אופציה 2 לתרגיל בית מספר 5
ביצוע ניסוי מעבדה לחקר ה-RAG

Dr. Yoram Segal
כל הזכויות שמורות

26 בנובמבר 2025

תוכן העניינים

1 מבוא **2**

1.1 הקדמה 2

1.2 מטרת הניסוי 2

2 מבנה הניסוי: 3 תתי-ניסויים **2**

2.1 ניסוי 1: "הבעיה" -- חלון הקשר וגודל ההקשר 2

2.2 ניסוי 2: "האשליה" -- גלישה וחוסר רלוונטיות 2

2.3 ניסוי 3: "הפתרון" -- שימוש ב-RAG 3

2.4 שיפורים ושידרוגים לטיוב ניתוח הניסויים -- נקודות להתייחסות בכל אחד מהניסויים 3

3 דף הוראות לסטודנט **3**

3.1 שלב 1 -- ניסוי "קלאסי" ללא RAG 3

3.2 שלב 2 -- הוספת רעש 3

3.3 שלב 3 -- ביצוע RAG 4

4 דגשים לחלק המעשי **4**

4.1 Ollama 4

4.2 LangChain 4

5 טבלת תכנון הניסוי **4**

6 הצגת תוצאות הניסוי **5**

1 מבוא

מסמך זה מציג תכנון ניסוי ללימוד והדגמת השפעת RAG (Retrieval-Augmented Generation) על בעיות חלון ההקשר.

1.1 הקדמה

מטלה זו עוסקת בניתוח ולימוד של RAG. שורת הניסויים המוצעת להלן מהווה מסגרת רעיונית כללית, ואתם מוזמנים לפרש, לפתח ולחקור את הנושאים בכל דרך שתמצאו לנכון. עבור כל ניסוי עליכם להגדיר שאלות מחקר, לבצע את הניסויים ולנתח את הממצאים, רצוי תוך הצגת ניתוח סטטיסטי ויזואלי (באמצעות גרפים או טבלאות). מומלץ לחזור על כל ניסוי מספר פעמים כדי להבטיח תוקף סטטיסטי לתוצאה.

שימו לב: המסקנות שלכם אינן חייבות לחפוף לחומר שהוצג בשיעור. אתם רשאים להגיע לתובנות עצמאיות, ובלבד שתנמקו אותן היטב; במקרים אלו מומלץ להיעזר בסימוכין חיצוניים ולהציע הסבר לפערים שגיליתם. קחו את הניסויים למקום שמעניין אתכם ולכיוון החקירה האישי שלכם -- ההנחיות הללו נועדו לשמש כ``סיעור מוחות" ואינן בגדר הגדרות סגורות.

1.2 מטרת הניסוי

להמחיש לסטודנט, בקוד Python קצר, מדוע נוצרות בעיות "איבוד מידע" (lost in the middle) והזיות במודל שפה כאשר הקשר ארוך מדי או בלתי רלוונטי -- וכיצד RAG פותר זאת, על-ידי אחזור ממוקד.

2 מבנה הניסוי: 3 תתי-ניסויים

2.1 ניסוי 1: "הבעיה" -- חלון הקשר וגודל ההקשר

-- **יצירת תוכן סינתטי:** שורת עובדות פשוטות, לדוגמה: ["Paris is the capital of France.", "Berlin is the capital of Germany.", ...] (אפשר 02 עובדות).

-- **שיוך שאלה:** כל פעם מוסיפים עובדות חדשות ל"קשר" ובודקים האם המודל מצליח לענות על שאלת עובדה שמוסתרת באמצע/בסוף/בהתחלה ("מהי בירת גרמניה?").

-- **מדידה:** אחוז הדיוק/הצלחת שליפת המידע לפי מיקום העובדה והאורך.

2.2 ניסוי 2: "האשליה" -- גלישה וחוסר רלוונטיות

-- **הוספת "רעש":** לרשימת העובדות נוסיף משפטים לא רלוונטיים או הרבה "פילרים".

