטי ולירי בין השנים 2017 ו-2018.
* תמיר ודורון ציינו, כי מרבית שינויי החקיקה מקורם בתוכן/במהות ולא בכותרות וראשי הפרקים. יחד עם זאת, ניתן לעשות שימוש בניתוח הכותרות ע"מ לבסס את הטענה כי על אף שנשמר המבנה הכללי, התוכן עצמו שונה מהותית (ניתוח הכותרות וראשי הפרקים יסייעו בביסוס הרישא של הטענה). דורון ציין, כי להערכתו מרבית השינויים חלו בתת-פרק N (Tax based on income from sources within or without the United States).
* במהלך בחינת התוצאות, עלה צורך בטיוב נוסף של הכותרות וראשי הפרקים (כגון: הסרת מספרי כותרות וכד') וכן, הצורך בשליפת מידע נוסף שיענה למכלול הדרישות לניתוח מורכבות הטקסט, כפי שעלו מהמסמך הראשוני (כגון: מבנה היררכי, מספר ההסתעפויות בכל פרק, אורך התיקונים, מספר מילים וכד'). כמו כן, בהמשך, ובהתאם לתוצאות ניתוח נוספות, נדרש יהיה לכייל את המדדים והספים (Threshold) לפיהם נקבע הדמיון הסמנטי/לירי.
* המורכבות של הטקסט תכלול התייחסות למספר פרמטרים, כאמור לעיל, בין אם במדד אינטגרטיבי ובין אם במספר תוצאות/מדדים נפרדים (מאחר שהתלות בין המדדים אינה ברורה דיה).
* דורון ציין, כי במקביל לבניית וכתיבת הקוד / האלגוריתם לניתוח קודי המס, יש צורך לכתוב ולתעד הסבר תמציתי, נהיר וברור לקוראי המאמר, שאינו טכני, באשר לקוד / האלגוריתם ולחיתוכים המבוצעים לצורכי הניתוח. הנ"ל יצורף כחלק בלתי נפרד מהמאמר.
* המשך עבודה וניתוח:
  + ניתוח השוואתי לבחינת דמיון (סמנטי ולירי) בין יחידות מידע קטנות יותר, במדגם מקרים שבהם הכותרות נמצאו דומות (קרי, עבור כותרות שנמצא דמיון ביניהן, יש "לפרקן" ולנתח את התוכן הפנימי ברזולוציה נוספת, ולהשוות פסקה מול פסקה, כמכפלה קרטזית).
  + בחינת מדדים סטטיסטיים של מילים/ביטויים מסוימים (כגון: חציון, ממוצע וכד').
  + חישוב שכיחויות – השוואת המילים/ביטויים השכיחים, בין הקוד לפני/אחרי הרפורמה (למשל, באמצעות שימוש ב- N-Grams).
  + בחינת פרמטרים נוספים בין פרקים/כותרות דומים – כמות מילים, כמות ראשי פרקים / תתי פרקים, כמות פסקאות לכותרות דומות – לצורך הצגת שוני מבני בין הקוד לפני/אחרי הרפורמה.