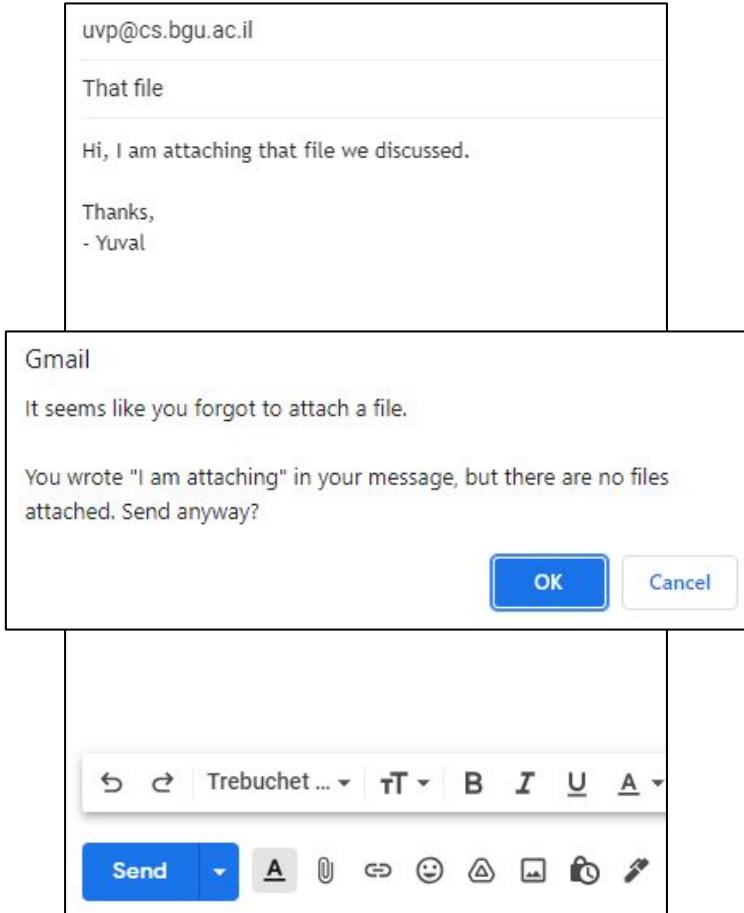


# עיבוד שפה טבעיות ש2:

## עיבוד טקוט

# עיבוד טקסט "שטחי"



- הפעלת כלים חישוביים שלא עושים שימוש ולא מתאימים לעשות שימוש בשום ידע לשוני
  - הרבה פעמים, זה כל מה צריך (שים לב להסבירנות!)
  -
- בתרח נל"פ, כפניות טרומ-עיבוד (pre-processing) - חשוב לטובות
  - ניקוי טקסט (תagi html, מספרי שורה)
  - האחדת טקסט (normalization / canonicalization)
  - פיצול למשפטים (sentence segmentation)
- חישוב סטטיסטיות ראשוניות ו"בדיקה שפהית"

# "דרישת קדם"

- להציג גישה ל-shell x11ux (בטעם :(bash :)
- מוק: יש מובנה, פשוט לפתח טרמינל
- חלונות 10: [ubuntu subsystem](#)
- מכל מחשב שהוא: גישה מרוחקת לשרת יוניקס, אוניברסיטאי או לא
- המלצה חמה: לעבור על [מדריך יוניקס למשוררים](#) (מ קישר גם ממודל), לראות שהכל עובד, ולהפניהם.

## מושגי יסוד בטקסט

types 4 6 tokens 6 6  
: נטול הנקודות  
do n't  
are n't  
were n't  
: נטול הנקודות  
don't do : Tokeners  
לפניהם

For lemma: Be  
Is for: Are  
Am  
Is  
Was

For Building → The lemma is:  
Build if it's  
a Verb  
Building if it's  
an object

- תבנית (type) מול תמנית (token)
  - ייחס התבניות/תמןיות (TTR)
  - Herdan's law:  $|V| = kN^\beta$
  - חלוקה לתרミニות = tokenization

צורת יסוד (lemma)

- גבעול (stem)
  - מילה: stories
  - צורת יסוד: story
  - גבעול: stori

The grassy field was full of grass, the flowers are laying on it en masse,  
each flower has a petal and the groundskeeper is listening to heavy metal.

- כמה תמניות? 28
  - כמה תבניות?
  - מה ה-TTR? קיומן גז-1
  - מה קורה עם פיסוק?
  - אילו מילים חולקות צורת יסוד?
  - אילו מילים חולקות גבעול?
- flower & flowers

אמרו לנו שעברית היא שפה שקשה לעבד אותה עם מחשבים, או להבין מה קורה אליה על-בסיס  
שיטות שפותחו לשפות כמו אנגלית, אבל אני לא חושב שהצלחנו לקלוט עד כמה.

- כמה תמניות?
- כמה תבניות?
- מה ה-TTR?
- מה קורה עם פיסוק?
- אילו מילים חולקות כורת יסוד?
- אילו מילים חולקות גבעול?

