מבוא לאיחזור מידע

מעבדה 2 - יצירת מילון

המשך עיבוד מקדים ובניית מילון

המטרה - בניית מילון מצומצם ככל האפשר לכל המושגים הרלוונטיים לחיפוש.

- במעבדה זו יש להציג את ההשפעות השונות של תהליכי עיבוד הנובעים מהשפה עצמה - תיקון שגיאות, זיהוי שורשים, זיהוי מילים נרדפות וכו'.
- במהלך המעבדה יש ללמוד שלוש ספריות מרכזיות בפייתון, הקשורות לתחום (מפורט בהמשך).
- במהלך למידת הספריות מומלץ לצמצם את היקף הטוויטים עליהם עובדים, כדי לצמצם את זמני הריצה במהלך הניסויים השונים.

מטלת המעבדה

סכמו על כל אחד מהכלים במילים שלכם (יש להתייחס לנקודות הבאות):

- תכונות ויכולות כלליות (בקצרה)
- תכונות ויכולות הרלוונטיות לבניית המילון שלנו
- מבני נתונים שימושיים

1. NLTK
2. TextBlob
3. spaCy

מטלת המעבדה

יש להדגים כל אחת מהפעולות הבאות, בעזרת כל אחת מהספריות שניתנו, או להסביר למה לא ניתן. הדוגמא צריכה להיות בקנה מידה קטן, ניתן להדגים על נתונים מלאכותיים.

- טוקניזציה
- תיקון שגיאות
- למטיזציה (lemmatization)
- סטמינג (stemming)
- ראשי תיבות (acronym or abbreviations)
- מילים נרדפות (synonyms)
- WSD - זיהוי מילה על סמך הקשר
  (bass – a type of fish or tones of low frequency)
- זיהוי ישויות בטקסט Named entity recognition - NER

מטלת המעבדה

בסיום ההיכרות עם הספריות ניתן להמשיך לעבודה על אוסף הטוויטים, כאמור מומלץ לנסות התהליך באופן מבוקר, ורק לאחר הבנת טווחי הזמנים, ניתן להריץ על הכל.

- לאחר עיבוד מקדים ראשוני, שבוצע במעבדה הקודמת, יש לבנות מילון ראשוני של מושגים. (טוקניזציה)
- יש לבצע את הפעולות השונות על המילון ולהציג את השפעתם באמצעות ויזואליזציות וטבלאות שונות לבחירתכם והבנתכם.
- יהיה עליכם לנסות את התהליך בסדר שונה של פעולות כדי להגיע למסקנה מהו הסדר העדיף.
- כחלק מהניתוח שלכם, יש למדוד זמנים עבור כל פעולה ועבור כלל התהליך.