Gagal Retrieval Search-Engine

Student:
Gal Tadir, Gal Azaria

Teacher:
Nir Greenberg

Course:
Information Retrieval

כללי:

תחילה על מנת להפעיל שאילתה בודדת או מספר שאילתות יחדיו על המנוע, נדרש להפעיל ולאתחל את האינדקס.

על מנת לאתחל את האינדקס המשתמש מתבקש להכניס את ה Corpus Path שזה הוא המיקום של מאגר המידע שעליו המשתמש רוצה לעבוד ולבנות את האינדקס. בנתיב זה גם צריך להיות קובץ ה Stop Words.

לאחר מכן על המשתמש נדרש להכניס את המיקום אליו הוא רוצה לכתוב את קבצי ה Post File. נתיב זה ישמש בהמשך גם על מנת לבצע את הטעינה.

אם הנתיב יהיה מלא בקבצים, המערכת תדרוש מהמשתמש או למחוק את הקבצים האלו או לבצע טעינה של המערכת בעזרת הקבצים הקיימים.

בשלב זה המשתמש צריך להחליט האם להשתמש באופציית הStemming בעזרת הCheck Box.

בעת לחיצה על כפתור ה Start המערכת תתחיל את פעולת האינדוקס.

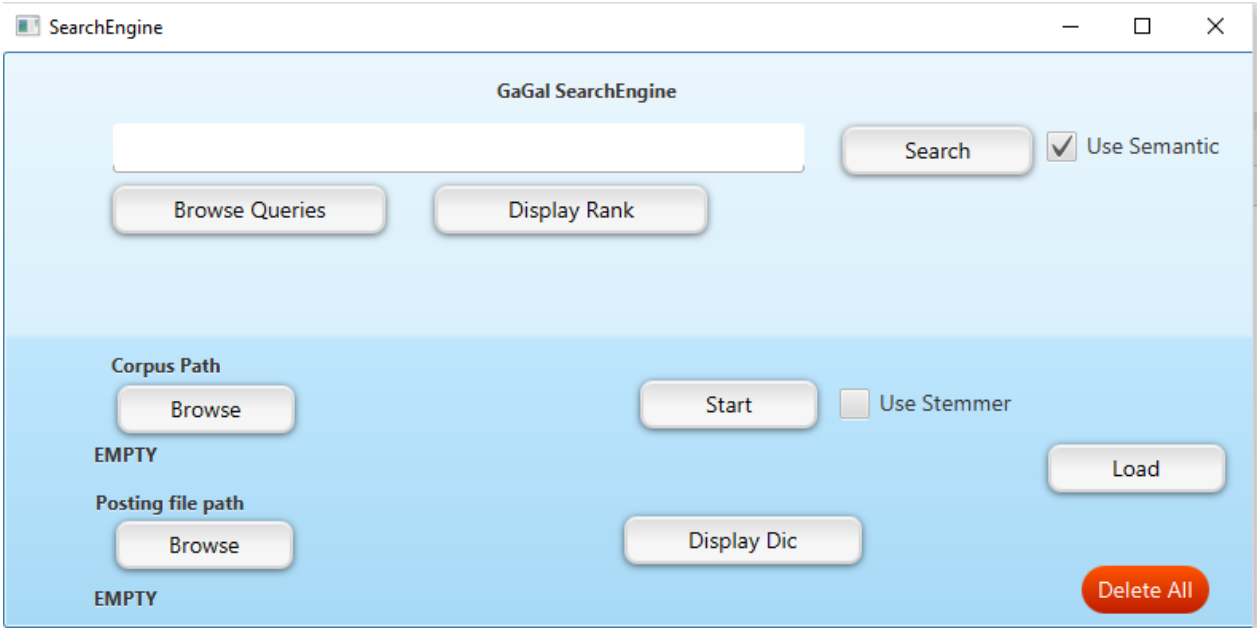
לאחר שהפעולה תסתיים, ניתן יהיה לבצע חיפושים.

ישנן 2 אופציות של חיפוש, אופציה ראשונה היא להריץ שאילתה בודדת בעזרת המקום המתאים לכתיבת שאילתה, במצב כזה המערכת תדע לנתב את החיפוש אך ורק לשאילתה הספציפית הזו ותציג את התוצאות על פיה.

