ניהול נתונים באינטרנט – תרגיל מס'2 – Crawling & Ranking

הוראות:

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, ורק 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 אחד מהמגישים (למשל: HW1_123.zip).
תאריך הגשה: **15.04.2018**

שאלה 1:

(א) ממשו תוכנית CRAWLER בשפת PYTHON עבור אתר ויקיפדיה, שמקבלת כתובת URL של עמוד התחלתי, ובונה גרף של 10 הצבעות **<u>הפנימיות</u>** הראשונות בין כל העמודים, כלומר לינקים שמתחילים ב "****/wiki/". על ה CRAWLER להקפיד לא לבקר באותו עמוד פעמיים, כמו כן, יש לעצור אחרי עומק 3 צעדים מהעמוד המקורי, כאשר באיטרציה האחרונה יש לשמור רק את הלינקים לעמודים שכבר ביקרנו בהם (על מנת למנוע מצב של המון קודקודים ללא הצבעות החוצה).
יש להדפיס את הגרף בצורה של רשימות שכינות. לדוגמא:

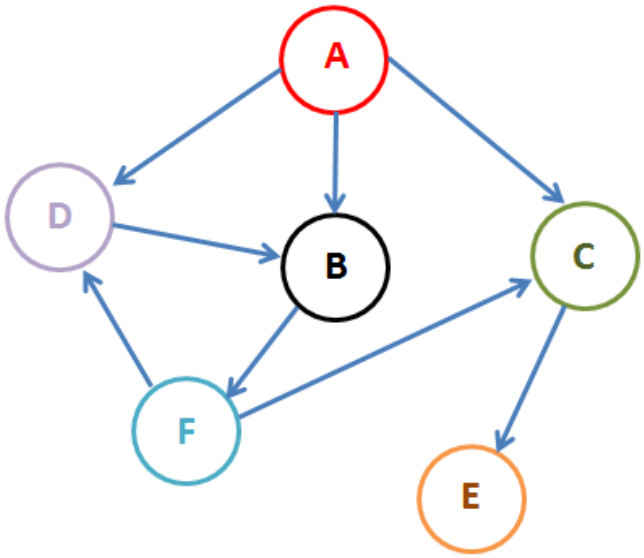
SiteA = {SiteB, SiteC, SiteD}
SiteB = {SiteA, SiteD}
SiteC = {SiteA, SiteD}

הגישו את התשובות בקובץ question1a.py

(ב) ממשו אלגוריתם PAGERANK, המקבל גרף מסעיף (א) ומחזיר את ערכי ה PR של העמודים. יש להשתמש ב DAMPING FACTOR של 0.3. על התוכנית להדפיס את ה PR ואת ה URL של כל אחד מהעמודים.
הגישו את התשובות בקובץ question1b.py

(ג) הריצו את התוכניות שפיתחתם בסעיפים (א) ו (ב) על האתר הבא:
https://en.wikipedia.org/wiki/Kirill_Nababkin
והדפיסו את התוצאות לקובץ question1c.txt

שאלה 2:



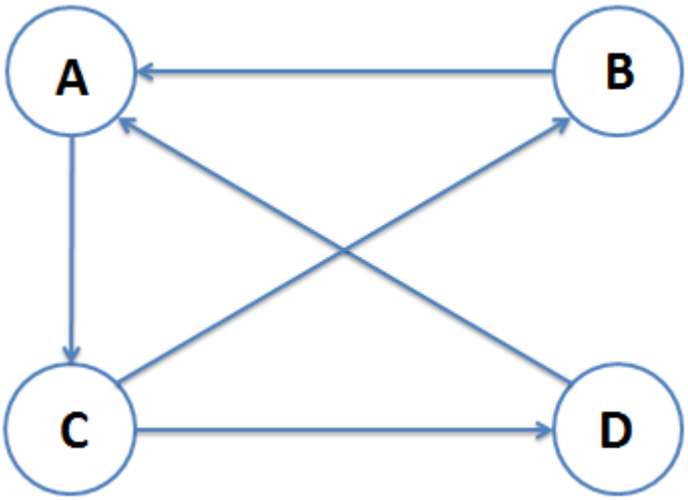
(א) נתון גרף של לינקים בין האתרים A, B, C, D, E, F. חשבו את ה PAGERANK של כל אחד מהאתרים ללא DAMPING FACTOR והסבירו את התוצאה שקיבלתם. הגישו את החישובים שלכם ואת ההסבר בתוך הקובץ answers.pdf

(ב) חשבו את ה PAGERANK של עם DAMPING FACTOR של 0.1 והגישו את החישובים בקובץ answers.pdf

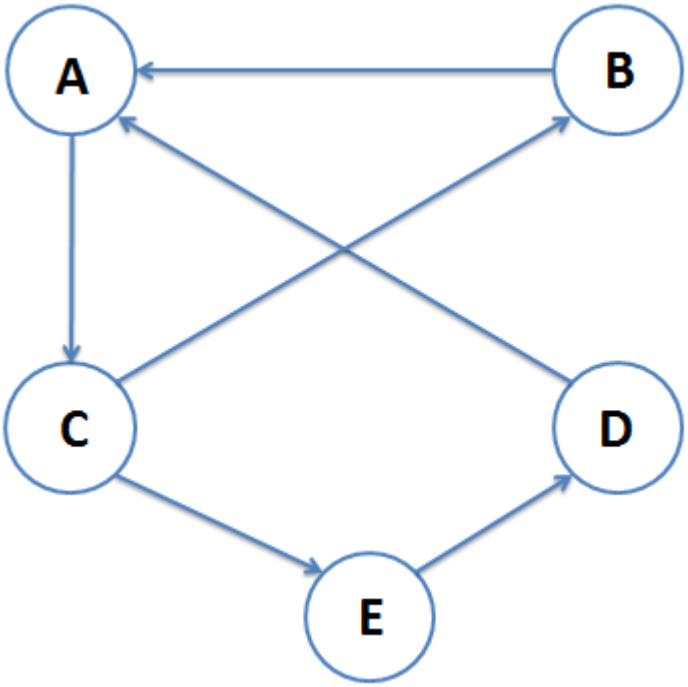
(ג) הסבירו את השוני בתוצאות של סעיף (א) וסעיף (ב) והגישו את ההסברים בקובץ answers.pdf

שאלה 3:

נתון גרף המייצג ארבעה דפי אינטרנט והקישורים ביניהם:



לאחר מכן, נוסף דף חדש (E), והקישורים התעדכנו. הגרף החדש מוצג להלן:



מה קרה לערכי ה PAGERANK של **כל אחד מהקודקודים** לאחר השינוי של מבנה הגרף? (הניחו שחישוב ה- PR נעשה תמיד ללא DAMPING FACTOR)

נמקו את תשובתכם והגישו את ההסבר בתוך הקובץ answers.pdf