ניהול נתונים באינטרנט 2022

פרויקט תכנות – Information Retrieval

הוראות כלליות

יש לעלות את הפתרונות בקובץ ZIP לMOODLE, שכולל קובץ PDF בשם answers.pdf ובו הפתרון וקבצי קוד נוספים (HTML, XML או Python) לפי הדרישה של כל סעיף וסעיף. ההגשה היא בזוגות, כך שרק אחד מבני הזוג יגיש את התרגיל, אך יש להקפיד לכתוב את השמות והת.ז. של שני בני הזוג בתוך הקובץ.
שם של הקובץ ZIP צריך לכלול את הת.ז. של שני המגישים, למשל: HW_IR_123_456.zip

תאריך פרסום: 28/04/2022 **תאריך הגשה: 02/08/2021**

רקע

זהו פרויקט תכנות בנושא אחזור מידע (Information Retrieval) בו תממשו מנוע חיפוש מבוסס Vector Space Model. עליכם לבנות מערכת אוטומטית שבהנתן שאלה בשפה טבעית ומאגר מסמכים, תחזיר למשתמש את אוסף המסמכים הרלוונטים ביותר לשאלה.
הפרויקט להגשה עד לתאריך שמצוין למעלה ומהווה 11% מהציון הסופי בקורס.

תיאור המערכת

התכנית שתכתבו תבנה מנוע חיפוש בשני שלבים:

א. התוכנית תבנה Inverted Index מתוך מאגר של מסמכי XML
ב. התכנית תקבל כקלט שאלה מהמשתמש ובעזרת האינדקס (שבנתה offline) תחזיר רשימה של מסמכים מהמאגר, מדורגים לפי הרלוונטיות לשאלה

למשל, עבור מאגר מאמרים אקדמיים בנושאי רפואה, התוכנית תבנה Inverted Index ותשתמש בו כדי להחזיר מאמרים רלוונטים לשאלות כגון:

- Is salt (sodium and/or chloride) transport/permeability abnormal in CF?
- What abnormalities of insulin secretion or insulin metabolism occur in CF patients?
- Can CF be diagnosed prenatally?
- …

שלב א': בניית Inverted Index

התכנית תקבל כקלט נתיב לתיקייה המכילה מאגר מסמכי XML שישמשו לבניית ה-Inverted Index. כל מסמך במאגר הוא אובייקט XML ייחודי, ראו בהמשך.
תכנית המחשב תעבור על כל המסמכים במאגר ותבנה מתוכם את האינדקס כפי שלמדנו בהרצאה:

1. בניית המילון לאינדקס
   a. מעבר על המילים במסמך, למשל בכותרת ובסיכום שלו (extract)
   b. ביצוע tokenization למילים במסמך
   c. הסרת מילים שכיחות stopwords, ניתן למצוא מאגר stopwords בחיפוש אונליין
   d. ביצוע stemming בעזרת Porter Stemmer למילים שאינן stopwords (שלב זה אינו חובה)
2. חישוב ושמירת ציוני TF ו-IDF לכל מילה במילון, עבור כל מסמך שבו היא מופיעה
3. חישוב ושמירת אורכי המסמכים במאגר (נחוץ לדירוג BM25 בשלב האחזור)

שלב ב': אחזור מידע בהנתן שאלה

התכנית תקבל כקלט את ה-inverted index שבנינו ושאלה בשפה טבעית מהמשתמש. בהנתן השאלה, המערכת תחזיר את רשימת המסמכים שרלוונטים לשאלה (אם ישנם כאלה). על המסמכים להיות מדורגים לפי ציון הרלוונטיות שלהם.
נשתמש בשתי פונקציות הדירוג שלמדנו בהרצאה, TF-IDF ו-BM25.

TF-IDF:
נשתמש ב-Vector Space Model כפי שלמדנו בהרצאה.
נחשב בצורה אינקרמנטלית את הציון של כל מסמך בהנתן המילים בשאלה:

1. נעבור על כל מלה רלוונטית בשאלה (או על ה-stem שלה אם בחרנו להשתמש ב-stemming)
2. נמצא כל מסמך שמכיל מילה זו
3. נחשב את ציון המסמך בעזרת נתוני ה-TF-IDF הרלוונטים ששמורים ב-index

לבסוף, נחזיר את רשימת המסמכים הרלוונטים מדורגים לפי TF-IDF + cosine similarity.

