**מגיש1 :** אופק ברנסקי **03683062 ניהול נתונים באינטרנט**

**ת.ז.** 311170138 **פרויקט תכנות 2**

**מגיש 2:** מתן טולדו **ת.ז.** 314023409

1. **תיאור קוד הפרויקט:**

הקוד שלנו מורכב מקובץ python ומאגר קבצי XML (ה Corpus)

**vsm\_ir.py-** מודול המקבל את הקלט מהמשתמש, מבצע pre-computationשל ה inverted index, ומוצא את הקבצים הדומים ביותר לשאילתא, במודל bm25 ובמודל tf-idf.

**Tester.**py- טסטר שבנינו לפרויקט, מפורט עליו בהמשך.

A picture containing graphical user interface

Description automatically generated

**כעת נרחיב על מודול:**

1. המודול מקבל את קלט משתמש:
   1. במידה והקלט שאילתא, המודול מבצע את הפעולות הבאות:
      1. מפרסר את סוג המודל המתבקש – **bm25** או **tfidf**
      2. מחשב ע"פ הנלמד בשיעור את הציון **bm25** או **tfidf** של השאילתא עם כל המסמכים שבוצע אליהם preprocessing.
      3. מחזיר את ה Top k מסמכים הדומים ביותר (כאשר כמות המסמכים שמוחזרים configurable בתוך הקוד.
   2. במידה והקלט create, הוא מבצע את חישוב ה inverted index באופן הבא:
      1. מבצעים איטרציה על כל המסמכים (עבור כל מסמך, שולפים בעזרת xpath את הTitle, Extract ו Abstract שלו.
      2. עבור כל מסמך אנו מוחקים את כל הstop words (בעזרת החבילה nltk), מבצעים stemming בעזרת nltk.PorterStemmer וכמובן מוחקים סימני פיסוק ותווים שאינם אותיות באנגלית.
      3. לאחר "טיפול" בטקסט, אנו מבצעים tokenize של המילים שנותרו לנו, ומחשבים את הinverted index כפי שהוצג בכיתה.
      4. לאחר מכן מבצעים חישוב של ה idf scores, וה document vector norms
      5. לסיום, אנו שומרים את התוצאה בקובץ json
2. הקוד מוחלק לאובייקט המייצג את ה Vector Space Model:

Text

Description automatically generated

1. והכנסנו עוד מספר פונקציות עזר לטובת שמירה וחישובים של ערכים שונים:

Text

Description automatically generated

1. **ניתוח התוצאות:**

ע"מ לנתח את תוצאות המודל, ולמקסם את תוצאותיו, ביצענו מספר דברים:

1. ראשית יצרנו קובץ test המריץ את כל השאילתות שניתנו (99 כאלו) על המודל שלנו,  
   עבור כל שאילתא, חישבנו את כלל הציונים האפשריים – NDCG, Precision, Recall, F value .  
   כאשר הפלט הסופי של הטסטר היה ממוצע הציונים של המודל על 99 השאילתות הרלוונטיות.
2. ע"מ למקסם, בדקנו את השפעת הפרמטרים הבאים על התוצאות, ובחרנו בפרמטרים שמקסמו את הציונים עבור המודל:
   1. **Bm25** **k** **value-** ההיפר-פרמטר שנלמד בשיעור
   2. **Bm25 b value-** ההיפר-פרמטר שנלמד בשיעור
   3. **Document threshold-** כמות המסמכים המקסימלי שיוחזר עבור שאילתא מסוימת,  
      היה ניכר כי ככל שהעלינו את כמות המסמכים שהוחזרו, ה Recall גדל אך ה Precision קטן:

Document threshold – TF-IDF and BM25:

|  |  |
| --- | --- |
|  |  |

BM25 – k value, b value:

|  |  |  |
| --- | --- | --- |
|  | A screenshot of a computer  Description automatically generated with medium confidence |  |

1. **סיכום:**

לאחר ניתוח של הפרמטרים וביצועי התוכנית, הגענו למסקנה כי הנתונים הטובים ביותר הינם:

Bm25\_K = 1.2

זהו הערך שנותן ציון NDCG גבוה ביותר.

Bm25\_b = 0

זהו הערך שנותן ציון NDCG גבוה ביותר.

Document\_threshold = 7

זה ערך שנותן ציון NDCG גבוה, אבל גם מבטיח כיסוי מסמכים מסוים שנותן גם ציון F גבוה.

אלו המדדים הסופיים עם הפרמטרים האלו על פני 99 שאילתות:

