תיאור הפרויקט

מגישים: אייל הוכשטד 315341289 יניב כוגן 208417204

1. בנינו offline את הinverted index, על בסיס מאגר מסמכי הxml שהיו נתונים לנו. לכל מסמך בתיקייה המתאימה, ולכל רקורד בו, השתמשנו בכותרת, ובקטעי הטקסט הנלווים לרקורד כדי לבנות vocabulary מעליו נעבוד, תוך עיבוד מקדים של הטקסט כפי שדיברנו בהרצאה (stemming, tokenization, remove of stopwords) וכן טיפול במקרים שונים (הסרת מספרים, הסרת מחרוזת ריקה ורווחים). לאחר מכן, חישבנו לכל token את ציון הidf שלו וציון הtf שלו ביחס לכל מסמך בקורפוס ובאמצעות כך את ציון הtf-idf.. בנוסף חישבנו את האורכים של כל המסמכים כפי שתואר בהרצאה. אחסנו את כל מבני הנתונים במילונים מקוננים באופן כזה, שכאשר יתבצע חישוב online של המסמכים הרלוונטים עבור שאילתה מסוימת, תוחזר תשובה כמה שיותר מהר.

2. כתבנו קוד שמחשב את המסמכים הרלוונטים עבור שאילתה נתונה:

ראשית, ביצענו עיבוד מקדים לשאילתה עצמה, בדומה לעיבוד בסעיף 1, והתעלמנו מהמילים שאינן בvocabulary אותו בנינו בסעיף 1.

שנית, חישבנו לכל token רלוונטי מהשאילתה לאחר העיבוד את ציוני הidf והtf שלו, ובאמצעות כך את ציון הtf-idf.

לאחר מכן, חישבנו את ציון הcosine similarity של כל מסמך ביחס לשאילתה, ובאמצעות מיון הציונים בסדר יורד, חילצנו את המסמכים הרלוונטים ביותר.

תוך ניסוי ותהייה, קבענו סף לציון הcosine similarity (הסף היה ממוצע משוקלל עם פרמטר קבוע, של חציון ציוני הרלוונטיות שקיבלנו במסמכים, ושל 1) כדי להחליט איזה מסמך רלוונטי ואיזה לא.

3. אמדנו את איכות מערכת אחזור המידע שבנינו, באמצעות מאגר המידע Cystic Fibrosis Database ובאמצעות קובץ מאגר השאלות שניתן לנו:

חישבנו ציוני precision וrecall על גבי השאילתות השונות, תוך שימוש באלמנטי הResults וה-Records הנלווים לכל שאילתה. החישובים התבצעו בדומה לנוסחאות שראינו בהרצאה. בנוסף, חישבנו ציוני cumulative gain וf score, גם כן בדומה למה שראינו בהרצאה.

4. קבצי עזר נוספים:

- הקובץ vectors.py שימש אותנו לחישוב הנורמה כחלק מחישוב ציון ה- cosine similarity.

- הקובץ tokenizer.py בו מימשנו את העיבוד המקדים לטקסט.

- הקובץ query.py באמצעותו ביצענו את חלק 2 המתואר לעיל.

- הקובץ corpus_index.py באמצעותו ביצענו את חלק 1.

- הקבצים test_queries_parser.py, combined_queries_scores.py באמצעותם ביצענו את חלק 3 המתואר לעיל.