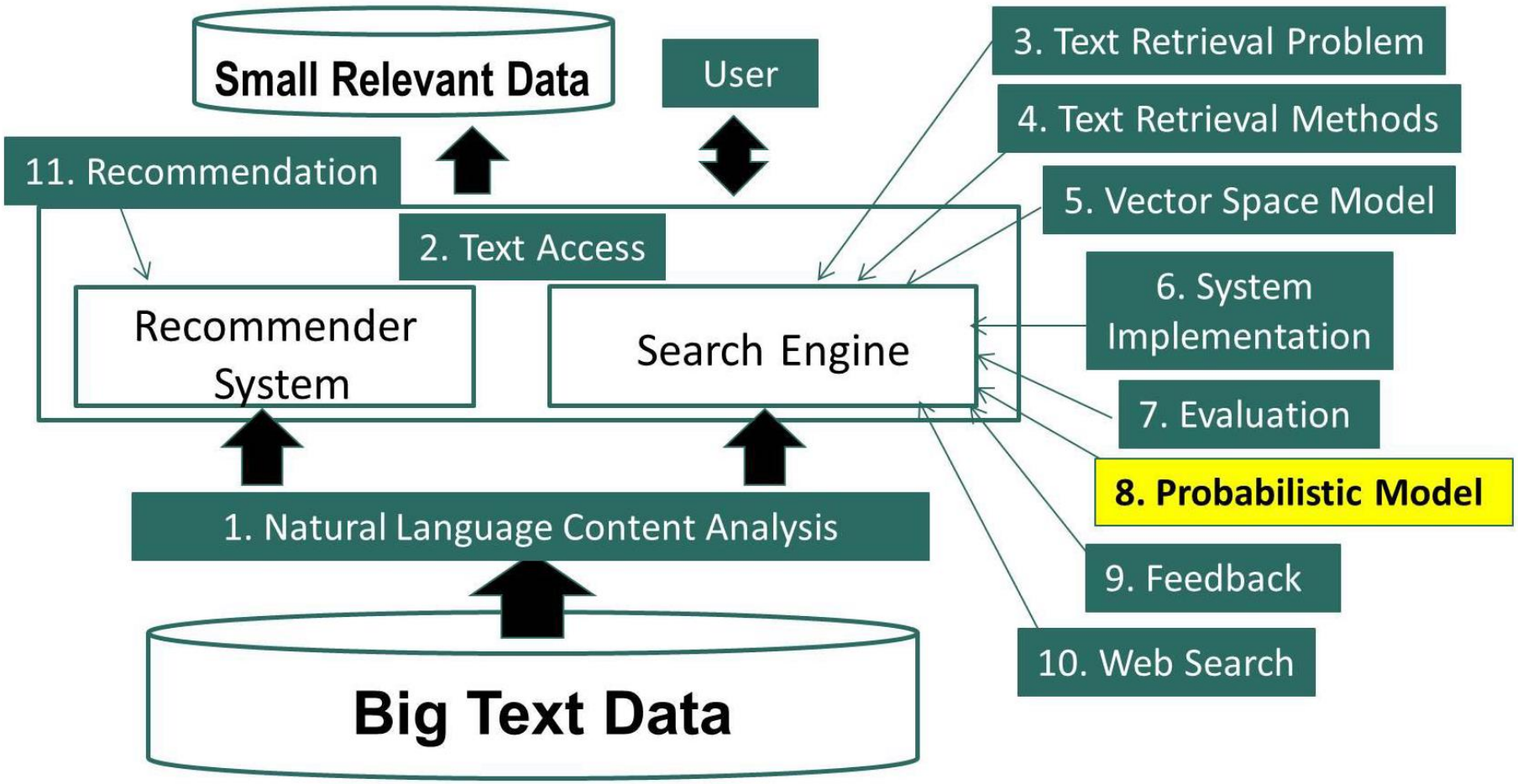
מודל אחזור הסתברותי


Small Relevant Data
User
3. Text Retrieval Problem
4. Text Retrieval Methods
11. Recommendation
5. Vector Space Model
2. Text Access
6. System Implementation
Recommender System
Search Engine
7. Evaluation
8. Probabilistic Model
1. Natural Language Content Analysis
9. Feedback
10. Web Search
Big Text Data

שיטות אחזור מידע שונות

כדי לדעת כן/לא

- מודל הסתברותי: $f(d,q) = p(R=1|d,q),\ R \in \{0,1\}$
  - מעריכים את ההסתברות של הרלוונטיות של המסמך לשאילתה.
  - BM25 הוא מודל הסתברותי קלאסי
  - מודל שפתי, ספציפית, סבירות שאילתות (Query Likelihood) במודל זה אנו נתמקד

כשאנחנו כותבים שאילתא יש לנו כוונה למסמך ולכן אנו כותבים את המילים שיודע, אלינו בלי...

מילים בלשון הציבורית.

