פרוייקט במבני נתונים ואלגוריתמים

מדמה מנוע חיפוש דוגמת Google Search

הוראות:

- קראו את המסמך עד סופו.
- יש לבצע את עבודה בשלשות בלבד.
- תאריך הגשה עד מועד א. לא תינתן הארכה.
- מומלץ לבצע את המשימות קרוב ככול האפשר לסיום למידת הנושא המתאים בכיתה.
- במידה והתבקשתם לממש אלגורתמים או מבנה נתונים כלשהם, אסור להשתמש בפונקציות מוכנות. ניתן רק להשוואת אותם למימוש שלכם.
- בחלק מהשאלות אתם מתבקשים להציג תוצאות, חישבו מהי הדרך הטובה ביותר להציגם. יש להעדיף ויזואליזציה של התוצאות.
- עבודות מועתקות או חשודות למועתקות- תיפסלנה!!!

הקדמה:

השימוש במבני נתונים ואלגורתמים.

מטרת הפרוייקט היא להדגים (באופן כללי) כיצד עובד האלגורתמים אליו מבוסס מנוע החיפוש של גוגל, בשם Page Rank. בפרוייקט זה נממש את האלגורתמים באופן חלקי, נפרק את הבעיה למספר תתי בעיות בהתאם לחומר הנלמד בקורס.

הערות:

1. פרוייקט זה מיועד להמחיש את הרעיון מאחורי מנועי חיפוש. בפעול <u>לא תתבקשו</u> ליצור מנוע חיפוש מקצה לקצה, עם זאת קבוצות אשרו יממשו מערכת שלמה מקצה לקצה יזכו לבונוס בהתאם לאיכות עבודותם (עד 25 נקודות נוספות לציון העבודה, ניתן לעבור את ה 100) למעוניינים כדאי לחפש לגבי Page Rank ולגבי Web Crawler.
2. לפעמים הקושי בהתאמת מבני נתונים או אלגורתמים נובע בעצם הגדרת הבעיה, לכן ההוראות יהיו לפעמים כלליות. בכול מקרה העבודה אינה נבדקת ע"י תוכנה אוטמטית, לכן ניתן ואפילו ממולץ לגשת לבעיות בפרספקטיבה יצרתית.
3. קושי נוסף במציאות של חקר אלגורתמים הוא מציאה/יצירה של בסיס נתונים לנסות עליו את האלגורתמים. בפרוייקט תאלצו להתמודד גם כן אם השגת בסיס הנתונים בעצמכם.

להזכירכם מכיוון שאינכם נדרשים לממש מערכת מקצה לקצה אתם יכולים לעבוד עם בסיס

נתונים שונה לכול מכלול במערכת.

4. לפעמים בחירת הפתרון המתאים לבעיה היא בעיה אמפרית ולא אנלטית. כלומר, לפעמים יש לבצע בחירה של פרמטרים באלגורתמים בשביל להשיג ביצועים אופטימלים, אשר נובעת מבדיקה נסיונית ולא בעקבות סיבה הניתנת להוכחה. במקרים כאלו ניתן לקחת את בסיס הנתונים ולחלקו **אקראית** לשתי קבוצות 80%-90% מבסיס הנתונים לקבוצה לימוד/בחירת הפרמטרים ואת שאר הנתונים לצורך בדיקת האלגורתמים, כמובן שניתן לחזור על תהליך זה מספר פעמים בשביל להימנע ממקריות.
5. ניתן ואף רצוי לחפש במקורות חיצונים כיצד לבצע ולשכלל את השיטות הדרושות.
6. אנו מעודדים את קבוצות הסטודנטים להתייעץ בניהם, אך אין לשתף קוד.
7. עבור כל שאלה בפרוייקט יש לתת הסבר קצר וממוקד, של המימוש שביצעתם ולהציג תוצאות לאיכות עבודתכם. במידה ונוסו מספר שיטות לפתרון הבעיה יש לבצע השוואה מסודרת בניהם ולציין מתי ולמה יבחר כל פתרון.