BM25:
ציון ה-BM25 של שאלה ומסמך יהיה שווה לסכום הציונים של כל מלה בשאלה והמסמך:

1. נעבור על כל מלה רלוונטית בשאלה (או על ה-stem)
2. נמצא כל מסמך שמכיל מילה זו
3. נחשב את ציון ה-BM25 של המלה והמסמך

לבסוף, נחזיר את רשימת המסמכים הרלוונטים מדורגים לפי ציון BM25

$$\mathrm{BM25}(D,Q) := \sum_{q_i \in Q} idf(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\mathrm{avgdl}}\right)}$$

$$idf(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

מאגר המידע לפרויקט: Cystic Fibrosis Database

מאגר המידע שלנו הוא Cystic Fibrosis Database. המאגר מכיל 1,239 תקצירי מאמרים על מחלת הסיסטיק פיברוזיס שפורסמו בין 1974-1979. בנוסף, מהמאגר מכיל 99 שאלות באנגלית. לכל שאלה מצורפים שמות המסמכים הרלוונטים אליה בצירוף ציון של 4 שופטים (0-2).

המאגר כולו זמין ב- Moodle כקובץ tar בגודל 1.54Mb הכולל את המאגר והשאלות כולן כקבצי XML. מסמכי ה-Document Type Definition (DTD) של קבצי ה- XML כלולים גם הם במאגר.

קבצי המסמכים:
ישנם 6 קבצים cf74.xml, ..., cf79.xml בהם כל המסמכים הרלוונטים לבניית האינדקס. בכל קובץ XML מסמך (מאמר) מיוצג כאובקייט XML בשם RECORD ובו המידע על כל מסמך. פירוט על כל אובייקט ניתן למצוא בקבצי ה-DTD. האובייקטים הרלוונטים הם:

- RECORDNUM מזהה המסמך בו נעשה שימוש גם בקובץ השאלות
- TITLE כותרת המסמך
- EXTRACT חלק מהמסמך
- ABSTRACT סיכום קצר של המסמך

ניתן, אך לא חובה, לעשות שימוש באובייקטים נוספים לצורך שיפור תוצאות החיפוש לבחירתכם (למשל ציטוטי מאמרים). במידה ועשיתם שימוש באובייקטים נוספים, הוסיפו הסבר קצר בקובץ ה-answers.pdf שאתם מגישים.

קובץ השאלות:
קובץ השאלות cfquery.xml מכיל 99 שאלות בשפה טבעית בנושאי סיסטיק פיברוזיס. כל שאלה נמצאת באובייקט XML בשם QUERY. האובייקט QueryText מכיל את תוכן השאלה. האובייקט Results מכיל את מספר המסמכים הרלוונטים לשאלה. האובייקט Records מכיל את כל מזהי המסמכים הרלוונטים לשאלה.

לכל מסמך יש ציון (score) של 4 שופטים בין 0-2. *הציון המשוקלל* של המסמך יהיה סכום ציוני השופטים בין 0 (0000) ל-8 (2222). שימוש לב ל"הערה על בדיקת הפרויקט" בנוגע לשימוש נאות בקובץ השאלות.

זכויות יוצרים:
כל זכויות היוצרים על המאגר שמורות למחברים מ- School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA. המאמר בו פורסם המאגר הוא:

Shaw, W.M. & Wood, J.B. & Wood, R.E. & Tibbo, H.R. The Cystic Fibrosis Database: Content and Research Opportunities. LISR 13, pp. 347-366, 1991

מסמך דרישות והרצת הקוד

- על הקוד להיות כתוב באמצעות Python 3 בלבד
- על קובץ ה-zip להכיל מסמך דרישות בשם requirements.txt ובו כל הספריות החיצוניות שנחוצות לשם הרצת הקוד, כל חבילה בשורה נפרדת. לדוגמה:

```
numpy
nltk
lxml
json
…
```

- נוכל לייצר את מסמך הדרישות באמצעות הרצת הפקודה freeze, שתדפיס את כל ספריות הפייתון בסביבת ההרצה שלנו
  - כנסו לסביבת הפייתון בה כתבתם את התרגיל
  - כתבו את שתי השורות הבאות:
    - `from pip._internal.operations import freeze`
    - `print('\n'.join(freeze.freeze()))`
  - תוכן קובץ הrequirements צריך להיות הפלט של ההדפסה
- פרויקט שלא יכיל מסמך דרישות ולא יסיים לרוץ ללא שגיאות <u>לא יקבל ציון עובר</u>
- הפרויקט יכיל קובץ פייתון בשם vsm_ir.py אשר ממנו תתבצע הרצת הקוד
  - קובץ ההגשה יכול לכלול קבצי פייתון נוספים אך יש לתעד את השימוש של כל קובץ במסמך answers.pdf שמצורף להגשה

- כדי ליצור את ה-inverted index התוכנית תרוץ משורת הפקודה באופן הבא:

```
python vsm_ir.py create_index [corpus_directory]
```

  - create_index משתנה שמצביע על יצירת ה-inverted index
  - corpus_directory הוא ה-path לתיקייה בה נמצאים קבצי ה-XML של מאגר המידע

