

Summarization



גלעד מרקמן
[פרויקט ב- Github](#)

סיכום טקסט באמצעות מודלי שפה

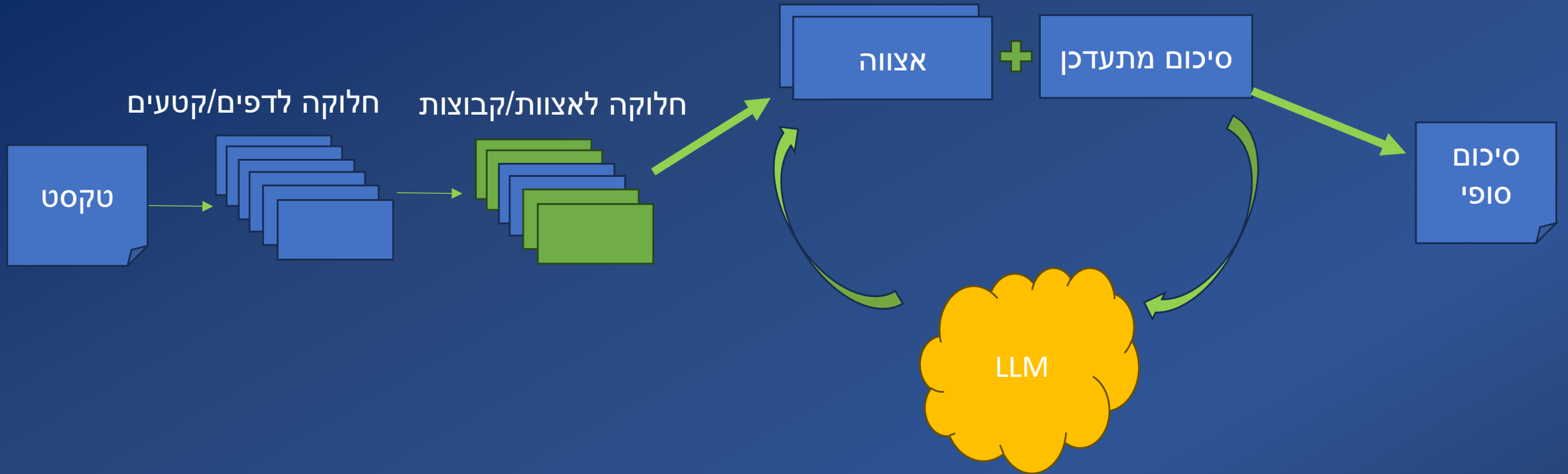
- מודלים של בינה מלאכותית (LLM) מצטיינים בסיכום טקסט קצר.
- המודלים מתקשים לסכם היטב טקסט ארוך:
 - אורך הקלט של המודלים מוגבל, ולכן טקסט ארוך מידי לא נכנס כולו למודל (context window).
 - ירידה באיכות - כאשר "דוחסים" כמות גדולה של טקסט, המודל מתקשה לשמור על מבנה העלילה או הטיעון, ולעיתים נוטה לפספס נקודות חשובות או לחזור על עצמו.
- המודלים האחרונים שיצאו לשוק מאפשרים Context Window של עד מיליון Tokens (כולל את הקלט והפלט). אולם, הסיכום של כמות גדולה של טקסט ב"מכה אחת" לא יהיה איכותי.

אלגוריתמים לסיכום טקסט

אנחנו נציג שני אלגוריתמים נפוצים לסיכום טקסט:

- **Refine Summarization** - כל קטע מסוכם ברצף, והסיכום מתעדכן תוך שילוב הקטע החדש. דרך זו מאפשרת שמירה על התפתחות הרעיון/העלילה לאורך הטקסט, אך היא מחייבת סיכום סידרתי.
- **MapReduce Summarization** - הטקסט מחולק לקטעים, כל קטע מסוכם בנפרד (Map). הסיכומים מאוחדים לסיכום כולל (Reduce). דרך זו מאפשרת עבודה מקבילה על טקסט גדול, אך עלולה לאבד הקשרים מתפתחים.
- ישנם אלגוריתמים נוספים ומתקדמים המנסים להתמודד עם הבעיות ולשפר ביצועים.

Refine Summarization



- בשיטה זו המודל מעדכן בכל פעם את הסיכום של הדפים.
- המודל מקבל אצווה של דפים וסיכום האצוות עד לשלב זה, ומוסיף את המידע באצווה לסיכום.

class RefineSummarizer

- **public RefineSummarizer(string language = "Hebrew") :**
 - בנאי היוצר את ה kernel ואת ה chat service.
- **private static List<string> ExtractPages(string path):**
 - הפעולה מקבלת נתיב של קובץ pdf וממחזירה רשימה של מחרוזות, כך שכל מחרוזת כוללת דף אחד.
- **private static List<string> MakePageWindows(List<string> pageTexts, int pagesPerChunk) :**
 - הפעולה מקבלת את רשימת הדפים ומחזירה רשימה של מחרוזות, כך שכל מחרוזת היא אצווה (chunk) של דפים בגודל pagesPerChunk.