-- **בדיקה:** שוב נבקש את המודל לענות -- האם ההצלחה יורדת כי הקשר "רווי" ומבלבל?

-- **מדידה:** איכות התשובה, הופעת "הזיה" או דיוק ירוד.

2.3 ניסוי 3: "הפתרון" -- שימוש ב-RAG

-- **הטמעה באמצעות Embeddings:** כל עובדה מקבלת embedding (על-פי nomic-embed-text ב-Ollama או Sentence Transformers עם LangChain).

-- **בניית Vector DB:** שימוש ב-Chroma/FAISS וכדומה -- לאחזור אנלוגי בסגנון RAG.

-- **שאילתה:** שואלים את אותה שאלה, אך הקוד שולף רק את העובדה הכי רלוונטית מה-Vector DB.

-- **מדידה:** האם עכשיו גם בשדה "רועש", RAG מוצא (כמעט תמיד) את העובדה הרלוונטית?

2.4 שיפורים ושידרוגים לטיוב ניתוח הניסויים -- נקודות להתייחסות בכל אחד מהניסויים

-- **Reranking:** דירוג משופר בעזרת מודל LLM.

-- **Contextual Retrieval:** הוספת הקשר לפסקאות או הגדרה מושכלת.

-- **טעויות RAG נפוצות:** החסרת עובדה, מסמך גדול מדי, מידע "נתקע" בהקשר.

3 דף הוראות לסטודנט

3.1 שלב 1 -- ניסוי "קלאסי" ללא RAG

1. צור 02--03 "מסמכי עובדה" סינתטיים.
2. שמור את כולם כחיבור-טקסט ארוך אחד.
3. הגש למודל בעזרת ollama.generate או chat והצג שאלות כמודגם לעיל.
4. בדוק: האם המידע באמצע או בסוף נשאר זמין ונשלף?

3.2 שלב 2 -- הוספת רעש

1. הוסף טקסטים רנדומליים באורך רב.
2. נסה שוב לשאול. בדוק: איך הביצועים?

3.3 שלב 3 -- ביצוע RAG

1. עבור כל עובדה: הפק embedding (Ollama/LangChain).
2. בנה Vector DB (Chroma/FAISS וכדומה).
3. בעת שאילתה: שאל את רשת הוקטורים, החזר קטע רלוונטי, ורק אותו תן למודל השפה כקשר.
4. מדוד: האם איכות התשובה משתפרת? האם המודל פחות מזייף?

4 דגשים לחלק המעשי

4.1 Ollama

השתמש בפקודות כמו ollama.embeddings, ollama.generate, ובמידת הצורך Chroma DB ב-Python.

4.2 LangChain

הפעל את ההתממשקות עם Chroma, EmbeddingModels, Chain ו-Retriever המובנים.
הערה חשובה: בכל צעד -- מדדו במפורש את מהירות המענה, אורך הקשר ו-**דיוק התשובה**.

5 טבלת תכנון הניסוי

יוסין-תת	סינוב המ	סילאוש המ	האוושה/הייפיצ
יוסין 1	תודבוע דורא דמסמ	הדבוע לע הלאש X סוקימב	דרוי קויד יוס/עצמאב
יוסין 2	+ שער אלמ דמסמ תודבוע	הדבועה לע הלאש	תויזה, קויד תדירי
יוסין 3 (RAG)	+ Embeddings ירוטקו רוזחא	רוזחא -> הלאש החילש, הדבועה ל-LLM	ההובג החלצה, רתוי רהמ

טבלה 1: טבלת תכנון תתי-הניסויים

6 הצגת תוצאות הניסוי

על הסטודנט לתכנן ולחשוב באיזה אופן משכנע להציג את תוצאות הניסוי, החקר של הניסוי, והמסקנות מהניסוי. מומלץ לתקף את התוצאות בגרפים לפי שיקול דעת הסטודנט.

הערה: האמור במסמך זה מיועד לנשים ולגברים כאחד, והשימוש בלשון זכר הוא מטעמי נוחות בלבד.