תסריט להתאמנות — מה שאני אומרת, ברצף

שקופית 9 — ניתוב

"כאן אנחנו בודקים את השאלה המרכזית שלנו: האם אפשר לחזות את ה-queue מתוך הטקסט בלבד. בדקנו את זה על 25% מהדאטה שלא שימש לאימון — 7,145 כרטיסים."

"בדקנו ארבע שיטות. הראשונה, majority baseline, היא בלי שום אינטליגנציה — פשוט מנחשים תמיד את ה-queue הכי נפוץ, Technical Support, וזה נותן 29.2%. זה קיים רק בשביל שתהיה לנו רצפה להשוואה.

השנייה, keyword rules — חוקים כתובים ביד, למשל אם מופיעה המילה 'refund' מנחשים Billing. זה נותן 27.4%, כלומר בפועל פחות מה-baseline, כי התאמות מילים מקריות מטעות יותר משהן עוזרות.

השלישית, k-NN על embeddings — הופכים את הכרטיס לוקטור של 384 מספרים שמייצג את המשמעות שלו, מוצאים את הכרטיסים הכי דומים לו שכבר מתויגים, ומצביעים ברוב על ה-queue שלהם. זה נותן 64.6% דיוק.

והרביעית, linear SVM על TF-IDF — נותן 64.8%, הכי טוב מכולן. בשונה מ-k-NN, הוא לא משווה לכרטיסים אחרים בזמן הניחוש. הוא לומד פעם אחת מראש אילו דפוסי-מילים הכי מפרידים כל queue מהשאר, ואז בזמן הניחוש זה רק חישוב מתמטי — לא השוואה לכלום."

"לא רק החלנו embeddings והנחנו שהם ינצחו — מדדנו אותם מול שלושה baselines. ההפתעה היא ששיטה קלאסית, SVM לינארי על תדירויות מילים, מנצחת בקושי את ה-embeddings הנוירוניים כאן — 64.8% מול 64.6%. זו התובנה האמיתית: embeddings כן מוסיפים ערך אמיתי מעבר לניחוש קבוע או חוקי מילות-מפתח, אבל הם לא בהכרח מנצחים כל חלופה קלאסית — והדרך היחידה לדעת היא למדוד."

שקופית 10 — שליפה

"בדקנו גם את החיפוש עצמו. BM25 עובד לפי התאמת מילים, semantic search עובד לפי דמיון בין embeddings."

"לפני המספרים — מה זה MRR? זה מודד כמה רחוק צריך לגלול כדי להגיע לתוצאה הראשונה שבאמת רלוונטית. אם היא במקום הראשון, הציון הוא 1. אם היא במקום השלישי, הציון הוא שליש. וזה פשוט הממוצע של זה על פני כל 250 השאילתות שבדקנו.

חשוב להבין ש-precision ו-MRR שואלות שאלות שונות. Precision שואלת מתוך 5 תוצאות כמה טובות, MRR שואלת כמה מהר מצאתי את הראשונה הטובה. אז שיטה יכולה לנצח באחת ולהפסיד בשנייה — וזה בדיוק מה שקרה: סמנטי מנצח ב-precision אבל מפסיד ב-MRR, כי הוא מביא יותר תוצאות טובות בסך הכל, אבל לא תמיד שם את הכי טובה במקום הראשון."

"BM25, החיפוש המילולי, מנקד כרטיס לפי כמה מילות-שאילתה הוא מכיל — אבל לא באופן שווה: מילה נדירה שווה הרבה יותר ממילה שכיחה. הוא לא מבין משמעות, רק אם המילים תואמות.

חיפוש סמנטי הופך את השאילתה לוקטור באותו מרחב כמו כל כרטיס, ומדרג לפי דמיון קוסינוס — כלומר תוצאה יכולה לדרג ראשונה גם בלי אף מילה משותפת עם השאילתה, אם המשמעות קרובה.

וההיברידי משלב את שני הדירוגים — לא את הציונים עצמם, כי הם לא ניתנים להשוואה ישירה — כך שכרטיס שמדורג טוב באחת השיטות, ובטח בשתיהן, עולה למעלה."

"סמנטי מנצח ב-precision — הוא מוצא את הכרטיס הנכון יותר פעמים — אבל ל-BM25 עדיין MRR מעט גבוה יותר, כלומר כשהוא כן צודק, הוא נוטה לדרג את זה

ראשון. במקום לבחור מנצח, שילבנו את שני הדירוגים עם RRF, וזה ניצח את שניהם בנפרד: precision@5 של 0.758 ו-MRR של 0.876."

שקופית 11 — RAG

"בנוסף לחיפוש עצמו, בנינו RAG. המשתמש שואל שאלה בשפה טבעית, המערכת קודם מביאה כרטיסים רלוונטיים, ורק אחר כך מייצרת תשובה על בסיס ה-context הזה.

זה קורה בארבעה צעדים. שואלים שאלה רגילה — למשל מה התלונות הכי נפוצות על חיוב. המערכת קודם מחפשת, בדיוק כמו בשקופית הקודמת, ומביאה את חמשת הכרטיסים הכי רלוונטיים. רק אז נכתבת תשובה, שחייבת להתבסס רק על חמשת הכרטיסים האלה — לא על שום ידע כללי אחר. ולבסוף, בדיקת אמינות: כל מזהה כרטיס שהתשובה מציינת נבדק מול מה שבאמת נשלף, וכל מזהה שלא תואם נמחק לפני שמישהו רואה אותו."

"אחרי שהמודל עונה, המערכת בודקת כל ticket ID שהוא מצטט מול מה שבאמת נשלף — וכל ID שלא תואם נמחק לפני שהוא מגיע למשתמש."

ההדגמה החיה

```
פותחים http://localhost:8000/docs
```

> "זו ה-API החי — 17,354 כרטיסים אמיתיים כבר באינדקס."

```
מריצים GET /compare, שאילתה: security incident
```

> "אותה שאילתה, שלושה דירוגים. התוצאה המובילה של מילולי היא כרטיס גרמני שמתאים רק כי הוא מכיל את המילה 'Incident' — לא באמת קשור לאבטחה. התוצאה המובילה של סמנטי לא חולקת אף מילה עם השאילתה והיא הנושא הנכון."

```
מריצים POST /route, טקסט: "Our production database has been unreachable for two hours and customers cannot check out"
```

> "64.8% דיוק מול 29.2% של ניחוש קבוע — המספרים משקופית 9, בזמן אמת, על טקסט שהמודל מעולם לא ראה."

```
מריצים POST /ask, שאלה: "What are the most common billing complaints?"
```

> "זו בדיקת הציטוטים משקופית 11 — כל ID כאן אמיתי, אפשר לבדוק בשדה citations."

```
מריצים GET /kpis
```

> "כרטיסים מתויגים 'גרמנית' שגויים ב-27.2% מהזמן מול 0.05% ל'אנגלית' — פגם אמיתי וחד-כיווני בדאטה. ו-Service Outages הוא 4% מהנפח אבל 71% high priority — איוש לפי נפח היה מרעיב את זה משמעותית."