- ב- QL אנו נעריך את ההסתברות $p(R=1|d,q) \approx p(q|d,R=1)$

הסיכוי שהמסמך מה הכי כשאילתא

אם משתמש מעוניין במסמך d מה הסיכוי שהמשתמש יזין את השאילתה q (על מנת לאחזר את המסמך d)?

מודל אחזור הסתברותי (הרעיון המרכזי)

מדגם אותם רלוונטיים מדגם שני לפי

לאיזה שני אנשים שונים.

Query (q)	Doc (d)	Relevance (R)
q1	d1	**1**
q1	d2	1
q1	d3	0
q1	d4	0
q1	d5	1
…	…	…
q1	d1	**0**
q1	d2	1
q1	d3	0
q2	d3	1
q3	d1	1
q4	d2	1
q4	d3	0

$$f(d,q) = p(R=1|d,q) = \frac{count(q,d,R=1)}{count(q,d)}$$

P(R=1|q1,d1) = ? 1/2

P(R=1|q1,d2) = ? 2/2

P(R=1|q1,d3) = ? 0/2

איך נתמודד עם מסמכים/שאילתות שלא נראו/הוזנו? (בעייתי, לא מתמודד עם בעיה זו, עדיין מציג את הרעיון הבסיסי) נצטרך להעריך בדרך כלשהי

בזמן שיש לנו את כל המידע הנ"ל, אפשר לחשב את ההסתברות לרלוונטיות.

מודל אחזור Query Likelihood

Query (q)	Doc (d)	Relevance (R)
q1	d1	**1**
q1	d2	1
q1	d3	0
q1	d4	0
q1	d5	1
…	…	…
q1	d1	**0**
q1	d2	1
q1	d3	0
q2	d3	1
q3	d1	1
q4	d2	1
q4	d3	0

המשתמש מעוניין במסמך

$$f(d,q) = p(R=1|d,q) \approx p(\mathbf{q}|d,R=1)$$

מה ההסתברות שהמשתמש
יזין את השאילתה q

הנחה: משתמש בונה שאילתה על
סמך "מסמך רלוונטי בעיני-רוחו"

מהו ככל-הנראה "מסמך רלוונטי בעיני-רוחו"

q = **“news about presidential campaign”**

דוגמא:

מהו ככל-הנראה "מסמך רלוונטי בעיני-רוחו"

q = "**news about presidential campaign**"

d1 … **news about** …

d2 … **news about** organic food **campaign**…

d3 … **news** of **presidential campaign** …

d4 … **news** of **presidential campaign** …
… **presidential** candidate …

d5 … **news** of organic food **campaign**…
campaign…**campaign**…**campaign**…

p(q|d1)

p(q|d2)

p(q|d4)

p(q|d5)

נניח שהמשתמש מעוניין
במסמך d4 מה הסיכוי
שהשאילתה תהיה q?

מהו ככל-הנראה "מסמך רלוונטי בעיני-רוחו"

q = "**news about presidential campaign**"

d1 … **news about** …

d2 … **news about** organic food **campaign**…

d3 … **news** of **presidential campaign** …

d4 … **news** of **presidential campaign** …
… **presidential** candidate …

d5 … **news** of organic food **campaign**…
campaign…**campaign**…**campaign**…

p(q|d1)

p(q|d2)

p(q|d4)

p(q|d5)

נניח שהמשתמש מעוניין
במסמך d5 מה הסיכוי
שהשאילתה תהיה q?

מהו ככל-הנראה "מסמך רלוונטי בעיני-רוחו"

q = "**news about presidential campaign**"

d1 … **news about** …

d2 … **news about** organic food **campaign**…

d3 … **news** of **presidential campaign** …

d4 … **news** of **presidential campaign** …
… **presidential** candidate …

d5 … **news** of organic food **campaign**…
campaign…**campaign**…**campaign**…

p(q|d1)

p(q|d2)

p(q|d4)

p(q|d5)

נניח שהמשתמש מעוניין
במסמך d4 מה הסיכוי
שהשאילתה תהיה q?

נניח שהמשתמש מעוניין
במסמך d5 מה הסיכוי
שהשאילתה תהיה q?