במצב בו מיקום השאילתה ריק, המערכת תנסה לחפש שאילתות דרך הנתיב המיועד לכך, כאשר המשתמש הוא האחראי להכנסת הנתיב.

לאחר לחיצה על כפתור ה Search יתבצע הדירוג המתאים לפי השאילתה, עם לחיצה על כפתור עם Display Rank ניתן יהיה לראות את כל המסמכים הרלוונטים ואת הדירוג שלהם עלפי כל שאילתה שחיפשנו עבורה.

הוראות הפעלה:



- Corpus Path – מיקום מאגר המידע המקורי ובנוסף מיקום ה Stop-Words.
- Posting File Path – מיקום קבצי המילון. לכתיבה או לקריאה.
- Start – התחלת ביצוע האינדוקס למאגר המידע הקיים בCorpus Path , הקבצים לאחר האינדוקס ישמרו ב- PostingFile Path.
- Display Dic – הצגת המילון הסופי, קיים רק לאחר לחיצה על Start.
- Load – ביצוע טעינה של המילונים ושל קבצי הPostFile הקיימים ב- Posting File Path. על מנת לבצע טעינה תקינה, יש להוסיף גם את מיקום הStopWords ב Corpus Path.
- Delete All – כפתור מחיקה של כל קבצי המילון וקבצי הPosting הקיימים. לא ניתן למחוק אם לא קיימים קבצים.
- Browse Queries – מיקום השאילתות.
- Display Rank – כפתור זה מציג תחילה את כל השאילתות אשר בוצע עליהן חיפוש, לכל שאילתה קיים כפתור RANK אשר מציג את כל המסמכים הרלוונטיים לאותה שאילתה, לכל מסמך קיים כפתור המציג את כל הישויות הרלוונטיות למסמך זה.
- Search – הפעלת החיפוש. כאשר קיים טקסט בחלון השאילתה, החיפוש יתבצע עליו ולא כל קבצי השאילתות!

תיאור המחלקות:

- **Searcher –**
- מחלקת searcher תפקידה לקבל נתיב לקובץ שאילתות או שאילת מהמשתמש ולהחזיר דירוג המסמכים הרלוונטיים ביותר. לשם כך המחלקה משתמשת במחלקת parse על מנת לפרסר את השאילתה. לאחר פרסור השאילתה המחלקה קוראת למחלקת rank ושולחת לה את השאילתה מפורסרת. מחלקת rank מחזירה דירוג שאילתות למחלקת searcher.
- handleQueryByPath(String path) - השיטה מקבלת נתיב לקובץ השאילתות ומבצעת דירוג מסמכים לשאילתות מהמסמך.
- initQueries(String path) – שיטה פרטית אשר מקבלת נתיב לקבוץ ויודעת לפרק את הקובץ לשאילתות. לאחר פירוק הקובץ לשאילתות השיטה מאחסנת את כלל השאילתות בתוך מבנה נתונים.
-
- getRankingMap() – השיטה מחזירה מפת דירוג השמורה בתור שדה במחלקה.
- writeTrec_Eval() – לאחר הרצת השאילתות הפונקציה רושמת לקובץ את תוצאות ה trec-eval.

- **Ranker –**
- מחלקת Ranker מקבלת שאילתה מפורסרת (לפי כותרת ודירוג) ומחזירה מפת דירוגים, כאשר לכל שאילתה מוחזרת רשימת מסמכים רלוונטיים אשר לא עולה על 50 מסמכים, ולכל מסמך מוחזר הדירוג שלו ביחס לאותה שאילתה.
- ranking(Map<String, Integer> titleMap, Map<String, Integer> descriptionMap) – השיטה הראשית של המחלקה, השיטה מקבלת שני מבני נתונים המתייחסים לשאילתה אותה יש לדרג, מבנה נתונים אחד שומר בתוכו את כותרת השאילתה ומבנה נתונים נוסף שומר בתוכו את תיאור השאילתה. השיטה מנתחת את הכותרת והתיאור ביחס למסמכים בקורפוס ומחזירה דירוג למסמכים הרלוונטיים.
- setUseSemmantic(boolean useSemmantic) – שיטה המאפשרת למשתמש להגדיר ל ranker האם להשתמש בטיפול סמנטי בהליך הדירוג או לא.
- setUseStemmer(boolean useStemmer) - שיטה המאפשרת למשתמש להגדיר ל ranker האם להשתמש בסטמינג בהליך הדירוג או לא.
- getScore(Map<String,Integer>query, String docNum, Map<String,Map<String,Integer>> posting) – השיטה מקבחלת שאילתה, מסמך ואת השורות הרלוונטיות מקובץ הפוסטינג אשר נשלפו לפני הקריאה לפונקציה, השיטה מחשבת דירוג של השאילתה למסמך ומחזירה את הדירוג.