- קלט לדוגמה:

```
python vsm_ir.py create_index /my_dir/data/cfc-xml
```

- התוכנית תשמור את inverted index לדיסק תחת השם vsm_inverted_index.json
  - הפורמט של הקובץ נתון לבחירתכם
  - על קובץ האינדקס להשמר בתיקיית ההרצה של הקוד (לא בתיקיית מאגר המידע)
  - הריצו את התכנית על המאגר וצרפו את הקובץ vsm_inverted_index.json להגשה

- כדי להחזיר את המסמכים הרלוונטים לשאלה התכנית תרוץ באופן הבא:

```
python vsm_ir.py query [ranking] [index_path] "<question>"
```

  - ranking מייצג את פונקציית הדירוג ויהיה אחד משני ערכים, `tfidf` או `bm25`
  - index_path הוא ה-path לקובץ ה-inverted index שיצרה התכנית
  - התוכנית תקבל כקלט שאלה באנגלית מהמשתמש, תטען את ה-inverted index מהדיסק ותחזיר את כל מזהי המסמכים הרלוונטים לשאלה, כשהם מדורגים לפי ציון הדירוג (סדר יורד)
    - התוכנית תשמור את התוצאות בקובץ ranked_query_docs.txt
    - המסמך יכול להיות ריק במידה ואף מסמך אינו רלוונטי לשאילתא
    - קובץ תוצאות לדוגמה:

```
494
536
1143
78
```

    - בדוגמה חזרו 4 מסמכים רלוונטים לשאלה. המסמכים מדורגים כך שמסמך 494 הוא הכי רלוונטי ו-78 הכי פחות
- הרצת הקוד תתבצע משורת הפקודה בלבד
- על התכנית להסתיים לאחר הרצת הפקודה (create_index או query), אין להשאיר את התכנית רצה

הוראות הגשה

יש להגיש קובץ ZIP יחיד בשם HW_IR_[id1]_[id2].zip עם מספרי תעודת הזהות של שני המגישים. על קובץ הזיפ לכלול את המסמכים הבאים:

- answers.pdf ובו תיאור כללי של הפרויקט. יש לכתוב את תעודות הזהות של המגישים בראש המסמך
- vsm_ir.py קובץ פייתון שמריץ את התכנית
- requirements.txt קובץ ובו כל ספריות הפייתון הדרושות לשם הרצת התכנית
- vsm_inverted_index.json קובץ ה-inverted index שיצרה התכנית כאשר הרצתם אותה על המאגר
- כל קובץ פייתון נוסף שבחרתם להוסיף לפרויקט. יש לכלול הסבר קצר בקובץ answers.pdf
- נא לתעד את כל קבצי הקוד בצורה סבירה

בדיקת הפרויקט

הפרויקט יבדק באופן אוטומטי על 30 שאלות באנגלית. לכל שאלה נחשב ציוני NDCG@10, Precision, Recall, F, עבור המסמכים שהוחזרו ודירוגם לעומת תוצאות האמת.
לפני ההגשה, מומלץ מאוד להריץ את המערכת שכתבתם על מאגר Cystic Fibrosis Database ועל כמה מהשאלות שבו. מאחר ודירוגי המסמכים ביחס לכל שאלה נתונים, אנו ממליצים לכתוב פונקציות שערוך לביצועי המערכת שלכם. ודאו שהמערכת שלכם מחזירה מסמכים הגיוניים בהינתן השאלה ושהמסמכים אכן מדורגים.

תזכורת למדדי השערוך:

- $NDCG@10 = \frac{DCG@10}{IDCG@10}$
- $Precision = \frac{|\{\text{retrieved documents}\} \cap \{\text{relevant documents}\}|}{|\{\text{retrieved documents}\}|}$
- $Recall = \frac{|\{\text{retrieved documents}\} \cap \{\text{relevant documents}\}|}{|\{\text{relevant documents}\}|}$
- $F = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)}$

קלטי הרצה לדוגמה:

- `python vsm_ir.py query bm25 /mydir/vsm_inverted_index.json "What factors are responsible for the appearance of mucoid strains of Pseudomonas aeruginosa in CF patients?"`
- `python vsm_ir.py query tfidf /mydir/vsm_inverted_index.json "What is the prognosis for survival of patients with CF?"`
- `python vsm_ir.py query tfidf /mydir/vsm_inverted_index.json "Are there abnormalities of taste in CF patients?"`

<u>הערה בנוגע לבדיקה:</u>

הקובץ cfquery.xml מכיל 99 שאלות בצירוף שמות המסמכים הרלוונטים לכל שאלה ודירוגם.

כפי שצוין, ניתן להשתמש בשאלות ובדירוגי המסמכים לצורך שערוך הביצועים ה-Vector Space Model שכתבתם. אסור לעשות כל שימוש בדירוגי המסמכים לצורך שאינו performance evaluation. למשל, קוד שלא יממש VSM או קוד שיבצע hardcoding משאלות למסמכים בהתבסס על דירוגי המאגר <u>יקבל ציון 0</u>.

בהצלחה!