אחזור מידע – תרגיל בית 1 – אביב התשפ"ד

יש להגיש למוודל עד לתאריך 5.7.24 בשעה 23:59

שאלה 1: מודלים לאחזור מידע

1. המודל הבוליאני:יש לענות על שאלות 4a+4b בתרגול 3.
2. tf-idf: נתונות הטבלאות הבאות:

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

► **Figure 6.8** Example of idf values. Here we give the idf's of terms with various frequencies in the Reuters collection of 806,791 documents.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

► **Figure 6.9** Table of tf values for Exercise 6.10.

יש לחשב את ערכי tf-idf עבור כל term, ועבור כל אחד מהמסמכים. יש להציג חישוב מפורט.

שאלה 2: בניית אתר HTML

עליכם לבנות אתר לסטודנטים בצוות, לפי המטלה אשר התחלתם לבצע בכיתה. האתר יכלול דף אישי לכל סטודנט, הכולל פרטי הסטודנט, רשימת תחביבים, טבלת קורסים, ומייל אישי להתקשרות. יש לכלול לפחות עוגן אחד (anchor) בכל דף.
כמו כן, האתר יכלול דף פתיחה אשר מפנה לשני הדפים של הסטודנטים. יש לכלול תפריט אשר מאפשר ניווט נוח בכל שלושת הדפים.
עליכם לפתוח ריפו קבוצתי ב -GIT, להגדירו כציבורי , ולהעלות אליו את האתר, לפי ההסבר על פתיחת דף גיט שנמצא בתיקיית תרגילי הבית.

שאלה 3: קדם פרויקט –חקר אתר

בחרו אתר אשר בו תרצו להתמקד. האתר צריך לכלול כמות מידע מספקת לצורך ניתוח ע"י זחלן רשת (בדרך כלל כמות של כמה אלפי רשומות תספיק).בהמשך הסמסטר, בפרויקט ,תבנו זחלן רשת (crawler)אשר ישלוף את המידע מהאתר שבחרתם. בפרויקט תנתחו את המידע ותציגו מסקנותיכם.

על זחלן רשת ניתן לקרוא כאן:

https://he.wikipedia.org/wiki/%D7%96%D7%97%D7%9C%D7%9F_%D7%A8%D7%A9%D7%AA

לאחר עיון באתר, ענו על השאלות הבאות:

1. מהו תחום העיסוק המרכזי של האתר? מהו המידע הזמין למשתמשי האתר? ענו בפסקה אחת. צרפו את הקישור לאתר.
2. רשמו שלוש שאילתות מעניינות שהייתם רוצים לקבל עליהן תשובה באתר, והאתר אינו **עונה עליהן** כעת. עבור כל שאילתא כזו, ציינו מהם פרטי המידע הנדרשים לצורך מענה על השאילתא.

יש להגיש קישור לריפו שלכם ב -GIT, הכולל את כל קבצי האתר, ותיקייה בשם 1HW שבה קובץ וורד ובו מענה על שאלות 1+3.ניתן להעלות שאלות הקשורות לתרגיל בפורום הקורס.