שנה שניה של הקורס.

אתר שיכול לעזור
https://gyungsunsonnykim.wordpress.com/category/data-science/natural-language-process/text-mining/page/2/

Information Retrieval

אחזור מידע - ינתר אחזור טקסטים (כמו בגוגל)

Course number: 250310

lecturer: Dror Mughaz

Course Requirements

- Python programming Assignments: ~~25%~~ - 30%.
- Final Exam: ~~75%~~ - 70%.

Motivation: Harnessing big text data

מיוצר

- Text data is produced by humans
  - Contains people`s opinions
  - Encodes human knowledge
  - Offers opportunity for discovering knowledge

נצרך

- Text consumed by humans
  - Need intelligent tools
  - Humans play an essential role in mining

צריך לדעת מה קהל היעד
ולפי זה לתבים.

אנשים יכולים להבין דברים שמכונה לא.

Course objectives:

איך מנועי חיפוש עובדים? איך למדתם?

- Basic concepts and practical techniques in text retrival
  - How search engines work
  - How to implement a search engine
  - How to evaluate a search engine
  - How to improve and optimize a search engine
  - How to build a recommended system
- Hands-on experience on
  - Creating a test collection to evaluate a search engine
  - Experimenting with search engine
  - Participating in a search engine competition

What do we hope to teach?

פתלמיקלס

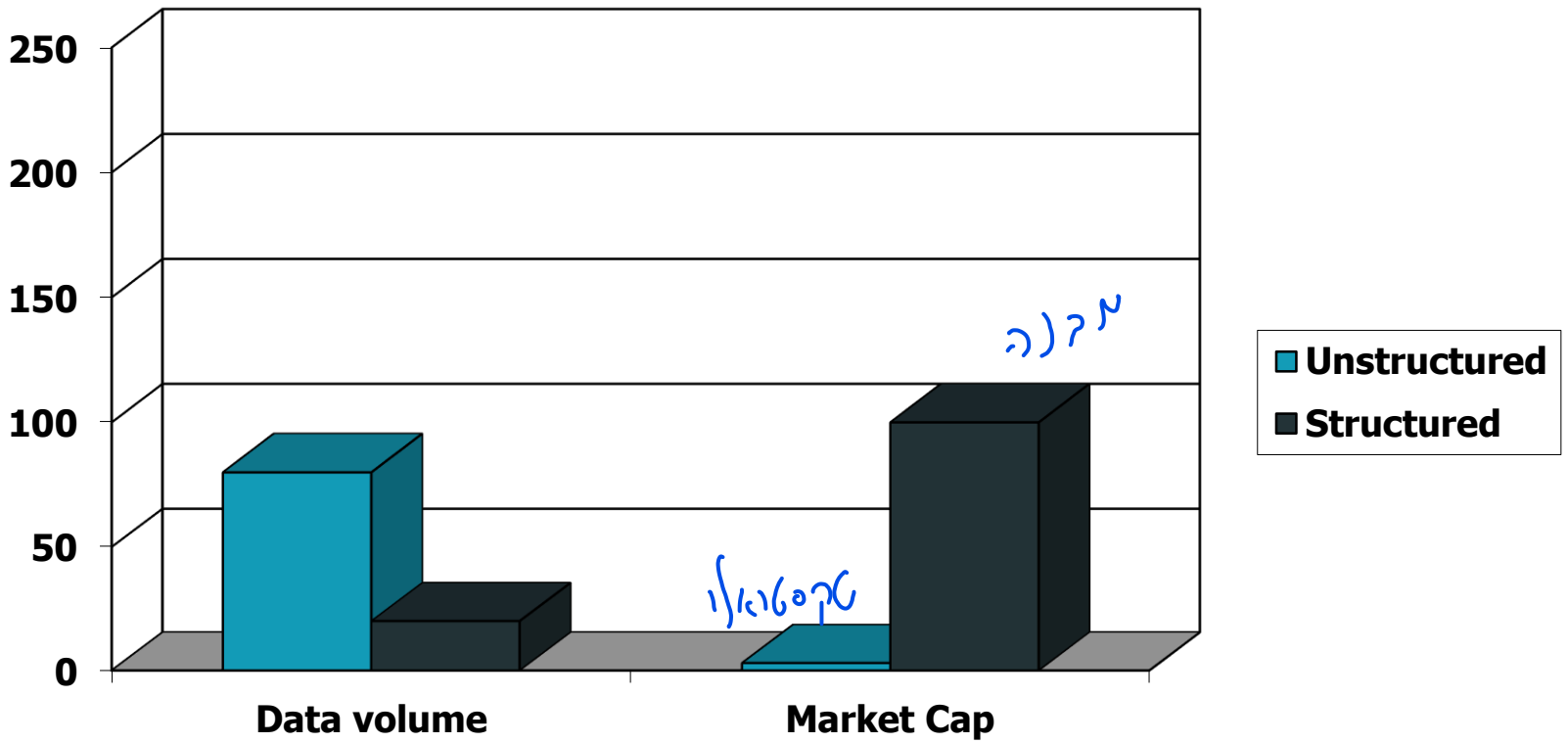
- כיצד לבצע אינדקס טקסט יעיל (מהיר, קומפקטי).
- מודלים לאחזור: מודלים בוליאניים, וקטוריים-מרחביים, הסתברותיים ולמידת מכונה.
- הערכות ובעיות ממשק של אחזור מידע.
- אישכול וסיווג מסמכים.
- חיפוש באינטרנט, כולל סריקה, אלגוריתמים מבוססי קישורים, משוב עקיף, נתוני-על והתאמה אישית. כל אחד אישי

