שלום וברכה

הפרויקט הוא בעצם מנוע חיפוש בעברית ובאנגלית. שלבי פעולת המנוע:

- בהתאם לשאילתא שהוכנסה, המנוע מאבחן את השפה ובוחר באיזה מערכת קבצים להשתמש (קבצי טקסט בעברית או באנגלית).
- מערכת הקבצים מעובדת מראש – כל המסמכים עובר עיבוד ראשוני להורדת תווים מיוחדים, ניקוד, stop words ולבסוף מעבר לצורת בסיס (lemmatization).
- כמו כן לכל קובץ מוקצה וקטור BOW בהתאם למילון של הקורפוס כולו, וכן וקטור מייצג על מודל doc2vec שאומן על קורפוס מבעוד מועד.
- המשתמש יכול לבחור האם להשתמש במערכת הקבצים הקיימת או לעדכן אותה על ידי מעבר מחדש על כל המסמכים ועיבודים, ולאחר מכן אימון המודלים (תהליך הלוקח זמן רב).
- לאחר הכנסת השאילתא, היא עוברת את אותו עיבוד זהה למסמכים ומוקצים לה הוקטורים הנדרשים.
- המערכת מחשבת את ערכי הדמיון בין כל מסמך לשאילתא על פי כל המדדים הבאים:
  - דמיון בין וקטורי BOW על בסיס מכפלה סקלרית.
  - דמיון מדד ג'קארד בין וקטורי BOW.
  - דמיון מבוסס שיטת TF-IDF עם טרנספורמציית BM25.
  - דמיון מבוסס מרחב וקטורי שאומן במודל doc2vec.
  - דמיון מבוסס על כלל המדדים.
- המערכת שואלת את המשתמש באיזה פונקציית דמיון להשתמש, ומדפיסה את 30 המסמכים הראשונים על פי דירוג זה (סדר יורד).
- לאחר מכן ניתנת אפשרות למשתמש להכניס שאילתא נוספת או לסיים.

המערכת נכתבה בשפת פייתון בגרסא 3.8.

כדי להריץ את המערכת יש להתקין את כל החבילות המיובאות, להתקין בנפרד אפליקציית yap (הוראות התקנה פה) ולהריץ אותה כשרת HTTP בפורט 8000, ולאחר מכן להריץ את הקובץ main בלבד.

ניתן לפנות אליי לכל שאלה

רון קינן

Ronke21@gmail.com