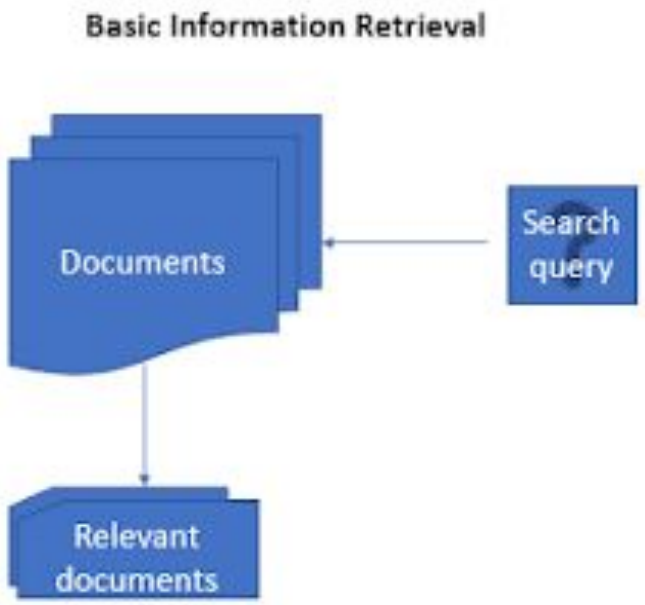
מבוא לאיחזור מידע

מעבדה 1

מבוא לאיחזור מידע

גילוי והצגה אוטומטיים של עובדות, חוקים, וקשרים החבויים בתוכן טקסטואלי.



מעבדה 1 - הכר את הנתונים

בשבועות הקרובים נעבוד עם נתונים שנקצרו מטוויטר.

הנתונים הינם טוויטים שעברו סיווג לדבר נאצה (תוכן גזעני או סקסיסטי)

הקובץ הינו קובץ CSV - עמודה ראשונה הינה ID, שניה הינה סיווג לדברי נאצה (מלל גזעני או סקסיסטי) , ושלישית היא הטקסט עצמו.

מעבדה 1 - חלק א - פייתון

כיתבו קוד בפייתון, מומלץ בעזרת שימוש ב pandas , על הקוד לנתח את קובץ הנתונים, וליצור קובץ חדש גם הוא מופרד בפסיקים.
עבור כל טוויט יתווספו הפרטים הבאים:

1. מספר מילים
2. מספר אותיות
3. גודל מילה ממוצע
4. מיספור של stopwords בעזרת שימוש בחבילה NLTK
5. מספר תווים מספריים
6. ספירת כמות של תווים מיוחדים יכולה לתת עוד מידע על אופי הטקסט. עבור טוויטים לדוגמא מס ה - # . הוסיפו עמודה עם מס ה - # ובמידה ויש עוד סימנים יחודיים שכדאי למספר הוסיפו גם אותם.
7. מספר מילים שנכתבו באותיות גדולות (ביטוי לכעס)