## עיבוד זרי

- שימוש בפקודות יוניקס "מלוכלות":

tr -sc 'A-Za-z' < wizard.txt | head

טראנסלט שורה אחת של אותיות אלייזר גולדשטיין. גולן

חלופי תווים tr

מיון sort

יחוד uniq ← נציגותם של כל אחד

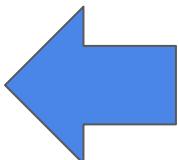
(הצגת עזרה ע"י man או --help)

(הצגה נגלת באמצעות less)

(בדיקות תור-כדי באמצעות head-tail)

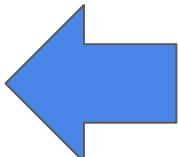
(ספרית שורות עם wc)

ה"קורפו" הועלה למודל - wizard.txt



# ביטויים רגולריים

- עובדים בד"כ ברמת השורה
  - יש כל מיני "טעמים" (flavors)
- <https://regex101.com/>
  - ממליץ לשחק איתו קצת לפני הבחן הבא
- <https://alf.nu/RegexGolf>
  - מה שambil אתנו לסוגי שגיאה



# מרחק ערך (edit distance)

- מטריקה להשוואה בין שתי מחרוזות
  - בرمת המילה או בرمת התו
- שימוש ציון למערכת: "כמה קרוב" היא הצליחה להגיע ל"מחרוזת הדروשה"
- יתרונות משמעותיים: מוגדרת היטב וקלה לחישוב
  - ומילאה קצת "מרחיב משחק" בדמות ציונים שונים לעדכויות שונות
- חסרון עיקרי: מוטיבציה לשונית, או אפילו נל"פית, מוטלת בספק
- בפועל נמצא אותה לרוב ב: משימות גנרטיביות בرمת התו (מציאת הטוית, ייצור מילים חדשות, תיקון שגיאות, וכו')
- אלה כן, וביולוגיה ("שור ריצופים למי שחוגג.ת")

# פעולות אוטומיות

- הוספה (insertion): חבלת  $\leftarrow$  חבלות
- מחיקה (deletion): חבלת  $\leftarrow$  בצלת
- שיכול (transposition): חבלת  $\leftarrow$  חבלת
- החלפה (substitution): חבלת  $\leftarrow$  חצמת
  - הרבה פעמים נספר כמחיקה + הוספה
- מרחק עירכה (לובנשטיין Levenshtein): מינימום הפעולות הנדרשות להגעה ממחוזת אחת למחוזת שנייה
  - לובנשטיין מושקל: פעולהות שונות יכולות לגורור מחיר שונה

## אופן הפתרון - תכונות דינמי

- נתונות מחרוזת א' בעלת אורך ח ומחרוזת ב' בעלת אורך ז
- נגידר את (j,i) D להיות מינימום הפעולות הנדרשות למעבר בין הראשית-באורק-ו של א'  
והראשית-באורק-ז של ב', וنبנה טבלה לחישוב איטרטיבי.
  - מהו הפתרון הכללי?
  - כמה חישובים נידרש לעשות?
  - כמה זכרון נדרש?
  - כמה זמן ייקח "ליישר" אח"כ את המחרוזות (=להתאים בין התווים ולסמן מחיקות ותוספות) אם שמרנו מצביעים בתוך הטבלה?

## (סעיף 14 מtower המציגת של SLP)

Initialization

$$D(i,0) = i$$

$$D(0,j) = j$$

Recurrence Relation:

For each  $i = 1 \dots M$

For each  $j = 1 \dots N$

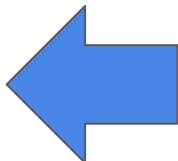
$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases}$$

Termination:

$D(N,M)$  is distance

## אופן הפתרון - תכונות דינמי

- נתונות מחרוזת א' בעלת אורך ח ומחרוזת ב' בעלת אורך ז
  - נגידר את (j,i) D להיות מינימום הפעולות הנדרשות למעבר בין הראשית-באורכו של א' והראשית-באורכו-ז של ב', ובננה טבלה לחישוב איטרטיבי.
    - מהו הפתרון הכללי?
    - כמה חישובים נידרש לעשות?
    - כמה זכרון נדרש?
    - כמה זמן ייקח "ליישר" אח"כ את המחרוזות (=להתאים בין התווים ולסמן מחיקות ותוספות) אם שמרנו מצביעים בתוך הטבלה?



- נדגים באמצעות התאמת בין  **מרפי למטריה**.
  - (ניחוש זרייז כמה י יצא בסוף, מיישריז?)

1 1 2 7 1 2  
2 3 6 1  
1 0 1

1	2	3	2	N	+	1
4	3	2	1	0	#	
3	2	1	0	1	1	
4	3	2	1	2	C	
3	2	1	2	3	2	
2	3	2	3	4	1	
3	4	3	4	5	2	