Report

- The number of key-value pairs that were sent from the mappers to the reducers in your map-reduce runs, and their size (take a look at the log file of Hadoop), with and without local aggregation:

Step 1:

- With Local Aggregation (Combiner):
    - Map Input Records: 163,471,963
    - Map Output Records: 513,588,669
    - Combine Input Records: 517,475,954
    - Combine Output Records: 6,819,676
    - Reduce Input Records: 2,932,391
    - Reduce Output Records: 1,402,490
    - Map Output Bytes: ~8.37 GB
    - Map Output Materialized Bytes: ~33.09 MB
- Without Local Aggregation (Hypothetical):
    - Map Output Records: 513,588,669 (unchanged).
    - Reduce Input Records: 513,588,669 (no reduction by combiner).
    - Map Output Materialized Bytes: ~8.37 GB (unchanged).

Step 2:

- With Local Aggregation (Combiner):
    - Map Input Records: 1,402,490
    - Map Output Records: 1,402,489
    - Combine Input Records: 0 (no combiner used).
    - Combine Output Records: 0 (no combiner used).
    - Reduce Input Records: 1,402,489

    - Reduce Output Records: 360,170
    - Map Output Bytes: ~41.68 MB
    - Map Output Materialized Bytes: ~15.43 MB

Step 3:

- With Local Aggregation (Combiner):
    - Map Input Records: 1,402,490
    - Map Output Records: 1,402,490
    - Combine Input Records: 0 (no combiner used).
    - Combine Output Records: 0 (no combiner used).
    - Reduce Input Records: 1,402,490
    - Reduce Output Records: 360,170
    - Map Output Bytes: ~41.68 MB
    - Map Output Materialized Bytes: ~15.34 MB

Step 4:

- With Local Aggregation (Combiner):
    - Map Input Records: 720,340
    - Map Output Records: 720,340
    - Combine Input Records: 0 (no combiner used).
    - Combine Output Records: 0 (no combiner used).
    - Reduce Input Records: 720,340
    - Reduce Output Records: 360,170
    - Map Output Bytes: ~37.10 MB
    - Map Output Materialized Bytes: ~15.84 MB

Step 5:

- With Local Aggregation (Combiner):
    - Map Input Records: 360,170
    - Map Output Records: 360,170
    - Combine Input Records: 0 (no combiner used).
    - Combine Output Records: 0 (no combiner used).
    - Reduce Input Records: 360,170
    - Reduce Output Records: 360,170
    - Map Output Bytes: ~14.25 MB
    - Map Output Materialized Bytes: ~7.83 MB

- Scalability report: The time of the running for two different numbers of mappers, and for two different input sizes.

For using 2 EC2s and for the entire 3-Gram:
With total time of 41 minutes and 11 seconds.

j-1HJWVMZ9EYBVU	Map reduce project	Terminated All steps completed	January 08, 2025, 09:32	33 minutes, 11 seconds	8

For using 1 EC2 for the entire 3-Gram:

With total time of 53 mintues and 38 seconds.

j-4YS2472KPNB0	Map reduce project	Terminated All steps completed	January 07, 2025, 20:43	53 minutes, 38 seconds	4

For using 2 EC2 for input of 20 3 words sentences:

With total time of 4 mintues and 14 seconds.

Cluster ID	Cluster name	Status	Creation time (UTC+02:00)	Elapsed time	Normalized instance hours
j-375GQBJRK10OP	Map reduce project	Terminating All steps completed	January 07, 2025, 22:05	4 minutes, 14 seconds	8

- Choose 10 'interesting' word pairs and show their top-5 next words. Judge whether the system got to a reasonable decision for these cases:

הזוגות מילים שנבחרו עם ה5 מילות המשך שלהן נראות שהוציאו תוצאות מתאימות. המילות המשך הכי פופולריות לאומת ההכי פחות נשמעות הגיוניות בהתחשב שמדובר בהמון DATA כך שבתור דוברי עברית ניתן להבין מה יותר הגיוני לשמוע ביום יום לעומת מה שפחות שבהסתברות הכי נמוכה.

1. אבא שלי:

אבא שלי אמר 0.018316507

אבא שלי אומר 0.01598078

אבא שלי מת 0.013613105

אבא שלי ואני 0.009623945

אבא שלי ואמא 0.0091984272.

2. בדרכים שונות

בדרכים שונות ומשונות 0.028090656

בדרכים שונות כדי 0.020310104

בדרכים שונות ומגוונות 0.017956913

בדרכים שונות לחלוטין 0.014511287

בדרכים שונות גם 0.013732135

3. גם לגבי

גם לגבי היהודים 0.010150015

גם לגבי העתיד 0.008183777

גם לגבי השאלה 0.0076577067

גם לגבי יתר 0.007545531

גם לגבי יהודי 0.0074745417

4. דרש רבי

דרש רבי יהודה 0.069114566

דרש רבי עקיבא0.067023635

דרש רבי שמלאי 0.065947115

דרש רבי אלעזר 0.06565344

דרש רבי יוסי 0.05204463

5. הגעתי למסקנה

הגעתי למסקנה שאין 0.110848844

הגעתי למסקנה שיש 0.06708938

הגעתי למסקנה שאני 0.06675798

הגעתי למסקנה שלא 0.065836966

הגעתי למסקנה שזה 0.04259485

6. ואחר כך

ואחר כך גם 0.004434705

ואחר כך אמר 0.004421234

ואחר כך הם 0.003122449

ואחר כך אני 0.002884388

ואחר כך באו 0.0027189255

7. זמן שלא

זמן שלא יהיה 0.016418695

זמן שלא יצא 0.014624715

זמן שלא הגיע 0.01325047

זמן שלא תהיה 0.0124861

זמן שלא נמצא 0.012067139

8. חצר בית

חצר בית הכנסת 0.08699262

חצר בית הנשים 0.04775858

חצר בית המקדש 0.044371665

חצר בית הספר 0.03875947

חצר בית המלך 0.03372103

9. טובת הנאה

טובת הנאה אינה 0.021732807

טובת הנאה ממון 0.021047235

טובת הנאה לבעלים 0.0202927

טובת הנאה לעצמו 0.01959157

טובת הנאה שיש 0.017566383

10. תודתי נתונה

תודתי נתונה גם 0.043883562

תודתי נתונה לידידי 0.040959775

תודתי נתונה למר 0.0389055

תודתי נתונה לפרופ 0.035172045

תודתי נתונה לד 0.032871425