**קורס איחזור מידע**

**README עבור שיעורי בית מספר 5**

**שיש להגיש למערכת המודל עד 3/5/2016 לפני השיעור**

הקוץ הזה מסביר מה יש בתיקיית הנתונים. מסמך נפרד יפרט את המטלה

בתיקייה יש את הקבצים הבאים

1. טקסט גולמי של 1,398 מסמכים
2. מטריצה המייצג את הטקסט של המסמכים אבל במספרים. למטריצה יש 1398 טורים (אחד לכל מסמך) ו9135 שורות (אחת לכל מילה הקיימת במסמך כלשהו). הערך של משבצת טור i שורה j משמעותו מספר הפעמים שמילה מספר j מופיעה במסמך i. המטריצה ניתנת לכם בצורת קובץ טקסט. בכל שורה בקובץ יש שלושה מספרים -- משמאל לימין, מסמך (כלומר טור במטריצה) מילה )כלומר שורה במטריצה) והערך (כלומר מספר הפעמים שאותה מילה מופיעה באותו מסמך). הקובץ כולל רק ערכים שהם לא אפס, כלומר, הוא לא רושם אל כל הצילים שאינן מופיעות בכל מסמך, רק אלה שכן.
3. טקסט גולמי של 225 שאילתות
4. מטריצה המייצג את הטקסט של השאילתות אבל במספרים, דומה למה שתואר למעלה עבור מסמכים
5. מילון שרושם את המילים לפי הסדר, כלומר, אם מכל סיבה שהיא תרצו לדעת "מה זה בכלל צילה מספר 8724?", היא המילה בשורה 8724 של המילון
6. מיפוי המאפשר לדעת איזה מסמך גלמי הולך עם שורה מסויימת במטריצת המסמכים. לדוגמא אם מכל סיבה שהיא תרצו לדעת "איזה מסמך זה בכלל מסמך 624?" הוא המסמך בשורה 624 של הרשימה הזאת (הטור הראשון במטריצה באמת מייצג את המסמך עם השם cran1 אבל הטור השני מייצג לא את המסמך cran2 אלא את המסמך cran10 ובגלל העניין הזה אני נותן לכם את הרשימה לפי הסדר שהתוכנה קבעה
7. מיפוי דומה לשאילתות
8. QRELS שהם התשובה הנכונה כלומר עבור כל שאילתא יש רשימה של (סדר גודל) 10 מסמכים שהם רלוונטיים ושכדאי שמנוע החיפוש ידרג אותם גבוה. צריך את זה על מנת להעריך את הביצועים של מנוי החיפוש לפי מדד average precision כמו שתיארנו בשיעור