- getPosting(Set<String> keySet) – השיטה מקבלת set של המילים מהשאילתה מחפשת את המילים הללו בקבצי הפוסטינג ושומרת את השורות הרלוונטיות.
- calcIdf(String term) – פונקציה פרטית אשר מחשבתאת ערך ה idf לביטוי.
- sortQueries() – הפונקתיה נגשת למבנה הנתונים בו שמורים הדירוגים עבור כל השאילתות, הפונקציה ממיינת את מבנה הנתונים לפי הדירוג ומשאירה את המסמכים הרלוונטיים בלבד.
- rankingDescriptionMap(Map<String,Integer> descQ) – שיטה פרטית המקבלת תיאור של שאילתה ומעדכנת את הדירוג בהתאם לכלל המסמכים הרלוונטיים.
- rankingTitle(Map<String,Integer> titleQ) - שיטה פרטית המקבלת כותרת של שאילתה ומעדכנת את הדירוג בהתאם לכלל המסמכים הרלוונטיים.
- calcAVGDocsSize() – שיטה פרטית המחשבת את הגודל הממוצע של המסמכים בקורפוס.
- getlineInPostFile(String term) – שיטה פרטית אשר מקבלת ביטוי ומחזירה את מס' השורה הרלוונטי בקובץ פוסטינג המייצגת את הביטוי.
- openFilePost(int line) – שיטה פרטית המקבלת מס' שורה ומחזירה את הקובץ פוסטינג ממנו יש לשלוף את המידע על הביטוי הרלוונטי.
- foundInDic(String term) – שיטה פרטית אשר מקבלת ביטוי מהשאילתה ותפקידה וחפש את המילה במילון במספר צורות שונות (אותיות קטנות\גדולות).

- Query –
- מחלקה אשר תפקידה לייצג שאילתה. המחלקה מחזיקה את השדות הבאים: כותרת השאילתה, תיאור השאילתה ומספר השאילתה. למחלקה אין שיטות או פונקציות מלבד הבנאי, גטרים וסטרים. תפקיד המחלקה הוא לרכז מידע עבור כל שאילתה במקום אחד.

אלגוריתמים:

אלגוריתם הדירוג – האלגוריתם מקבלת בתור ארגומנטים שני Maps, Map אחד מייצג את כותרת השאילתה עליה יש להוציא דירוגים ו Map שני המייצג את תיאור השאילתה, האלגוריתם שולף תחילה את קבצי הפוסטינג הרלוונטיים לדירוג השאילתה המדוברת. על מנת לשלוף את השורות הרלוונטיות מהפוסטינג נבדוק במילון השמור בזיכרון ה RAM את מספרי השורות הרלוונטיות. לכל פוסטינג מוקצות 200,000 שורות, כלומר אם עבור מילה מסוימת נצטרך לשלוף את שורה 500,012 ניגש לשורה 100,012 בקובץ הפוסטינג השלישי. סה"כ יצרנו 10 קבצי פוסטינג בהתאם לכמות ה term במילון.

לאחר ששלפנו את המידע הנחוץ מקבצי הפוסטינג יש לנו את כלל המידע הנחוץ להפעלת נוסחת הדירוג BM25. נפעיל את נוסחאת הדירוג ובהתאם נעדכן את דירוג השאילתה. קבענו את ערכי הפרמטרים בצורה הבאה: $b = 0.4$, $k=1.2$ זאת כיוון שלאחר מספר הרצות ערכי הפרמטרים הללו החזירו את התוצאות המיטביות. ביחס בין הדירוג שהתקבל מהכותרת ולדירוג אשר התקבל מהתיאור הענקנו לכותרת יחס 0.7 ולתיאור יחס 0.3. בנוסף ל Term המופיעה כולו באותיות גדולות, נגדיל את ערכו בדירוג השאילתה פי 3, ולכן ישות נגדיל את ערכה בדירוג פי I*3 כאשר I זה כמות המילים ממנה בנויה הישות.