מכלולי המערכת:

- ספירת מספר המילים המופיעות במסמך, ניתן למימוש לאחר תרגול Python.
- בניית ותחזוקת מבנה נתונים למיפוי מילה למסמכים בהם מופיע המילה, ניתן למימוש לאחר תרגולי Red Black Trees & Hash Tables.
- דחיסה ופריסה ללא איבוד מידע, על מנת לחסוך במקום בזכרון, ניתן למימוש לאחר תרגול Dynamic Programming.
- דירוג רלוונטיות של האתר, ניתן למימוש לאחר תרגולי Graphs.
- יכולת למיין באופן מהיר את האתרים לפי מאפיינים שונים : (מידת רלוונטיות, שפה או זמן כתיבה), ניתן למימוש לאחר תרגולי Sort.

תיאור פעולות המערכת:

נרצה "לזחול" (Crawl) ברשת ולמצוא מיפוי בין מילים לבין עמודי רשת בהם מופיע המילה. כלומר, בהינתן עמוד אינטרנט נפרק את הטקסט שבעמוד למילים ונוציא כפלט ווקטור של המילים המופיעות בו.

לאחר מכן, נרצה להשתמש במבני נתונים כך שבמידה ונחפש מילה מסויימת, מבני הנתונים יחזיר לנו את כתובות העמודים בהם מופיעה המילה המבוקשת.

ה"קסם" והפופלריות של מנוע החיפוש של גוגל נובע בעיקר מיכולתו למיין את העמודים לפי מידת הרלוטיות של עמוד האינטרנט. ניתן לייצג את עמודי האינטרנט והקשרים בניהם בעזרת גרפים, לאחר יצוג דפי האיטרנט כגרפים ניתן להשתמש בטכניקות מתורת הגרפים כדי לשערך את מידת הרלוונטיות של העמוד.

לאחר שנשלפו העמודים הרצויים, ניתן למיין את הערוצים לפי מספר מאפיינים כפי שצוינו קודם, בהתאם לרצון המשתמש.

<u>חלק א: ספירת מספר המילים המופיעות במסמך</u>. (ניתן לביצוע לאחר תרגול Python)

(15 נקודות)

א) בחלק זה תתבקשו לממש פונקציה שתקבל כקלט קובץ TXT ופלט הפונקציה יהיו המילים המופיעות במסמך, ובנוסף את מספר המופעים של כל מילה במסמך.

ב) הוסיפו לפונקציה משתנה קלט נוסף, רשימה של מילים.

במידה וניתן קלט זה לפונקציה לספור במסמך רק את המלים המופיעים ברשימת הקלט.

ג) ממשו פונקציה שבהינתן מספר מסמכים מפיקה כפלט רשימה של מילים איתן כדאי לתייג (למפות) את המסמכים. כדי לבדוק את הפונקציה כדאי להוציא כפלט מילים אשר "נפסלו".

ד) ציינו את הסיבוכיות של מימושכם.

זכרו להדגים את איכות עבודתכם.

<u>חלק ב: מיון סדר הופעת האתרים לפי מאפיינים שונים.</u> (ניתן לביצוע לאחר תרגולי Sort)
(15 נקודות)

הערה: בחלק זה לא ניתן להשתמש בפונקציות מיון מוכנות.

א) ממשו פונקציה אשר בהינתן רשימה/מערך של אתרים תחזיר רשימה/מערך ממוינית לפי שפת הטקסט (הניחו כי שפת המסמך נתונה). באיזה סוג מיון כדאי להשתמש ומדוע?
ב) ממשו פונקציה אשר בהינתן רשימה/מערך של אתרים תחזיר רשימה/מערך ממוינית לפי תאריך כתיבת הטקסט (הניחו כי תאריך העלאת הטקסט נתון). באיזה סוג מיון כדאי להשתמש ומדוע?
ג) בסעיף זה אתם מתבקשים לממש פונקציה למיון הרשימה/מערך לפי קריטריון הרלוונטיות של עמוד האינטרנט. אך מכיוון שעדיין לא מימשתם את האלגורתמים למדידת הרלוונטיות של העמוד. אתם מתבקשים לכותב פונקציה/סקריפט אשר ישווה בין לפחות בין שנ יאלגורתמי מיון ותבחר את המיון המועדף. שימו לב ובמידה ואתם משתמשים באלגורתמים מיון אשר משתמש בפרמטרים אליכם לקבוע אותם כפי שצויין בהערות קודם.
ד) ציינו מה הסיבוכיות של מימושכם.