מלא - עיבוד שפה טבעית / שיטת טיבול.

אחזור מידע

- אחזור מידע (Information Retrieval - IR) הוא מציאת חומר (בדרך כלל מסמכים) בעל אופי לא מובנה (בדרך כלל טקסט) העונה על צורך במידע מתוך אוספים גדולים (בדרך כלל מאוחסנים במחשבים).
- בימים אלה אנו חושבים תחילה על חיפוש באינטרנט, אך ישנם מקרים רבים אחרים:
  - חיפוש בדואר אלקטרוני
  - חיפוש במחשב האישי
  - חיפוש בבסיסי נתונים של תאגידים
  - שליפת מידע משפטי

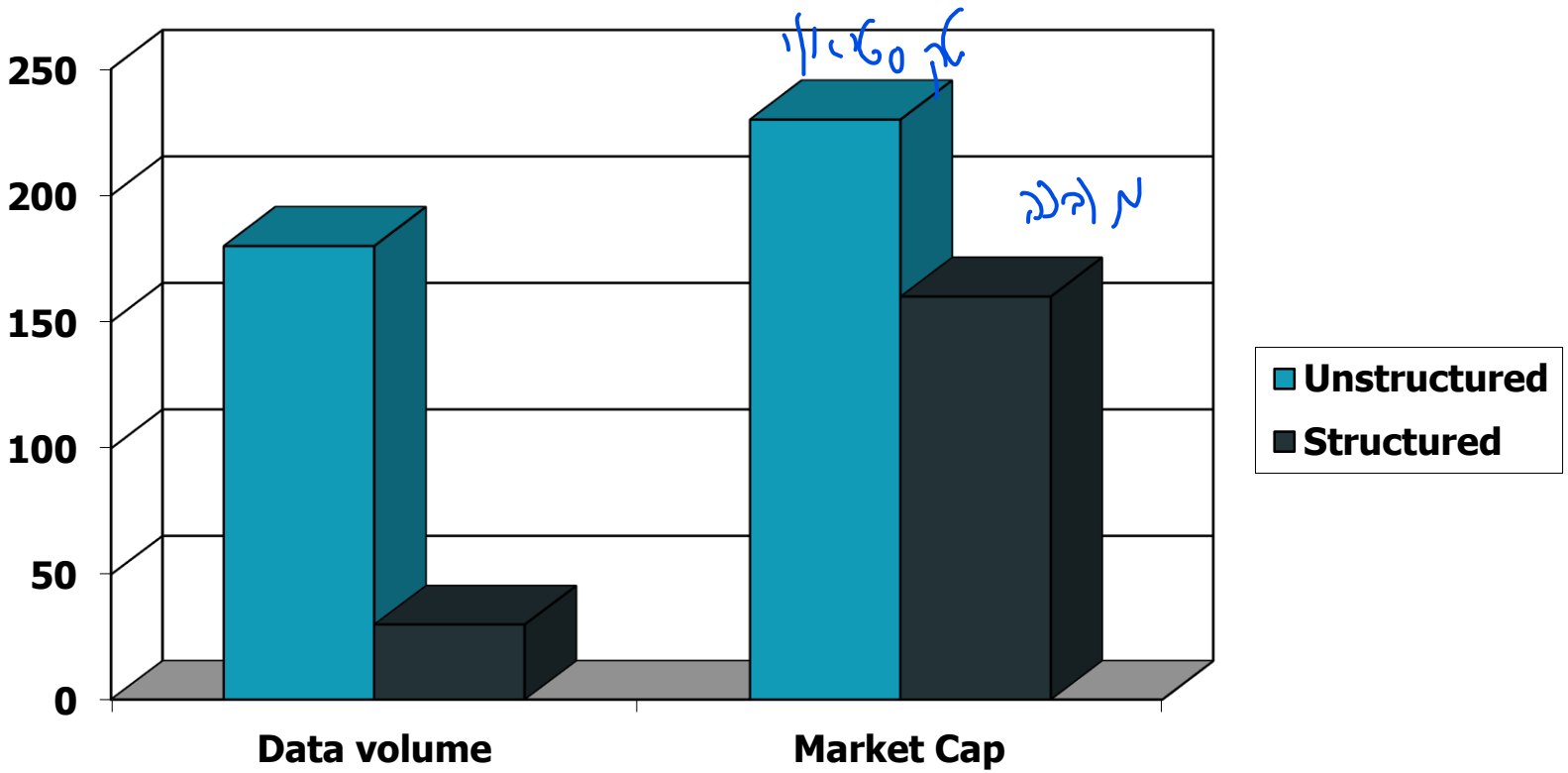
נתונים לא מובנים (טקסט) לעומת נתונים מובנים (בסיס נתונים) - שנות התשעים