אלגוריתם למציאת 5 הישויות –

במהלך הליך האינדוקס יצרנו מבנה נתונים אשר מכיל לכל מסמך את רשימת הישויות שהופיעו בו וכמה פעמים כל ישות הופיעה. בסוף הליך האינדוקס כאשר עברנו כל ה terms במסמך סיננו את מבנה הנתונים כך שישמור עבור כל מסמך את 5 הישויות שהופיעו בו בתדירות הכי גבוהה בלבד, זאת מתוך הבנה שככל שישות מופיעה במסמך יותר פעמים כך הדומיננטיות שלה במסמך גדלה. לבסוף נתנו לכל ישות דירוג בטווח 1-10 בהתאם למסמר ההופעות של הישות במסמך. כאשר 10 הוא הדירוג המקסימאלי ו – 1 המינימאלי. דירוג הדומיננטיות נקבע גם הוא באופן ישיר בהתאם לתדירות הישות במסמך. כאשר המשתמש בוחר לאחזר מסמך מוצגת לו האפשרות לראות מי 5 הישויות הדומיננטיות ביותר במסמך.

אלגוריתם לשיפור סמנטי –

האלגוריתם לשיפור סמנטי תחילה בודק האם למחשב חיבור פעיל לאינטרנט, במידה וכן הוא ישלוף עבור כל ביטוי בשאילתה את 3 המילים בעלות המשמעות הסמנטית הקרובה ביותר מתוך מקור ה API. במידה והחיבור לאינטרנט אינו זמין הוא ישלוף את המידע מתוך קובץ JAR אשר הוכן מראש.

בביטויים אלו נשתמש על מנת "להגדיל" בצורה מלאכותית את כמות ה term המקורי בכל מסמך. לדוגמא אם ב doc1 המילה 'maybe' הופיעה 3 פעמים והמילה 'perhaps' מופיעה 2 פעמים. ובשאילתה Q1 המילה 'maybe' מופיעה פעם אחת נייצר 'סביבת עבודה' בה המילה 'maybe' מופיעה ב doc1 5 פעמים. מאותה נקודה נריץ את אלגוריתם החיפוש בצורה זהה לצורה הרגילה. שיטה זאת הראתה שיפור עקבי בתוצאות החיפוש.

קבצים:

הסבר על הנתונים בקבצי ה PostFile -

קבצי הPostFile מחלוקים בצורה הזו – כל קובץ מכיל במקסימום 200 אלף מילים, ולכל מילה מופיע שם המסמך אשר בו היא מופיע ומספר הפעמים שהמילה מופיעה בו. ההפרדה בין מילה לבין מסמכים הינה @. לדוגמא:

Term@FBIS3-1=1

הסבר על הנתונים במילון –

ישנם 4 קבצים:

1. DocsDicFile – קובץ המכיל את כל מספרי המסמכים הקיימים במאגר המידע, לכל מסמך אנו שומרים את הכותרת שלו, מילים ייחודיות, מספר מילים כללי והתדירות של המילה שמופיע הכי הרבה פעמים. ההפרדה בין מסמך לפרטים הינה ~.

   Headline~UniqueWords~TotalWords~MaxFreq~

   לדוגמא: FBIS3-8843~CTK Profiles New Cabinet Members~212~310~14

2. TermsDicFile – קובץ המכיל את כל המילים המופיעות במאגר המידע, לכל מילה אני שומרים את מספר המסמכים בהם המילה הופיע, מספר ההופעות הכולל, מספר שורה בקובץ הPostFile והתדירות הכי גבוה בה המילה מופיע באחד המסמכים.

   DocsCounter~TotalAppearance~PointerLine~MaxApperance

   לדוגמא: flourishment~3~3~964338~1

3. EntitiesFile – קובץ המכיל לכל מסמך את 5 הישויות הדומיננטיות ביותר בו.
4. EntitiesFile2 – קובץ המכיל את כלל הישויות בקורפוס, זאת על מנת לאתחל את הפרסר כך שיידע מי ישות ומי לא.

