אוניברסיטת בן-גוריון בנגב
הפקולטה להנדסה
המחלקה להנדסת מערכות מידע
אחזור מידע 2019-2020

צוות הקורס: ד"ר ניר גרינברג
רועי דור, מקס שפירו ויפים פייטרברג

פרויקט תכנות – חלק ב'

בניית מנוע לאחזור מסמכים

באתר הקורס תמצאו:

- **קובץ queries topics עם 15 שאילתות** – שאילתות עליהן המנוע צריך לענות. שימו לב כי לכל שאילתה יש מספר מזהה בן שלוש ספרות שנמצא בתגית <num>. השאילתה שאתה תעבדו לצורך העבודה עצמה נמצאת בתגית <title>.
- **קובץ qrels לשימוש ב TREC-EVAL** - הקובץ מכיל הערכות רלוונטיות (Relevance judgment) כלומר לכל זוג של שאילתה ומסמך הערכה האם המסמך רלוונטי לשאילתה. 1- רלוונטי 0 - לא רלוונטי.
- **תוכנת הערכה TREC EVAL** – משמשת לחישוב מדדים לצורך הערכת ביצועי המנוע.
- **הוראות שימוש ל-TREC EVAL**.
- **קובץ clickstream -** קובץ טקסט המכיל לוג של חיפושים שבוצעו ע"י משתמשים והמסמכים שהם לחצו עליהם בתוצאות החיפוש. הרשומות בפורמט:

  user_id, doc_id, query

  הרשומות בקובץ מייצגות שאילתות (query) שמשתמשים (user_id) ביצעו והמסמכים (doc_id) מתוצאות החיפוש שהם לחצו עליהם. השורה הראשונה בקובץ היא כותרות לעמודות. הלוג מכיל 1000 רשומות וקיימים בו 100 משתמשים. כל שורה היא רשומה ובכל שורה המפריד הוא ',' (פסיק).

- **הערה- במידה ואתם מעוניינים כעת להוסיף או לשנות דברים שפתחתם בחלק א' של המנוע אתם רשאים לעשות זאת ולתעד בדוח שתגישו.**

עליכם לממש את המחלקות הבאות:

מחלקת Searcher

תפקידה לבצע את השאילתות. המחלקה תקבל שאילתה (מילה או אוסף של מילים עם רווחים ביניהם), המחלקה תנתח את השאילתה בהתאם לניתוח הטקסט שנעשה על המסמכים ותחזיר את המסמכים הרלוונטיים ביותר לשאילתה באופן מדורג (באמצעות שימוש במחלקת Ranker המפורטת בהמשך). השאילתות שאתם אמורים לעבוד אתם הם

- מספר המסמכים המוחזרים לשאילתה מוגבל ב-50 (כלומר יש להחזיר את 50 מסמכים הרלוונטיים ביותר לפי הדירוג רלוונטיות).
- הערה - בקובץ ה-qrels נמצאות התשובות הנכונות לשאילתות– המסמכים הרלוונטיים – כדי שתוכלו לבחון את המסמכים שהמנוע החזיר כרלוונטי ולכוונן את הביצועים של המנוע בהתאם

כאשר מוחזר מסמך, יש לאפשר למשתמש אופציה של "זיהוי יישויות" במסמך, שתחזיר את 5 הישויות הדומיננטיות ביותר במסמך מדורגות על פי החשיבות שלהן. "ישות" מוגדרת כביטוי שנשמר כאותיות גדולות בלבד (כפי שהוגדר בסעיף "אותיות גדולות/קטנות" ו"שמות וישויות" במחלקת Parse בחלק א')

- הערה: אם המסמך **מכיל פחות מ 5** ישויות שונות יש להחזיר את כל הישויות הקיימות במסמך.
- יש להפעיל נוסחה לחישוב הדומיננטיות של הישויות ולהציג למשתמש את הציון לפי האלגוריתם שלכם עבור כל ישות.

מחלקת Ranker

תפקידה לדרג את התשובות לשאילתות על פי נוסחת דירוג שאתם תפתחו. בחלק זה אתם רשאים להשתמש בכל אינפורמציה ששמרתם בinverted index או מהמסמך שאתם מוצאים לנכון.