בהחלט ייתכן שהמספרים אינם מדוייקים אבל הם משקפים:

- בשנות ה- 90 היו נתונים (DV) מובנים, מסדי נתונים רלציונים וכדומה
- היו נתונים בצורה שאינה מובנת, טקסט חופשי
- בתעשיה היו הרבה מסדי נתונים בעוד בתחום של טקסט חופשי היו מעט מאוד כלים וחברות

נתונים לא מובנים (טקסט) לעומת נתונים מובנים (בסיס נתונים) – עשור שני של שנות האלפיים



היום המצב הוא שונה לחלוטין

- ה- DV גדל גם במובנה וגם בשאינו מובנה, במיוחד בשאינו מובנה (בלוגים, טוויטים, פייס-בוק ...)
- התעשייה "נכנסה להילוך" בתחום של טקסט חופשי, למשל מנועי החיפוש ועוד

הנחות יסוד לאחזור מידע (IR)

- אוסף: סט מסמכים
  - כרגע אנו נניח שזה אוסף סטטי

- מטרה: אחזור מסמכים עם מידע הרלוונטי למשתמש בהתאם לדרישת/בקשת המשתמש ועוזר למשתמש להשלים משימה

Examples of information system applications

- Search
- Filtering – קורה
- Categorization – כמו אשקלון
- Mining/extraction בין השורות בתוך השורות
- Many others …

