בניית מסווג למציאת משמעות הבניין של מילים

פרויקט א: נפעל

פרויקט ב: הפעיל

פרויקט ג: פיעל

פרויקט ד: התפעל

1. קורפוס
   - התנ"ך
   - עברית מודרנית

2. בניית אוסף הדוגמאות לאימון ובדיקה [שבועיים]
   - 500 דוגמאות מהתנ"ך (על פי קבצי ה TEI של ספרי המקרא, באתר הקורס), 500 דוגמאות ממאגר העברית המודרנית (Modern Hebrew Treebank באתר הקורס)

3. תיוג הדוגמאות
   - על פי שתי שתי גישות: בלאו, גלינרט [6 שבועות]
   - תיוג 100 דוגמאות ובניית 'מדריך תיוג'
   - תיוג שאר הדוגמאות

4. ייצוג כל דוגמא כווקטור של מאפיינים [שבועיים]
   - עברית מודרנית: על פי המאגר המנותח
   - ארבעה סוגי מאפיינים בסיסיים
     - המילים
       - תנ"ך: החלק השני בשדה dtoken, אחרי ה '_'
       - עברית מודרנית: הטור השני
     - ערך מילוני
       - תנ"ך: השדה lemma
       - עברית מודרנית: הטור השלישי
     - המאפיינים המורפולוגיים
       - קטגוריה לקסיקאלית (שם, פועל, תואר, יחס, קישור,...) ומאפייני הטיה (מין, כמות, זמן, גוף)
       - תנ"ך: החלק השלישי בשדה dtoken, אחרי ה '__'
       - עברית מודרנית: טורים 5-6
     - המאפיינים התחביריים
       - נושא, נשוא, מושא,...
       - תנ"ך: מוגדרים במבנה syntactic info בסוף כל פסוק. השדה function עבור על מילה – ניתן לשיוך ע"פ ה phraseId
       - עברית מודרנית: הטור השמיני
   - הרכבים של מאפיינים
     - למה + מורפולוגי
     - למה + תחבירי
     - למה + מורפולוגי + תחבירי
     - למה + חלק דיבר
     - מורפולוגי + תחבירי
     - חלק דיבר + תחבירי
   - משחק עם 'חלון' ההקשר

    - מילה קדימה ומילה אחורה
    - שתי מילים קדימה ושתי מילים אחורה
- בשורה התחתונה: לייצר ווקטורים שונים לכל דוגמא, על פי שילובים שונים של מאפיינים וחלונות

5. אימון מסווג על הדוגמאות ובדיקת איכותו [שבועיים]
- כלי: חבילת WEKA (אפשר לשחק עם האלגוריתמים ועם סוג הפונקציות שיש שם), או scikit-learn
- 10-fold cross validation

6. דו"ח ממצאים [שבוע]
- איכות המסווג (F-score), עבור אוספים שונים של מאפיינים.
- סטטיסטות
    - כל מה שנראה לכם רלבנטי. לדוגמא:
        - מספר המילים בבניין 'שלכם' ביחס לבניינים האחרים במאגר
        - מספר המילים השונות בדוגמאות שתייגתם ביחס למספר המילים השונות במאגר
        - ...
- ניתוח
    - אילו מאפיינים אפקטיביים יותר
    - איזו שיטה (רוזן, גלירט) קלה יותר לסיווג
    - האם יש הבדלים באיכות הסיווג התנ"ך מול סיווג העברית המודרנית
    - האם יש פעלים בעייתיים לסיווג? מדוע?
    - כל שאלה מעניינת אחרת, או תובנות כלשהן

יש להגיש בסוף:

1. אוסף הדוגמאות
2. קוד המסווג
3. דו"ח מסכם

- תיאור כללי של העבודה
- אופן בחירת הדוגמאות
- מדריך התיוג
- תיאור המאפיינים השונים שנבדקו
- סוג המסווג
- דו"ח הממצאים