ניהול נתונים באינטרנט - תרגיל מס' 2 - Searching and Ranking

הנחיות הגשה

ההגשה תתבצע בזוגות, דרך אתר moodle, שם גם מוגדר מועד ההגשה. יש להגיש את הפתרונות בקובץ zip יחיד בשם username_wdm.zip (למשל, daniel2_wdm.zip) שיכיל את answers.pdf ואת קבצי הקוד. רק אחד מבני הזוג יגיש את המטלה (ניתן להשתמש ב-username שלו\ה), אבל הקפידו לציין בראש הקובץ answers.pdf שמות ות.ז. של שני בני הזוג.

שאלה 1

א. כתבו מחלקת WebGraph שמנהלת את מבנה הנתונים "גרף הלינקים" בין דפים שונים.

ב. כתבו מחלקת InvertedIndex שמנהלת ברשימה ממוינת לקסיקוגרפית של מילים, כאשר עבור כל מילה נשמרת רשימה ממוינת של מזהי דפים (url) ולכל אחד מהם ציון: מס' המופעים של המילה מחולק באורך הדף (מס' מילים).
נגדיר מילה כרצף תווים מופרד ב- whitespaces מרצפים אחרים. ניתן לשמור כל מילה בדיוק כפי שהופיעה (אין חובה לבצע טיפול בתווים מיוחדים, case וכו', אך ניתן לעשות כן - וצפוי שיפור בתוצאות)

ג. כתבו מחלקה Search עם מתודת crawl. המתודה תבצע crawling על דפי ויקיפדיה (בלבד). היא מתחילה מכתובת ה- url
http://simple.wikipedia.org/wiki/Albert_einstein

וע"י מעקב אחר לינקים (מהצורה `href="/wiki/xxx"` בלבד; כדי לבצע crawling הוסיפו http://simple.wikipedia.org בתחילת הלינק),
הפונקציה תמלא את מבני הנתונים שכתבתם בסעיפים הקודמים במידע עבור 100 דפים שונים (לפחות, לבחירתכם). יש לשמור את **כל** הקישורים הקיימים בין דפים בקבוצת 100 הדפים הללו.

היעזרו בתיעוד המחלקה URL
http://docs.oracle.com/javase/7/docs/api/java/net/URL.html

ובדוגמא (למשל):

http://stackoverflow.com/questions/238547/how-do-you-programmatically-download-a-webpage-in-java

שימו לב שמרגע קבלת ה- InputStream, תוכלו לעבוד איתו כרגיל.

שימו לב: בחרו **אחת** מבין שאלות 2,3.

שאלה 2

ממשו אלגוריתם לחישוב PageRank של הדפים, במתודה בשם pageRank בתוך מחלקת Search. הקלט הוא אובייקט מטיפוס WebGraph, וערך damping factor. הפלט הוא רשימה ממוינת לפי rank (מבנה הנתונים המדויק נתון לבחירתכם) של זוגות כאשר כל זוג הוא (ID,rank). ID הוא מזהה המסמך (url), ו- rank הוא ציון ה- PageRank שלו.

כזכור באלגוריתם איטרטיבי עשויה להתקבל תשובה שאינה מדויקת - רמת הדיוק בה תשתמשו כתנאי עצירה נתונה לשיקולכם (נסו ערכים שונים).
הסבירו בקצרה את תהליך הבחירה של תנאי העצירה (כהערה בקוד).

שאלה 3

ממשו את אלגוריתם HITS במתודה בשם hits בתוך מחלקת Search. הקלט הוא שוב אובייקט מטיפוס WebGraph, והפלט הוא שוב רשימה ממוינת של זוגות (ID,rank), אך הפעם rank הוא ציון ה- authority של הדף (ציוני ה- hubness משמשים תוך כדי ריצת האלגוריתם אך אינם מופיעים בפלט).

כזכור באלגוריתם איטרטיבי עשויה להתקבל תשובה שאינה מדויקת - רמת הדיוק בה תשתמשו כתנאי עצירה נתונה לשיקולכם (נסו ערכים שונים).
הסבירו בקצרה את תהליך הבחירה של תנאי העצירה (כהערה בקוד).

שאלה 4

ממשו את אלגוריתם TA (threshold algorithm) במתודה בשם TA בתוך מחלקת Search. הקלט של המתודה הוא רשימה **(באורך כלשהו)** שמכילה רשימות ממוינות של זוגות (ID,rank) (באותו מבנה נתונים שבחרתם עבור שאלות 2,3). פונקציית האגרגציה יכולה להיות hard-coded: בחרו אותה באופן שמתאים לדעתכם לאגרגציה של ציונים לפי מילים בודדות + ציון לפי ה- WebGraph. **הסבירו את בחירתכם כהערה בקוד.**

שאלה 5

כתבו את פונקצית main של המחלקה Search, שמשתמשת בפונקציות שכתבתם בשאלות הקודמות כדי לבצע חיפוש פשוט. תחילה, הפונקציה תריץ את crawl כדי למלא את מבני הנתונים. היא תדפיס לקובץ urls.txt את רשימת הדפים שבמבנה הנתונים. לאחר מכן, הפונקציה תריץ את אלגוריתם הדירוג שבחרתם לממש (HITS או PageRank), ותדפיס את 5 הדפים הטובים ביותר (לפי סדר) לפי האלגוריתם לקובץ rank.txt (אם מימשתם את HITS המיון הוא רק לפי authority).

לאחר שלב זה, התוכנית תרוץ בלולאה, בכל איטרציה היא תקבל כקלט מ-System.in שורה של מילות חיפוש (מופרדות ברווח) ותדפיס למסך רשימה ממוינת של חמש התוצאות הגבוהות ביותר, לפי אגרגציה בין הציונים לפי מילות החיפוש וציון ה- PageRank או HITS. חשבו את הרשימה תוך שימוש ב- TA. בקבלת השורה exit התוכנית תצא.

הריצו את התוכנית עם מחרוזות חיפוש לדוגמא (3 דוגמאות שונות לפחות) והעתיקו את תוצאות ההרצה מכל קבצי הפלט ומהפלט הסטנדרטי, ביחד עם הקלט שהוביל אליהם, לקובץ **answers.pdf**