חובה עליכם לממש את החלקים הבאים בכדי לדרג את המסמכים (עליכם להחליט לבד כיצד לשקלל כל חלק עבור הדירוג הסופי של מסמך, לדוגמא לתת משקלים שונים לכל רכיב או להוסיף רכיבים נוספים):

- BM25 - עליכם לממש נוסחה זו של מודל שפה לדירוג מסמכים. (במידה ויש נתונים חסרים הציבו אפס במקום, לגבי הפרמטרים- ניתן להריץ מספר ניסויים כדי לבחון אילו פרמטרים נותנים תוצאות טובות יותר).
- עליכם לממש או להשתמש באלגוריתם לטיפול סמנטי על פי בחירתכם, כלומר להבין את הקשר הסמנטי בין השאילתה למסמכים . יש להריץ את השאילתות עם ובלי טיפול סמנטי. יש להוסיף בממשק אפשרות להרצה עם ובלי טיפול סמנטי, יש לדווח על התוצאות עם ובלי טיפול סמנטי.
- בנוסף, עליכם לאפשר לבצע חיפוש תוך שימוש במידע ב-clickstream data. עליכם לחשוב איך ניתן להשתמש במידע זה, ולהפעיל יצירתיות באופן השימוש במידע זה. נציין כי אנו מחפשים לראות את צורת החשיבה והמימוש, תוך שימוש בכלים וקונספטים שלמדנו במהלך בסמסטר ולאו דווקא שיפור בתוצאות חיפוש.

מחלקת GUI

יש **להוסיף** למחלקה שפותחה בחלק הראשון של הפרויקט את היכולות הבאות:

א. שתי אפשריות הכנסה לשאילתה (שאילתה הנה אוסף של מילים המופרדות ברווחים) -

1. הכנסה של שאילתה בודדת בחלון חיפוש (כולל מקום להכנסה של שאילתה וכפתור RUN להרצת השאילתה).
2. בחירה של קובץ שאילתות (טקסט) באמצעות כפתור חיפוש בשם "Browse" במערכת הקבצים. פורמט הקובץ שיבחר יהיה זהה לקובץ ה-queries שניתן לכם. השאילתות ירוצו אחת אחרי השנייה לפי הסדר שלהן בקובץ.

- בקובץ ה queries יש תגיות שונות נוספות, במידה ואתם מעוניינים להשתמש במידע נוסף שנתון על שאילתה (לדוג' מידע שמצוי בתגית < desc>) אתם רשאים, תארו במה השתמשתם וכיצד זה תרם לשיפור תוצאות בדוח.
- לידיעתכם, במעמד ההגנה על המנוע תידרשו להריץ על שאילתות חדשות. אתם רשאים להניח שבקובץ ה queries יהיה תמיד באותו פורמט עם אותן תגיות.

ב. יש להוסיף פונקציונאליות של "חיפוש ישויות" - כפי שהוגדר לעיל

ג. יש להוסיף פונקציונאליות של הוספת טיפול סמנטי בשאילתה - כפי שהוגדר לעיל

ד. הממשק יציג את התוצאות עבור השאילתה (כמות המסמכים הרלוונטיים שנמצאו ואת מספר הזיהוי (המופיע תחת התגית <DOCNO> של כל אחד מהם).

ה. בנוסף, יש לאפשר שמירה של כל התוצאות בקובץ אחד. הקובץ צריך להיות בפורמט שיאפשר הערכת התוצאות באמצעות תוכנת TREC_EVAL. יש לאפשר לשמור את הקובץ באמצעות folderDialog כך שיהיה אפשר לבחור נתיב, את שם הקובץ אתם רשאים לבחור כרצונכם. הקובץ יכיל את כל התוצאות של השאילתות, כאשר ניתן להבין איזו תוצאה שייכת לאיזו שאילתה עפ"י ה-ID של השאילתה. במידה ומדובר בשאילתה שהרצנו דרך חלון החיפוש ניתן לתת מספר ID רנדומלי.

ו. אם לוחצים על לחצן איפוס לפני שאנדקסנו את המאגר (כשלא קיימים קבצי posting) או לחילופין כשלוחצים פעמיים על כפתור האיפוס יש להציג הודעה מתאימה שאין מה למחוק.

ז. **אנא הימנעו מנפילה של התוכנה והציגו הודעות בהתאם.**

ח. **לידיעתכם, אין קריטריון של הערכה לעבודה שמתחשב באסתטיקה או העיצוב של הממשק.** אין טעם להשקיע בזה. הממשק צריך להיות פונקציונאלי, לספק את מה שצריך **ולהיות יציב.**

ט. יש לאפשר למשתמש לבצע חיפוש תוך שימוש ב-clickstream data או בלעדיו..

הערכה

יש להריץ את המנוע על השאילתות שהוגדרו לכם. את התוצאות יש לשמור בקובץ תוצאות אחד מרוכז. יש לבצע שתי סדרות של הרצות: אחת עם stemming ואחת ללא stemming .

