מועד אחרון להגשה: 22/5/2018

יש להגיש דרך המודל: <https://moodle2.cs.huji.ac.il/nu17/mod/assign/view.php?id=235141>

# שאלה 1: (Word sense disambiguation (WSD, הפגת עמימות לקסיקלית

נתונים 5 פירושים שונים עבור המלה "ספר", ומידת הנפוצות (בסוגריים מרובעים) של כל אחד מהם:

| 1. קובץ דפים מאוגד יחד, עטוף בכריכה רכה או קשה, המיועד לקריאה ומכיל מידע או סיפור [54] 2. סכם, מנה, קבע את המספר או הכמות של אוסף פרטים באמצעות מנייתם [46] 3. גלב, אדם אשר מקצועו הוא הסרת שערות ועיצובן באמצעות גזירה, גילוח וכדומה [12] 4. רישום החשבונות של עסק [4] 5. אזור או ישוב הנמצא ליד הגבול [1] |
| --- |

1. בצעו טוקניזציה (tokenization) לכל אחת מההגדרות במסגרת, כלומר, הפרידו אותן לטוקנים (tokens).  
   דוגמה: מהמשפט "ולאחר מכן, ראינו שהוא יצא מהבית" נקבל [ו ל אחר מ כן , ראינו ש הוא יצא מ ה בית].
2. לכל אחד מ-6 המשפטים הנתונים מטה,

* בצעו טוקניזציה למשפט.
* ציינו מהו הפירוש המתאים (1-5) עבור המלה "ספר" במשפט, לדעתכם.
* הסבירו על סמך אילו מלים במשפט מלבד המלה עצמה הסתמכתם כדי לקבוע את משמעותה.
* האם אלגוריתם Lesk (ראו [במצגת](https://moodle2.cs.huji.ac.il/nu17/mod/resource/view.php?id=226254)) היה מצליח לגלות את הפירוש הנכון בהינתן ההגדרות לעיל, או שהוא היה בוחר בפירוש שגוי? אפשר להניח שהאלגוריתם סופר כל מילה רק פעם אחת, ומתעלם ממילות יחס וקישור (ב, ו, מ, ה, ל, ש, או, את, של) ומסימני פיסוק, אבל אינו מצויד במנתח מורפולוגי ("לקרוא" ו-"קרא" הן מילים **שונות**).

דוגמה: בהינתן המשפט "בחרתי **ספר** ממדף ספרי הקריאה, והתחלתי לקרוא", אחרי שנבצע טוקניזציה נקבל [בחרתי ספר מ מדף ספרי ה קריאה , ו התחלתי לקרוא]. הפירוש המתאים הוא פירוש מספר 1. אנחנו יודעים על סמך המלים "מדף", "קריאה" ו-"לקרוא". האלגוריתם היה מצליח, כי ההגדרה לפירוש 1 היא בעלת החפיפה הגדולה ביותר עם המשפט (מילה אחת: קריאה. לשאר הפירושים אין חפיפה).

המשפטים:

1. ההתיישבות ב**ספר** שינתה את אופיים של הצפון והדרום.
2. עוביו של **ספר** מוגבל לסנטימטרים מעטים, ולכן כאשר תכולתו עולה על 1,000 דפים, נהוג לפצלו.
3. **ספר** מקצועי כיום מטפל בדרך כלל רק בשיער הראש.
4. הילד **ספר** את הבתים בשכונה.
5. אזור ה**ספר** הוא מקום מסוכן יחסית למרכז
6. ב**ספר** הבישול יש עשרה מתכונים למנה ראשונה ועשרים למנה עיקרית.

# שאלה 2: (Semantic role labeling (SRL, תיוג תפקידים סמנטיים

עבור הפועל write קיים roleset יחיד ב-PropBank, אשר התפקידים בו הם:

| Arg0: writer, agent (מי שכותב)  Arg1: thing written, theme, topic (הדבר שנכתב)  Arg2: benefactive, recipient (מי שאליו נכתב הדבר, הנמען) |
| --- |

סמנו במשפטים הבאים את הארגומנטים Arg0, Arg1, Arg2 (אם קיימים). יש לסמן את **כל הצירוף** המתאים.

דוגמה:

[The brave soldier]Arg0 **wrote** [a heartbreaking letter]Arg1 to [her best friend]Arg2.

המשפטים:

1. The famous author **writes** every single one of his books on a typing machine.
2. His sister **wrote** him a note yesterday.
3. The young children **write** slowly and carefully.
4. This novel was **written** by Jane Austen.
5. Her story was never **written**.
6. Please **write** to me when you get home.

# שאלה 3 (רשות):

כתבו אלגוריתם overlap אשר מקבל כקלט שני משפטים (או באופן כללי רשימות של מלים) ומחזיר כפלט כמה מלים משותפות יש בין המשפטים. למשל overlap(first name for a girl, he knows this girl 's name) = 2