הצעת פרויקט: קורס עיבוד שפה טבעית (NLP)

מגיש: יניק גוטיה 318553526

שם זמני: הנגשת זכויות חברתיות באמצעות NLP: פיתוח מערכת אחזור סמנטי לשאלות בשפה טבעית.

1. מסלול הפרויקט

בחרתי לבצע את הפרויקט במסגרת **מסלול 3 – Data-Driven NLP Project**. מסלול זה נבחר היות שהוא מאפשר התמודדות עם Pipeline מלא של עבודה יישומית: החל מאיסוף עצמאי של נתונים (Data Collection), דרך עיבוד מקדים (Preprocessing) והנדסת תכונות, ועד להשוואה אמפירית בין גישות מידול שונות, ניתוח שגיאות והערכה.

2. נושא הפרויקט ומטרותיו

מטרת הפרויקט היא לפתח מערכת לאחזור מידע סמנטי בעברית, שתסייע למשתמשים למצוא מידע רלוונטי בנושאי זכויות, קצבאות והליכים בירוקרטיים. המערכת תאפשר למשתמש להזין שאלה חופשית בשפה טבעית (לדוגמה: "מה מגיע לי אחרי לידה?" או "איך מקבלים קצבת נכות?"), ותאתר את הפסקאות או המסמכים הרלוונטיים ביותר מתוך קורפוס ייעודי שייבנה מאתרי "כל-זכות" והמוסד לביטוח לאומי.

לב הפרויקט הוא הניסיון לגשר על הפער שבין "שפת היומיום" של הפונים לבין הניסוחים הפורמליים והמשפטיים המופיעים במקורות המידע הרשמיים, תוך שימוש בשיטות NLP מתקדמות.

3. שאלת המחקר

שאלת המחקר המרכזית בפרויקט היא: **באיזו מידה אחזור סמנטי מבוסס Sentence Embeddings משפר את איכות איתור המידע בתחום הזכויות בעברית, בהשוואה לאחזור לקסיקלי קלאסי (TF-IDF)?**

בנוסף, אבחן האם שילוב היברידי בין שתי השיטות (Lexical + Semantic) מוביל לשיפור נוסף בביצועי המערכת.

4. נתונים (Data)

- **מקורות:** הקורפוס יתבסס על דפי תוכן ומידע מאתרי "כל-זכות" וביטוח לאומי. המידע ייאסף באמצעות כלי איסוף נתונים (Scraping), בדקתי את robots.txt של שני האתרים והם בהחלט מאפשרים בכפוף לתנאי השימוש, ולמגבלות הרישוי של האתרים.
- https://www.kolzchut.org.il/robots.txt | https://www.btl.gov.il/robots.txt .
- **מבנה:** הנתונים יעובדו ויישמרו במבנה מסודר הכולל כותרת, טקסט גולמי, חלוקה לפסקאות ומקור הנתונים.
- **סט הערכה (Evaluation Set):** אבנה באופן ידני סט מבחן הכולל שאלות בשפה טבעית והצמדתן למסמכי ה-Ground Truth הרלוונטיים, כדי לאפשר הערכה כמותית של המודלים.

5. שיטות עבודה (Methodology)

הפרויקט יבוצע בשלבים הבאים:

1. **איסוף ועיבוד מקדים:** ניקוי טקסט, הסרת רעשים וחלוקה ליחידות תוכן (Chunking).
2. **מידול:** יישום שתי גישות לפחות:
   - מודל Baseline מבוסס **TF-IDF** (אחזור מבוסס מילות מפתח).
   - מודל סמנטי מבוסס **Sentence Embeddings** (שימוש במודלים כגון BERT בעברית או מודלים רב-לשוניים, **אשמח להמלצה!**).
3. **הערכה:** השוואת הביצועים באמצעות מדדים מקובלים בעולם ה-Information Retrieval, כגון **Recall@k** ו-**MRR** (Mean Reciprocal Rank).
4. **ניתוח שגיאות ווויזואליזציה:** הצגת השגיאות הנפוצות (למשל, מקרים של מילים נרדפות שהמערכת פספסה) והצגת התוצאות בגרפים השוואתיים.

6. אתגרים צפויים

- **מורכבות השפה:** התמודדות עם מורפולוגיה עברית עשירה ופער סגנוני בין שפת המשתמש לשפה הרשמית.
- **ניקוי נתונים:** טיפול במבני אתרים שונים ומיצוי הטקסט הרלוונטי בלבד.
- **בניית סט הערכה:** יצירת סט שאלות-תשובות איכותי ומייצג בזמן מוגבל.

7. סיכום

הפרויקט מציע פתרון יישומי לבעיה חברתית ובירוקרטית ממשית תוך שימוש בכלים טכנולוגיים מתקדמים. העבודה משלבת את כל שלבי ה-Pipeline של פרויקט NLP מבוסס נתונים, ותורמת להבנת היכולות של אחזור מידע סמנטי בשפה העברית.