זכרו להדגים את איכות העבודה של מימושכם.

<u>חלק ג: מימוש עץ אדום שחור, לפי מפתח מילים.</u> (ניתן לביצוע לאחר תרגולי Trees)
(15 נקודות)

הערה: אין להשתמש במימושים מוכנים ליצירת עצים.

א) ממשו עץ אדום שחור לאגירת עמודי האינטרנט, כאשר מפתח המיון הוא המילים (אשר נבחרו לשימוש) המופיעים בעמודי האינטרנט. בחרו כיצד להתמודד עם כפילויות.
ממשו את כול הפעולות אשר עץ אדום שחור תומך בהם, כולל מחיקה.
ב) הדגימו את האלגורתמים אשר מימשתם.
ג) ציינו מה הסיבוכיות של מימושכם.

חלק ד: אגירת עמודי אינטרנט במבנה נתונים Hash Table. (ניתן למימוש לאחר תרגול Hash table)

(15 נקודות)

הערה: אין להשתמש במימושים מוכנים דוגמת dict.

ממשו מבנה נתונים מסוג Hash Table לאגירת עמודי האינטרנט בדומה לסעיף קודם.

בחרו פונקציות ערבול מתאימה, והסבירו.

ציינו מה הסיבוכיות של מימושכם.

<u>חלק ה: השוואה בין עץ אדום שחור לבין Hash Table.</u> (ניתן למימוש לאחר תרגולי Hash tables ו Trees)

(10 נקודות)

השוו בין המימושים שלכם בסעיף ג וד.

כיצד יבחרו אתרים אפשריים בהינתן שאילתה בת מילה אחת או יותר (ניתן להניח כי אין שגיאות כתיב)?

העזרו בהסבר מילולי וגרפים.

<u>חלק ו: דחיסת טקסט.</u> (ניתן למימוש לאחר תרגול Dynamic Programming)

(10 נקודות)

אגירה של כול עמודי האינטרנט, המסמכים והמילים דורשת זכרון רב. ברצוננו לצמצם את כמות הזכרון הדרושה לאגירת טקסטים.

א) ממשו אלגורתמים דחיסה/פריסה ללא איבוד מידע כפי שנלמד בכיתה.

ב) ממשו אלגורתמים דחיסה/פריסה ללא איבוד מידע כאשר רק חלק העמודי האינטרנט גלויים מראש לדוחס.

ג) ערכו השוואה בין ללא דחיסה ודחיסה לפי סעיף א וב.

חלק ז: חישוב דירוג עמודי האינטרנט. (ניתן למימוש לאחר תרגול Graphs)

(20 נקודות)

בחלק זה אתם מתבקשים לממש אלגורתמים לחישוב ציון הרלוונטיות של עמודי האינטרנט.

מספר רמזים:

- ערכו של אתר עולה כול שיש יותר קישורים אליו מאתרים אחרים.
- ערך ההפניה של אתר פרופרציונלי לדירוג של האתר עצמו (המפנה). כלומר ככול שלאתר המפנה לאתר אחר יש דירוג גבוהה יותר בפני עצמו, האתר אליו מופנים מקבל ערך גבוהה יותר. (אל תשכחו נרמול)
- ניתן להניח כי הגרף יתכנס לאחר מספר סופי של איטרציות בגרף. כיצד תתמודדו עם מעגלים?

האלגורתמים של גוגל עושה שימוש בעקרונות אשר צויינו.

1) ממשו אלגורתמים למתן ציון לאתרי אינטרנט לפי העקרונות הללו. מומלץ כמובן להיעזר במאמרים לצורך מימוש האלגורתמים.

2) הסבירו והדגימו את מימושכם.

3) ציינו את הסיבוכיות של מימושכם.

בהצלחה

מקווה כי נהנתם והפקתם ערך נוסף

איתן