בס"ד

**שם הפרויקט: מערכת לניתוח מורכבות וקריאות טקסטים וחוקים: יצירת מדד חדשני ואימון על קוד המס האמריקאי**

שמות הסטודנטים: (ת.ז. אימייל וטלפון):

|  |  |  |  |
| --- | --- | --- | --- |
| ת.ז | שם מלא | אמייל | טלפון |
| 043139419 | יפעת גסטפריינד | [yifatb18@gmail.com](mailto:yifatb18@gmail.com) | 052-3124611 |
| 322355264 | רויטל יוסופוב | [revital2409@gmail.com](mailto:revital2409@gmail.com) | 054-6727722 |
| 318161767 | תומר סורוז'ון | [tomer.surujon@gmail.com](mailto:tomer.surujon@gmail.com) | 052-2292110 |
| 207813007 | אורי בראל | [oribarel80@gmail.com](mailto:oribarel80@gmail.com) | 054-7981113 |
| 205884927 | גיא פרבשטיין | [guymkc7@gmail.com](mailto:guymkc7@gmail.com) | 050-6310001 |

התמחות:

תואר מדע נתונים

מאשר הפרויקט:

ד"ר אור חיים אנידג'ר

מטרת הפרויקט:

מטרה העל:

פיתוח מערכת שתשמש לניתוח מורכבות וקריאות של טקסטים, בדגש על חוקים ומסמכים משפטיים. המערכת תשלב מדד כמותי חדשני להערכת המורכבות והקריאות, ותציע כלים ניתוחיים להשוואת טקסטים שונים.

כחלק מאימון המערכת, יתבצע פיילוט על קוד המס האמריקאי לפני ואחרי חקיקת חוק TCJA משנת 2017, חלק ממחקר שנערך בנושא זה ע"י פרופ' דורון נרוצקי מאוניברסיטת אקרון בארה"ב וד"ר תמיר שאנן מהמכללה למינהל. קרי, הפרויקט ינתח את השפעת חקיקת חוק TCJA על מבנה, קריאות ומורכבות קוד המס האמריקאי.

בעזרת טכניקות עיבוד שפה טבעית (NLP) וניתוחים כמותיים, נבצע השוואה בין קוד המס שלפני החקיקה ולאחריה ונציג תובנות שימושיות, באופן שיסייעו להשגת מטרות המחקר וביניהן - זיהוי מגמות ותובנות מרכזיות בנוגע לאופן שבו ה-TCJA השפיע על מבנה הקוד והיכולת להבינו.

מטרות הפרויקט:

1. יצירת מדד כמותי המשלב מדדים כמו מספר מונחים מוגדרים, תנאים מרובדים וקריאות סמנטית.
2. השוואת מורכבות וקורלציה בין טקסטים באמצעות טכניקות מתקדמות של עיבוד שפה טבעית.
3. פיתוח מערכת אינטראקטיבית להצגת התוצאות בצורה ידידותית, כולל גרפים, דוחות והשוואות.

רקע:

חקיקה ומשפטים כוללים טקסטים מורכבים הדורשים ניתוח מעמיק כדי להבין את השפעתם על משתמשי הקצה, כמו אזרחים, רגולטורים, יועצים משפטיים ועוד. קוד המס האמריקאי מייצג דוגמה בולטת למורכבות כזו, והשינויים שהוכנסו במסגרת TCJA של 2017 מציבים הזדמנות ייחודית להעריך את ההשפעה של שינויים בחקיקה על המורכבות והקריאות. קוד המס האמריקאי ידוע במורכבותו הגבוהה, וה-TCJA מ-2017 הציג שינויים משמעותיים. ניתוחים מתקדמים של מבנה ומורכבות קוד המס יכולים לסייע ביצירת חוקים נגישים יותר בעתיד.

מה קיים היום:

מערכות קיימות:

* מאגרי מידע, כמו מערכות חיפוש, מערכות משפטיות, חוקים ומסמכים משפטיים.
* כלים בסיסיים לניתוח קריאות והפניות (בתוך טקסט).

חסרונות:

* היעדר מדד כמותי מקובל למדידת וניתוח רמת מורכבות וקריאות טקסט.
* תלות גבוהה במילות מפת וכלים מוגבלים המסתמכים על מילות מפתח בלבד.
* חוסר ניתוח איכותי של הקשר בין טקסטים.
* חוסר ביכולת לנתח את ההשפעה של שינויים מבניים.

תיאור הפרויקט:

תיאור הבעיה:

חקיקה חדשה משנה לעיתים קרובות את מבנה החוק הקודם, וכתוצאה מכך את מורכבותו. אין כיום כלי המאפשר השוואה כמותית ואיכותית בין גרסאות שונות של חוקים וטקסטים רגולטוריים. יתרה מכך, אין כיום מדד ברור ומקובל המאפשר למדוד את מורכבותם של חוקים וטקסטים רגולטוריים.

הפתרון המוצע:

מערכת לניתוח והשוואת המורכבות של חוקים וטקסטים רגולטוריים (ובפרט של קוד המס לפני ואחרי TCJA), לרבות:

1. פיתוח מדד כמותי לרמת מורכבות וקריאות טקסט שיכלול ניתוח משתנים, כגון: מספר תנאים, הפניות צולבות, ספירת מילים והשוואת אורך טקסט, קריאות טקסטואלית, ותנאים מרובדים.
2. השוואת טקסטים על בסיס ניתוח סמנטי (למשל, דמיון בין סעיפים).
3. ניתוח קריאות באמצעות מודלים מתקדמים (כמו Flesch-Kincaid).
4. פיתוח מערכת המאפשרת חישוב מדדי מורכבות והשוואת טקסטים.