המשך

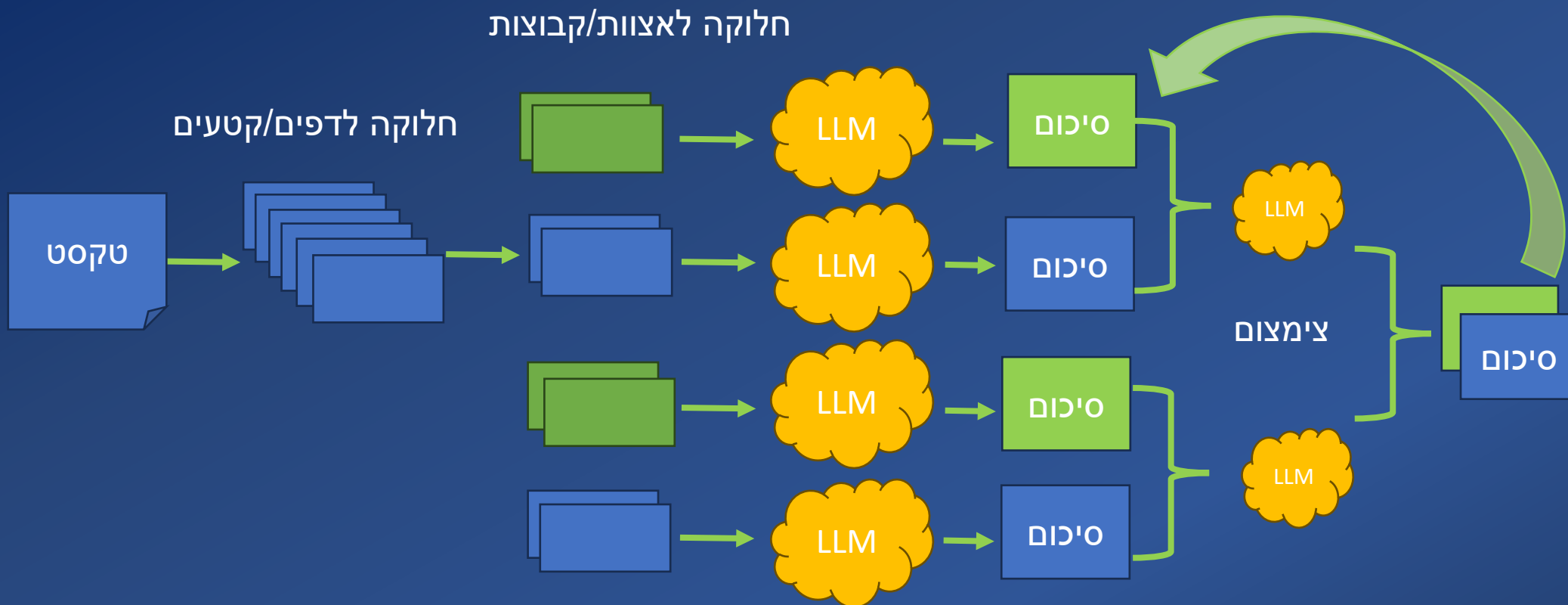
- `private async Task<string> RefineWithChunkAsync(string existingSummary, string newChunk):`

• הפעולה מקבל סיכום נוכחי ואצווה של דפים, שולחת אותם למודל בינה מלאכותית אשר מעדכן את הסיכום בהתאם למידע הנוסף הקיים באצווה.

- `public async Task<string> Summarize(string pdfPath, int pagesPerChunk = 5):`

• הפעולה הראשית המבצעת את הלולאה של `refine summarization`.

MapReduce Summarization



• שיטה זו מאפשרת לסכם מסמכים ארוכים באופן מקבילי.

class Summarizer

- המחלקה מממשת את השיטה Map Reduce Summerization.

- **public Summarizer(string language = "Hebrew"):**

- chat service ואת ה kernel הבנאי היוצר את ה

- **public async Task<string> Summarize(string pdfPath, int pagesPerChunk = 5, int groupSize = 5):**

- **private async Task<string> MapWindowAsync(string text):**

- פעולה המבצעת בעזרת מודל בינה מלאכותית את סיכום הטקסט. הפעולה מקבלת מחרוזת ומחזירה סיכום של המחרוזת.

- **private static List<string> ExtractPages(string path):**

- הפעולה מקבלת נתיב של קובץ pdf וממחזירה רשימה של מחרוזות, כך שכל מחרוזת כוללת דף אחד.

- `private static List<string> MakePageWindows(List<string> pageTexts, int pagesPerChunk):`

- הפעולה מקבלת את רשימת הדפים ומחזירה רשימה של מחרוזת, כך שכל מחרוזת היא אצווה (chunk) של דפים בגודל `pagesPerChunk`.

- `private async Task<string> ReduceManyAsync(List<string> parts, int groupSize):`

- הפעולה עוברת בלולאה על כל הסיכומים הנפרדים, מחלקת אותם לאצוות, ושולחת כל אצווה ל"צימצום".

- `private async Task<string> ReduceOnceAsync(IEnumerable<string> parts):`

- הפעולה מקבלת רשימה של סיכומים ו"מצמצמת" אותם לסיכום אחד.