דוגמאות שונות לחיפושים שונים

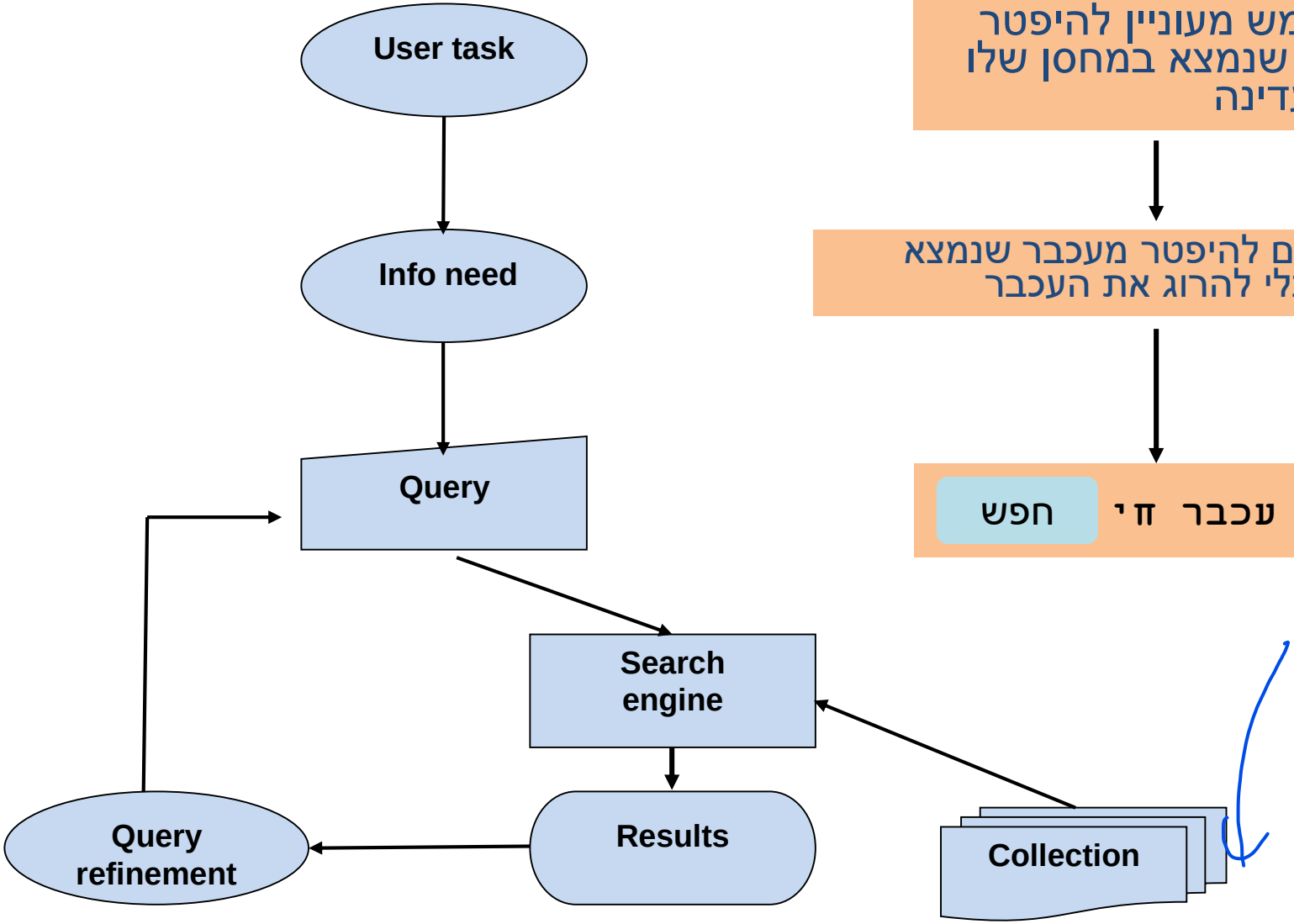
- אתר אינטרנט של חדשות או מאמרים חדשותיים
- בלוגים שונים
- מאמרים מדעיים
- הודעות בטוויטר (כרגע ממש נשלחות הרבה הודעות)
- מיילים (כרגע ממש נשלחים הרבה מיילים)
- יצירת/חיפוש קבוצת "חברים" בפייסבוק
- תמונות הקשורות לנושאים מסויימים