דוגמא לפונקציונליות המערכת:

1. קלט: שני טקסטים / מסמכים רגולטוריים להשוואה וניתוח.
2. חישוב וניתוח:

* חישוב מדד מורכבות וקריאות עבור כל טקסט.
* ניתוח סיבות שורש לרמת המורכבות, עבור כל מהטקסטים/קודי המס והשוואה בין הסיבות לכל טקסט.
* סיכום והצגת תובנות על בסיס ניתוח סמנטי של שני הטקסטים

1. הפקת דו"ח/פלט: דו"ח הכולל את כל פרט המידע (החישובי, מסקנות, סיבות שורש וכיו"ב), ניתוחים תובנות גרפיות.

**שלבי הפרויקט:**

1. **איסוף והוצאת הדאטה (Data Extraction):**

* שימוש בגרסאות ציבוריות של קוד המס לפני ואחרי TCJA.
* שימוש במקורות פתוחים של חוקים וטקסטים.
* שימוש בכלים לחילוץ נתונים ממסמכים טקסטואליים (PDF, Word).

1. **ניתוח ותחקור הדאטה (Data Exploration):**

* ניתוח מבני של סעיפים ותת-סעיפים
* זיהוי קשרים בין הפניות צולבות
* גילוי אנומליות, אי-התאמות, ומידע חסר

1. **בדיקת ובחירת מודלים רלוונטיים:**

בחירת מודלים - בחינה של המודלים הבאים והחלטה אילו מתאימים ורלוונטיים יותר:

* BERT: מודלים המותאמים לעיבוד שפה רגולטורית, המציעים תובנות עמוקות יותר על טקסטים ארוכים ומורכבים.
* GPT: לביצוע סיכומים אוטומטיים, יצירת תשובות לשאלות או כתיבת הצעות לשיפור ניסוחים
* TF-IDF: למדידת וחישוב שכיחות מילות מפתח.

תהליך בדיקת המודלים:

* אימון ראשוני על סט נתונים קטן כדי להעריך ביצועים (ראשי פרקים וסעיפים).
* מדידת דיוק, שלמות, וזמני ביצוע עבור כל מודל.
* בחירת המודל המתאים ביותר.

1. **פיתוח מערכת אינטראקטיבית:**

קלט:

* פיתוח ממשק משתמש ידידותי להוספת קבצי טקסט להשוואה.

פלט:

* הצגת תוצאות מדד מורכבות וקריאות עבור כל טקסט
* הצגת תוצאות ההשוואה וסיבות השורש, לרבות באמצעות ויזואליזציה של תוצאות הניתוח

1. **הערכת ביצועים (Testing & Evaluation):**

* בדיקות איכות ודיוק למודל
* משוב משתמשים (עורכי המחקר בנושא קוד המס האמריקאי), לגבי תוצאות המודל

טכנולוגיות וסביבת עבודה:

פיתוח המערכת יתבצע תוך שימוש בטכנולוגיות מתקדמות, מתודולוגיות וכלים מעולם עיבוד השפה הטבעית (NLP), פיתוח יישומים מבוססי ענן, ובניית ממשקים ידידותיים למשתמש.

1. שפות תכנות: Python, JavaScript
2. פלטפורמות שיתופיות: Jira או Monday לניהול משימות ופרויקטים
3. ספריות שימושיות:
   * pandas, Matplotlib
   * Hugging Face Transformers: ספרייה למודלים מבוססי Transformers עם ממשקים פשוטים לאימון ושימוש במודלים קיימים.
4. פלטפורמות ענן - אחסון ואימון מודלים:
   * AWS (Amazon Web Services): פתרון מבוסס ענן הכולל את S3 לאחסון נתונים ו-EC2 להרצת מודלים.
   * Google Cloud Platform (GCP): פלטפורמה מתקדמת לאימון מודלים על GPUs ולניהול יישומים בענן.
   * Azure: פתרון נוסף לעיבודים מקביליים ותמיכה ב-Data Science.
   * ניהול גרסאות: שימוש ב-Git ו-GitHub לניהול קוד, גרסאות, ושיתוף פעולה בין חברי הצוות.

תוצרים צפויים:

1. מערכת אינטראקטיבית לחישוב והשוואת מדדי מורכבות טקסט
2. דו"חות הכוללים תובנות, השוואות וניתוחי סיבות שורש למורכבות טקסט
3. גרפים המציגים שינויים באורך, קריאות ומורכבות מבנית (שינויים בין טקסטים)

תוצרים נוספים אפשריים:

1. פרסום מאמר למחקר בנושא קוד המס האמריקאי לפני ואחרי חקיקת חוק TCJA משנת 2017
2. פרסום מאמר אקדמי/אחר הנוגע למדד מורכבות טקסט שיפותח

צפי תרומה:

המערכת תתרום להבנת מורכבות טקסטים רגולטוריים ותסייע לשיפור נגישותם. המדד הכמותי יהווה כלי שימושי לחוקרים, יועצים ומקבלי החלטות.

בהיבט של המחקר העוסק בקוד המס בארה"ב - הפרויקט יתרום לשיפור ההבנה של השפעת חקיקה חדשה על קוד המס. התובנות יסייעו למקבלי החלטות ליצור חוקים נגישים יותר, ולמשתמשים להבין את קוד המס בצורה יעילה.