סיכום ביניים

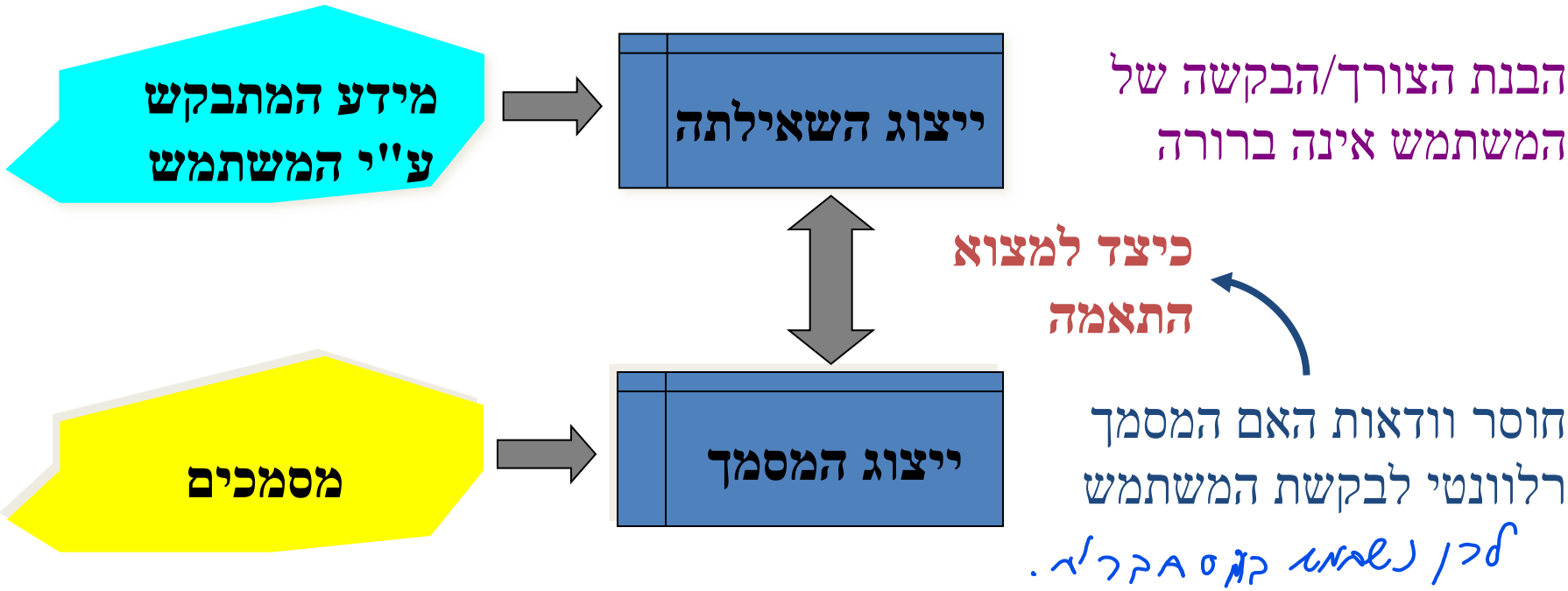
- הרעיון הכללי של המודל ההסתברותי, אנו מניחים שיש ערך אקראי R "וסביבו" בונים את פונקציית הדירוג.
- בגלל בעיות מסויימות אנו משתמשים בפונקציית קירוב, ההסתברות של q, ספציפית, בהינתן המסמך (כך לכל אחד מהמסמכים הרלוונטיים)
- Relevance(q,d) = p(R=1|q,d) → p(q|d,R=1)
- פונקציית הדירוג, **Query likelihood**: f(q,d)=p(q|d)
- אבל, כיצד נחשב p(q|d)? כיצד מחשבים הסתברות של טקסט בכלל?, לצורך זה יש את המודל השפתי, Language Model

p(q="presidential campaign"**|d=**"...**news** of **presidential campaign**
...**presidential** candidate ... " **)**

מה הסיכוי שהוא יזין
את השאילתה הזו

ספציפית, אם המשתמש
מעוניין במסמך זה

מדוע להשתמש במודל הסתברותי בשביל IR



במערכות IR מסורתיות, מנסים להתאים בין כל מסמך לשאילתה כאשר מרחב מונחי האינדקס במרחב אינו ברור לחלוטין.

ההסתברויות בבסיסה "נכנסת לעבודה" במצבי אי וודאות.
האם אנו יכולים להשתמש בהסתברויות בכדי לכמת את אי וודאות החיפוש שלנו?

נושאים הסתברותיים ב- IR

- מודל הסתברותי קלאסי של אחזור מידע
  - עקרון הדירוג ההסתברותי
  - Binary independence model (≈ Naïve Bayes)
  - BM25 (Okapi)
- רשתות בייסיאניות לאחזור טקסט
  - בהנתן אירוע שהתרחש, ניבוי הסבירות של כל אחת מכמה סיבות אפשריות גרם לאירוע.
- מודל שפתי ל- IR
- *שיטות הסתברותיות הן אחד הנושאים הוותיקים ביותר אך גם אחד הנושאים החמים כרגע ב- IR*
  - *מסורתית: רעיונות פשוטים/מדוייקים, אך בעבר לא הציגו את הביצועים הטובים ביותר*
    - *היום המצב שונה*

בעיית דירוג המסמכים

- יש בידינו מאגר של מסמכים
- המשתמש מזין שאילתה
- צריך להחזיר רשימה של מסמכים
- **שיטת דירוג מסמכים חיונית למערכות IR מודרניות**
  - **באיזה סדר נציג למשתמש את התוצאות?**
  - אנו מעוניינים שהתוצאות הטובות ביותר יהיו ראשונות
- **אידיאלי, דרג מסמכים בהתאם לרלוונטיות של מסמך בהתאם לשאילתה**
  - $P(R=1|document_i, query)$