<u>עיצוב המערכת</u>

המערכת בנויה מ 3 שלבים של Map-Reduce:

1. **בשלב הראשון** אנחנו סופרים את כל מספרי ההופעות הרלוונטיות לחישוב הקשר association בין lexeme ל feature. הם: count(L), count(F), Count(l,f), Count(L=l), Count(F=f)

הרשומות שהתוכנית מקבלת הן מהצורה: <מספר שורה, שורה המתארת עץ תחבירי>
הרשומות שהתוכנית מוציאה הן מהצורה: <lexeme feature-label, count>
ובנוסף התוכנית שומרת בזיכרון את Count(L) ואת Count(L=l) עבור כל לקסמה, דבר שמתאפשר לפי הגבלות הזכרון כיוון שבמאפר מבוצע סינון כך שרק מילים מה Golden-Standard עוברות.

2. **בשלב השני** אנו מחשבים את מדדי הassociation ומקבצים לכל lexeme את כל הפיצ'רים הרלוונטיים וציוני association שלהם.

שלב זה מוציא רשומות בצורה lexeme , HashMap<feature,[count, freq, PMI, T–Test]>

3. **בשלב השלישי** אנו מבצעים 3-way join על הרשומות עם עצמן ועם קובץ הזוגות (סינון) ומחשבים עבור כל זוג רלוונטי את הוקטור הסופי בגודל 24 מדדי דמיון!

הרשומות הסופיות היוצאות משלב זה הן מהצורה הבאה (דוגמא):

lexeme lexeme [0.0 , 0.0 , 0.999998 , 1.0 , 1.0 , 0.0 , 0.0 , 0.0 , 0.999999, 1.0 , 1.0 , 0.0 , 0.0 , 0.0 , 0.9999999 , 1.0 , 1.0 , 0.0 , 0.0 , 0.0 , 0.9999998 , 1.0 , 1.0 , 0.0]

<u>נתוני מידע - ריצת 10%:</u>

שלב 1:
רשומות שהתקבלו במאפר: 159,033,091
רשומות שיצאו מהמאפר: 178,902,389 בגודל 4,137,941,736 Bytes
רשומות שיצאו מהקומביינר: 7065206 בגודל 63,256,749 Bytes
רשומות שיצאו מהרדיוסר: 3811348

שלב 2:
רשומות שהתקבלו במאפר: 3811348
רשומות שיצאו מהמאפר: 3261318 בגודל 173,162,120 Bytes
רשומות שיצאו מהקומביינר: 0 (לא התרחש)
רשומות שיצאו מהרדיוסר: 414

שלב 3:
רשומות שהתקבלו במאפר: 414
רשומות שיצאו מהמאפר: 2349864 בגודל 11,150,372,668,474 Bytes
רשומות שיצאו מהקומביינר: 0 (לא התרחש)
רשומות שיצאו מהרדיוסר: 48

מדדים: שימוש בSGD Text

```
Correctly Classified Instances          44               91.6667 %
Incorrectly Classified Instances         4                8.3333 %
Kappa statistic                          0.75
Mean absolute error                      0.0833
Root mean squared error                  0.2887
Relative absolute error                 21.8193 %
Root relative squared error             66.386  %
Total Number of Instances               48

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.667    0.000    1.000      0.667    0.800      0.775    0.833     0.750
similar
                 1.000    0.333    0.900      1.000    0.947      0.775    0.833     0.900
not_similar
Weighted Avg.    0.917    0.250    0.925      0.917    0.911      0.775    0.833     0.862

=== Confusion Matrix ===

  a  b   <-- classified as
  8  4 |  a = similar
  0 36 |  b = not_similar
```

- **דיוק כולל:** 91.67% – 44 מתוך 48 מופעים סווגו נכון.
- **מטריצות למחלקת "similar":**
  - True Positive Rate (Recall): 66.7%
  - Precision: 100%
  - F-Measure: 80%
- **מטריצות למחלקת "not_similar":**
  - True Positive Rate (Recall): 100%
  - Precision: 90%
  - F-Measure: 94.7%
- **מטריצות משוקללות:**

  Precision: 92.5%, Recall: 91.7%, F1: 91.1%