סמינר בשיטות סטטיסטיות לעיבוד שפות טבעיות

עבודות

1. זיהוי סוף משפט: כתוב תכנית לומדת והרץ אותה על טקסטים עבריים ואנגליים. בדוק אם התכנית משפרת את ביצועיה כאשר קובץ האימון גדול יותר, ואיך פרמטרים שהותאמו לאנגלית מתאימים לעברית והפוך. Manning and Schuetze p. 147 (ראה גם תרגיל 7.15 עמ' 263 ותרגיל 10.8 עמ' 351).
2. זיהוי שפה: כתוב תכנית שקוראת טכסטים משפות שונות ולומדת לזהות את השפה. השווה בין גישות שונות: מילים אופייניות, סיומות, אורכי מילים (באוציות), אורכי משפטים (במילים) וכו'. Manning and Schuetze p. 227
3. Tagger לעברית: מצא ברשת tagger שנכתב לאנגלית, והרץ אותו על עברית. Manning and Schuetze p. 380 (אתר http://www.sultty.arts.usyd.edu.au/fsnlp/tagging).
4. בניית מנתח תחבירי סטטיסטי Manning and Schuetze p. 460.
5. Allignment: המאמר של Gale and Church (מס' 17 ברשימת הספרות) מכיל תכנית. הגרסה שלהם מכילה בגים. יש לי גרסה אלקטרונית לא מדובגת. המשימה: לדבג את התכנית ולערוך ניסויים בטקסטים אנגליים/עבריים.
6. חלוקת טכסט לפרקים ופרקי משנה.

האתר של הספר של Manning and Schuetze ,
http://www.sultry.arts.usyd.edu.au/fsnlp
מכיל מחוונים להמון תכניות, קורפוסים, מאמרים וכו'.