עיבוד שפה טבעית - תרגיל בית 2 - רטוב

תיאור המשימה

בתרגיל בית זה תממשו מספר מודלים שנלמדו בביתה כדי לפתור את המשימה של זיהוי ישויות בטקסט (Named Entity Recognition), תבצעו משימות עיבוד שפה על נתונים אמיתיים ותנתחו את טיב הצלחתכם. את התרגיל יש לבצע בשפת python3.

בתרגיל תידרשו לממש שלושה מודלים – מודל פשוט, מודל הכולל רשתות ליניאריות ומודל לבחירתכם.

לאורך התרגיל הכוונה במדד F1 היא למדד F1 בינארי ברמת המילה. כלומר, שרשור כל הפרדיקציות של המודל שלכם והתיוגים האמיתיים על פני כל המשפטים, וחישוב binary F1 בין שתי הרשימות.

הסבר על מבנה הציון בתרגיל:

- **45%** - מימוש מלא של שני המודלים הראשונים ועמידה ברף של ציון F1 של לפחות 0.5 על קובץ ה development
- **20%** - תחרות מבוססת F1 בתיוג קובץ התחרות
- **10%** - כתיבת דו"ח תמציתי (עד 3 עמודים) אשר יכלול את הסעיפים הנדרשים ועמידה בתנאי פורמט ההגשה (יפורטו בהמשך המסמך)
- **25%** - מענה על השאלות בקובץ **HW2-Dry**

נתונים:

קבצי הנתונים של התרגיל הם בפורמט הבא:

כל שורה מייצגת מילה, ומכילה את המילה והתיוג שלה, מופרדים על ידי טאב. בסוף כל משפט יש שורה ריקה. אנחנו נתייחס לתיוגים של המילה בתור תיוג בינארי – התיוג הוא שלילי אם המילה קיבלה את התיוג O, וחיובי אחרת. סטודנטים שמעוניינים בכך יכולים להשתמש במידע הנוסף שנמצא בתיוגים, אך אין חובה לעשות זאת.

קובץ האימון הוא הקובץ train.tagged, קובץ ה development הוא dev.tagged וקובץ המבחן הוא test.untagged.

אימון

מודל ראשון (פשוט):

אתם נדרשים לממש מודל פשוט לבחירתכם מבין המודלים KNN/SVM או כל מודל אחר שנתמך על ידי הספרייה sklearn. כל מילה יש לייצג על ידי וקטור שמורכב מהייצוג שלה ב word2Vec/GloVe (ניתן להשתמש בוקטורים מאומנים מראש אך אין חובה) והייצוג של הסביבה שלה. מימוש אפשרי הוא כזה שמכיל את המילה ושתי המילים הסמוכות לה מכל צד, אך ניתן לממש בכל דרך שהיא.

מודל שני (רשת Feed Forward):

הדרישה במודל זה דומה לזו שתוארה במודל הקודם, אך הפעם המודל צריך להיות מודל מבוסס רשת FF (מעל ייצוג כפי שתואר בסעיף הקודם).

מודל תחרותי:

במודל זה אתם יכולים לבחור לממש כל מודל שתרצו. אתם יכולים לשנות את אופן הייצוג, המודל עצמו או כל דבר אחר שתראו לנכון. המודל יכול להיות אחד משני המודלים שהוגשו בסעיפים הקודמים.

מבחן (**Test**):

עבור שלושת המודלים יש לתייג את קובץ ה development, ולדווח את ציון הF1.

תחרות:

יש לתייג את קובץ התחרות על ידי המודל התחרותי. יש להגיש את התוצאות בקובץ בשם
comp_987654321_123456789.tagged כאשר 123456789 ו987654321 הם תעודות הזהות של בני הזוג.

סביבת עבודה:

לכל זוג הוקצתה מכונה בה מותקנות הספריות הנדרשות לתרגיל. על התרגיל לרוץ בסביבת הקונדה py38_default. אין להתקין ספריות נוספות לסביבה זו ללא אישור מסגל הקורס דרך הפורום.

הגשה:

קובץ zip בלבד, בשם HW2_123456789_987654321.zip(עבור שני סטודנטים שמספרי הזהות שלהם הם 123456789 ו- 987654321). הקובץ הנ"ל יכלול:

1. דו"ח קצר (עד 3 עמודים בפורמט PDF) המכיל הסברים תמציתיים, דיווח וניתוח תוצאות. שם הקובץ צריך להיות report_987654321_123456789.pdf. הדו"ח צריך לכלול:
   a. שמות המחברים ות"ז
   b. אימון - הסבר על כל מודל שמימשתם.
   c. מבחן - דיווח מדד F1 על קובץ ה development עבור כל אחד מהמודלים
2. קבצי הקוד של התרגיל. על הקוד להיות מתועד וקריא. בנוסף, הקוד צריך להיות מסוגל לרוץ על כל מכונה שהיא. אנא כתבו ממשקי הרצה פשוטים לאימון, מבחן וייצור קבצי התחרות המתויגים.
3. קובץ התחרות המתויג – על קובץ התוצאות להיות בפורמט tagged (כפי שמפורט בחלק "נתונים"), הכולל את המילים והתגים. על מנת לוודא נכונות ולהימנע מאי נעימות בנוגע לציון, אנא ודאו כי אם משמיטים את הטאב והתגים מהקובץ המתויג שאתם מגישים, מקבלים בדיוק את אותם משפטים (ולפי אותו סדר) ובאותו פורמט כמו בקובץ התחרות. חוסר התאמה פירושו ציון 0 בחלק הזה.
4. ממשק לתיוג קבצי התחרות **-** על קבצי התחרות להיות ניתנים לשחזור (Reproducible).
הדרישה היא שניתן יהיה לקחת את הקוד והמודל המאומן שהגשתם ולייצר באמצעותם קובץ תחרות מתויג זהה לחלוטין לקובץ שהגשתם.
לטובת שחזור הקובץ, יש לכתוב ממשק הרצה פשוט, בקובץ נפרד בעל השם –generate_comp_tagged.py להרצת Inference בלבד על המודל המאומן ויצירת קובץ התחרות המתויג.
5. חלק יבש - קובץ מוקלד עם תשובות לחלק היבש (HW2-dry) בשם dry_123456789_987654321.pdf

העתקות:

בשל אופי המשימה והמורכבות שלה, קל לבדוק העתקות של קטעי קוד \ קבצים מלאים. למען הסר הספק אנו מדגישים כי אין להעביר קוד בין סטודנטים, בין אם להגשה ובין אם לא. אין להעתיק קטעי קוד מוכנים מהאינטרנט, ובכלל אין להסתמך על שום מקור אחר לקוד מלבד פרי יצירכם והחבילות החיצוניות אשר צוינו בסעיף הרלוונטי.