**סוגי ישויות (עם רמות וודאות שונות)**

1. NAME
   1. Person
   2. Company
   3. Location
2. Numbers
   1. Date
   2. Currency
   3. Other numbers

**אינדיקטורים- ממושקלים, אופציונלים**

* ניתוח תחבירי- שע"צ
* פייסבוק- פרופיל, page, places
* google maps- מקומות
* linkedin- חברות, אנשים
* שם domain- חברה
* אותיות גדולות
* מרכאות
* מילים מרמזות- mr. dr., from, to,
* חיפושים קודמים- אולי במקום לחפש
* לבדוק מילה, ולהרחיב למילים לפני / אחרי (או- להסתכל על זוגות, שלשות)- פרמטר?
* מילון
* the לפני מילה (כנראה לא שם, אבל עם ה-the אולי כן)

**פעולות הכנה**

* למצוא מאגרים
* חלוקה למשפטים
* להוריד upper case ?
* tokenization
* להסיר stop words

**ניסויים**

* בדיקה מול קורפוס (אם יש)
* בדיקה מול סטנפורד / כלי קיים
* בדיקה מול בדיקה טריוויאלית של אות גדולה

**פינצ'רים נוספים**

* מידע נוסף שנמצא בגוגל- למשל על אדם- מין, גיל, חברה- במה מתעסקת...

**מידע נוסף לאחר הישיבה עם לנה ונחום**

* n-grams- עד 5 מילים
* APIs-
  + ללנה יצא להשתמש ב-duck duck go- חינמי, ללא הגבלות
  + לנה תברר לגבי גוגל דרך האוניברסיטה
  + לכפיר יצא לעבוד על facebook
* השוואה-
  + לכלי קיים של NER
  + precision ו-recall לפי מאגר מתויג (נחום שלח לינק לטכניון)
* טיפים כלליים
  + לעבוד שפה-שפה
  + להתחיל מתשתית גנרית עם מעט פיצ'רים, לנתח אותם, ולהוסיף עוד בהמשך לפי הזמן שנשאר
  + לתאר גם מה חשבנו לבצע ובסוף לא הספקנו
* אם לא מסתדר עם מנוע חיפוש, לשקול לעבוד מול ויקיפדיה- להוריד עותק ולחפש ערכים. אולי מסווג לפי סוגים, גם אם לא- אפשר לחפש מילות מפתח בפסקה הראשונה. לא בנוסף למנוע חיפוש- יותר מידי API ללמוד לפרויקט.