מודל החיפוש הקלאסי

המשתמש מעוניין להיפטר מעכבר שנמצא במחסן שלו בדרך עדינה

מידע על דרכים להיפטר מעכבר שנמצא במחסן אבל בלי להרוג את העכבר

כיצד ללכוד עכבר חי חפש



יתר מדי משמעי

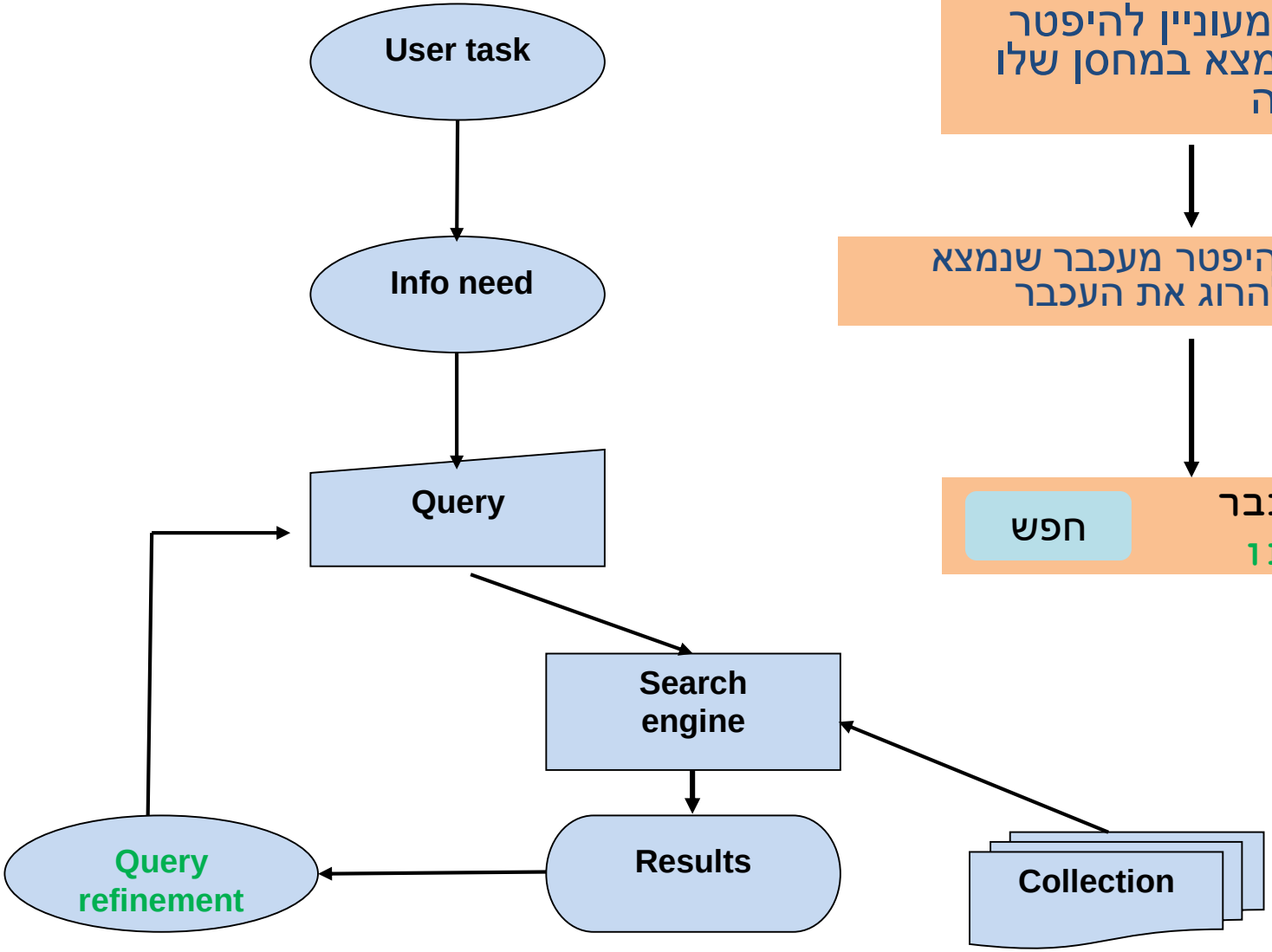
מודל החיפוש הקלאסי

המשתמש מעוניין להיפטר מעכבר שנמצא במחסן שלו בדרך עדינה

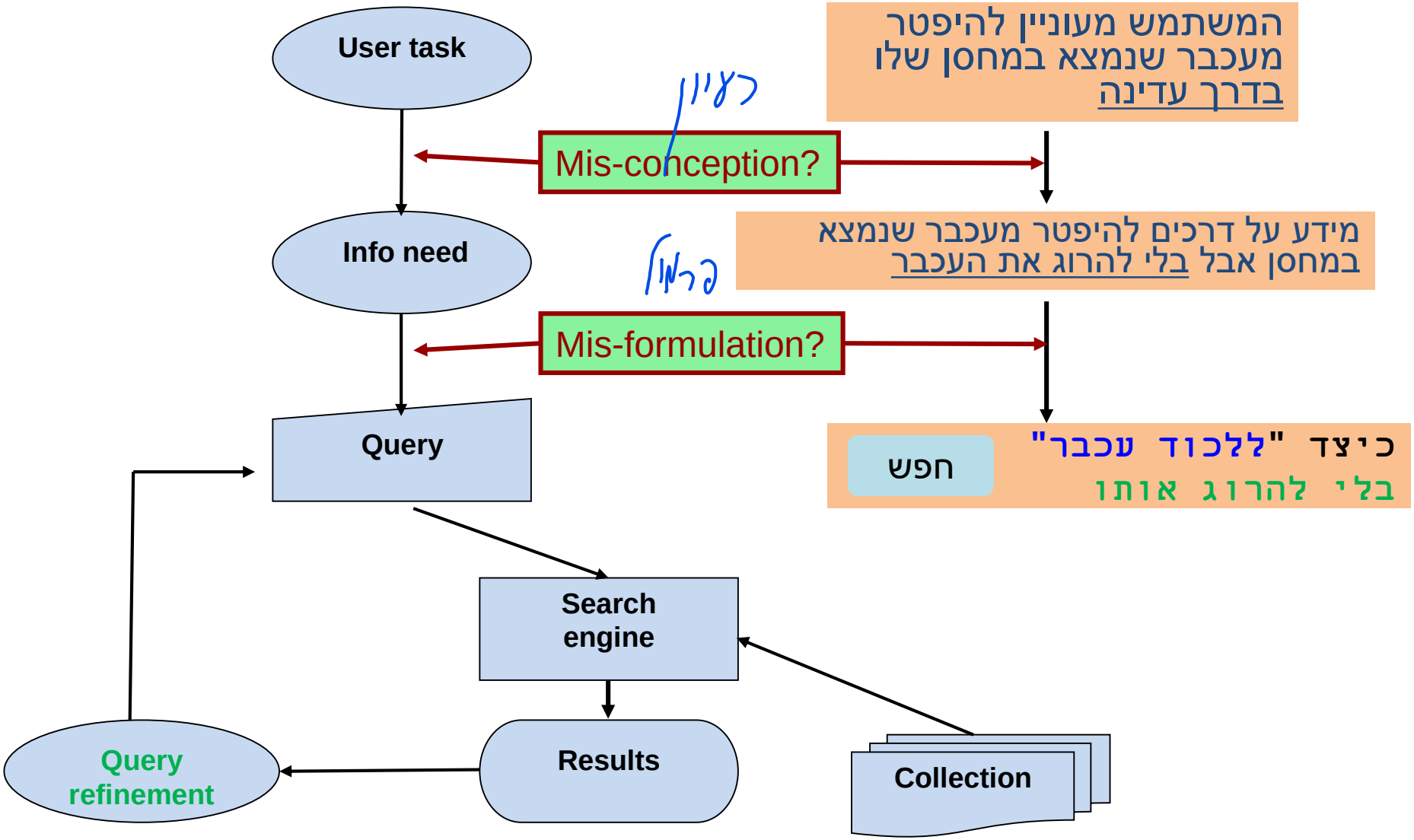
מידע על דרכים להיפטר מעכבר שנמצא במחסן אבל בלי להרוג את העכבר

כיצד ללכוד עכבר בלי להרוג אותו

חפש



מודל החיפוש הקלאסי



מתוך 1000 יש 100 רלוונטיים ואוחזרו 80, ש 50 הינו רלוונטי
לפי המערכת

כמה טובים המסמכים שאוחזרו?

מגדרים לטיב המערכת.
הגדרות יותר מדויקות בהמשך.

- Precision : קבוצה של מסמכים שאוחזרו **ורלוונטיים** למשתמש מתוך כלל המסמכים שאוחזרו (בהתאם לבקשתו של המשתמש)

$\left(\frac{50}{80}\right)$

- Recall : קבוצה של מסמכים שאוחזרו ורלוונטיים למשתמש מתוך **כלל המסמכים שרלוונטיים** (בהתאם לבקשתו של המשתמש)

$\left(\frac{50}{100}\right)$

- F1 : פונקציה של שני המדדים דלעיל — נלמד בהמשך

  – הגדרות מדויקות יותר בהמשך הקורס

נרחיב בהמשך...

שקף שילוה אותנו בקורס.

מודל כללי של מנוע חיפוש