כל סדרת הרצה תשמר בקובץ נפרד. יש להשתמש בקובץ qrels ובתוכנת TREC_Eval כדי לחשב

- Precision ו- Recall כוללים לכל שאילתה
- Precision ו- Recall כוללים למנוע
- Precision ב 5, 15,10 , 30 ו-50 (Precision@N) מסמכים לכל שאילתה ובממוצע למנוע
- MAP למנוע

הנחיות הגשה

יש להגיש לספריית ה-ftp (הכתובת תפורסם בהמשך) את כלל הפרויקט והדוח והמסכם כתיקייה אחת מכווצת. לפי שמות תז המגישים 123456789_123456789. **בנוסף, את קבצי הקוד יש להגיש למערכת ההגשות המחלקתית תחת לשונית הקורס - חלק ב'**. אתר המערכת: https://subsys.ise.bgu.ac.il/submission/login.aspx. יסופקו בדיקות בסיסיות לקוד במערכת וזאת על מנת לאפשר בדיקה שלכם כי הקוד אכן עובד ללא שגיאות. עם זאת, יודגש כי אלו בדיקות בסיסיות ובאחריותכם לוודא כי הקוד שאתם מגישים ניתן להרצה תקינה על כל חלקיו.

בתיקייה יש לשים את הקבצים הבאים:

1. חבילת הפרויקט (קוד מקור מתועד) כולל קובץ מוכן להרצה
2. קובץ הוראות הפעלה (Readme).
3. את קבצי ה-posting שהורצו על ה-data שנמצא באתר.
4. **יש להגיש דו"ח (קובץ Word) על פי הפירוט הבא :**
   a. הסבר מפורט על אופן פעולת המנוע - אם הוספתם מחלקות יש להסביר את מטרתן ואיך הן פועלות.
   b. יש לכלול הסבר מפורט של כל המחלקות הרלוונטיות לחלק זה (אין צורך להסביר על מחלקות שבניתם בחלק א' אך לא ביצעתם בהן שינויים במסגרת ביצוע חלק זה).
   c. הסבירו בצורה מפורטת את האלגוריתמים הכלולים במנוע. בפרט:
      i. אלגוריתם הדירוג
      ii. אלגוריתם למציאת 5 הישויות הדומיננטיות במסמך, כולל 2 דוגמאות
      iii. אלגוריתם לשיפור סמנטי
      iv. אלגוריתם ואופן שימוש ב-clickstream data.

   פרטו את הרכיבים השונים של כל אלגוריתם/נוסחא ,את ההצדקה לבחירה שעשיתם, למשל, כיצד קבעתם משקולות ולמה, והוסיפו דוגמא להשפעה של כל אחד מהאלגוריתמים.
   d. יש להסביר על הנתונים בקובצי ה-posting ובמילון התומכים באלגוריתמים שמימשתם.
   e. אם השתמשתם במהלך העבודה בקוד פתוח לפרט את השירות, כתובת, היכן השתמשתם, כיצד השתמשתם.

2. **הערכה של המנוע**- ניתוח הפלט של תכנת TREC EVAL (פלט לאחזור עם stemming ופלט לאחזור בלי stemming) לשאילתות שפורסמו לכם באתר לחישוב המדדים כמפורט בסעיף "הערכה". הפלט צריך להכיל שתי טבלאות, אחת עבור אחזור עם stemming ואחת בעבור אחזור בלי stemming. על כל אחת מן הטבלאות הבאות להכיל את העמודות הבאות:
   1. מספר השאילתה
   2. מילות השאילתה
   3. Precision לשאילתה
   4. Recall לשאילתה
   5. precision@5
   6. precision@15
   7. precision@30
   8. precision@50
   9. משך הזמן להרצת השאילתה.

כמו כן, יש להציג בסוף כל טבלה את מדד ה-map (כפי שלמדתם מדובר בממוצע של ה-precision על פני כל השאילתות – מדד זה נמצא בפלט של ה- TREC EVAL תחת הכותרת average precision over all rel docs).

3. **סיכום**: בעיות שנתקלתם בהם וכיצד התמודדתם איתן. מה האתגר הגדול לדעתכם בפרויקט. המלצות לשיפור האלגוריתם שלכם/מה הייתם עושים אחרת..?

- אין צורך לצרף את הפלטים הגולמיים של ה- TREC EVAL.
- אין צורך לצרף את הפלטים של השאילתות הנשמרים על הדיסק (קובץ עם המסמכים הרלוונטיים והדירוג שלהם)
- בהגשת המנוע יש לצרף את קבצי ה-index שלכם (קבצי posting והמילון) על מנת לחסוך זמן בבדיקות הפרונטליות. פרט לצירופם, אנא ודאו שהטעינה מהקבצים הללו ישירות לזיכרון עובדת כראוי על כל מחשב.

אנא וודאו לפני הגשה: שהגשתם את הדוח במלואו, שהקוד מתועד, עובד ולא נופל, זרקו שגיאות בהתאם. וודאו שהקוד עובד במעבדה. אי מילוי של הוראות אלו יפגע בציונכם בצורה משמעותית.

בהצלחה!