הערכת מנוע:

*המספרים באדום מסמנים את התוצאות לאחר הפעלת האלגוריתם לניתוח סמנטי.

Qid	מילות השאילתה	Precision	Recall	Precision@5	Precision@15	Precision@30	Precision@50	Time(ms)
351	Falkland petroleum exploration	0.4 0.4	0.41 0.41	0 0	0.4 0.4	0.433 0.433	0.4 0.4	1942 1918
352	British Chunnel impact	0.32 0.22	0.06 0.04	0.4 0.2	0.33 0.266	0.3 0.3	0.22 0.22	2110 2086
358	blood-alcohol fatalities	0.44 0.44	0.43 0.43	0.2 0.2	0.46 0.46	0.5 0.5	0.44 0.44	1656 1656
359	mutual fund predictors	0.06 0.1	0.1 0.1	0 0	0.06 0.06	0.1 0.1	0.1 0.1	2022 2013
362	human smuggling	0.18 0.18	0.23 0.23	0.2 0.2	0.2 0.2	0.2 0.13	0.18 0.18	1491 1470
367	piracy	0.367 0.38	0.1 0.1	0.8 0.8	0.46 0.46	0.36 0.36	0.367 0.38	2217 2193
373	encryption equipment export	0.06 0.06	0.187 0.187	0.2 0.2	0.2 0.2	0.1 0.15	0.06 0.06	2380 2358
374	Nobel prize winners	0.54 0.54	0.13 0.13	0.6 0.6	0.6 0.6	0.63 0.63	0.54 0.54	3710 3698
377	cigar smoking	0.2 0.24	0.27 0.33	0 0	0.133 0.133	0.16 0.2	0.2 0.24	1373 1363
380	obesity medical treatmen	0.1 0.1	0.71 0.71	0 0	0.2 0.2	0.13 0.13	0.1 0.1	1819 1816
384	space station moon	0.2 0.2	0.19 0.19	0.4 0.4	0.2 0.2	0.166 0.166	0.2 0.2	2699 2692
385	hybrid fuel cars	0.3 0.34	0.17 0.2	0 0	0.266 0.33	0.33 0.4	0.3 0.34	2860 2841
387	radioactive waste	0.26 0.26	0.17 0.17	0.4 0.4	0.2 0.2	0.233 0.233	0.26 0.26	4079 4064
388	organic soil enhancemen	0.14 0.18	0.14 0.18	0.2 0.2	0.2 0.266	0.16 0.233	0.14 0.18	3721 3693
390	orphan drugs	0.32 0.26	0.1 0.1	0.4 0.4	0.266 0.266	0.33 0.3	0.32 0.26	3005 2993
								Map =0.0692

עם ביצוע סטמינג								
Qid	מילות השאילתה	Precision	Recall	Precision@5	Precision@15	Precision@30	Precision@50	Time(ms)
351	Falkland petroleum exploration	0.3	0.31	0	0.26	0.26	0.3	1575
352	British Chunnel impact	0.22	0.04	0.2	0.266	0.3	0.22	1549
358	blood-alcohol fatalities	0.4	0.39	0.2	0.33	0.33	0.4	1655
359	mutual fund predictors	0.04	0.07	0	0.06	0.06	0.04	1772
362	human smuggling	0.16	0.2	0	0	0.1	0.16	819
367	piracy	0.38	0.1	0.6	0.266	0.366	0.388	1794
373	encryption equipment export	0.04	0.125	0	0.13	0.06	0.04	2736
374	Nobel prize winners	0.32	0.07	0.6	0.466	0.33	0.32	2317
377	cigar smoking	0.24	0.33	0	0.133	0.166	0.24	1863
380	obesity medical treatmen	0.1	0.71	0	0.133	0.133	0.1	2531
384	space station moon	0.2	0.19	0.4	0.2	0.133	0.2	3059
385	hybrid fuel cars	0.4	0.235	0	0.33	0.43	0.4	3405
387	radioactive waste	0.3	0.2	0.4	0.266	0.2	0.3	4341
388	organic soil enhancemen	0.2	0.2	0.2	0.33	0.233	0.2	3856
390	orphan drugs	0.3	0.122	0.2	0.133	0.33	0.3	7180
MAP =0.058								