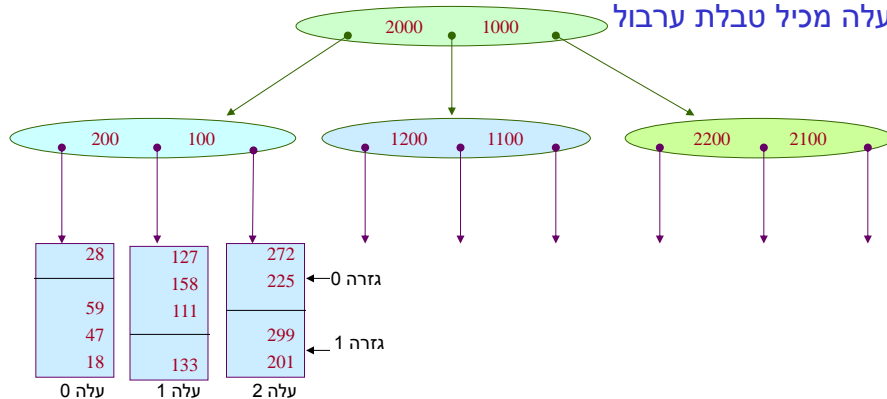


מבני אינדקס נוספים

קבצי אי-סדר חסום: Bounded Disorder (BD)

צירוף של עץ B^+ וטבלת ערבול
 האינדקס הוא עץ B^+ , כך שכל האינדקס נשאר בזיכרון הראשי
 הרשומות יושבות בעלים שגודלם מסילה
 כל עלה מכיל טבלת ערבול



קבצי אי-סדר חסום: Bounded Disorder (BD)

צירוף של עץ B^+ וטבלת ערבול
 האינדקס הוא עץ B^+ , כך שכל האינדקס נשאר בזיכרון הראשי
 הרשומות יושבות בעלים שגודלם מסילה
 כל עלה מכיל טבלת ערבול

פעולה	עצי B^+	ערבול	BD
חיפוש	2 גישות	גישה אחת	גישה אחת
מעבר סדרתי	קל	יקר	קל
תת-תחום	קל	יקר	קל
גודל הקובץ	גמיש	קבוע מראש (EH גמיש)	גמיש

אירגון טבלת הערבול

ההסתברות שגזרה מסוימת תגלוש: P_1
 ההסתברות שבעלה יש גזרה שגלשה (כאשר m מספר הגזרות בעלה):

$$1 - (1 - P_1)^m = mP_1 - \binom{m}{2} P_1^2 + \dots$$

אינה זניחה, ולכן

- כאשר **גזרה** גולשת, נשים את הרשומות העודפות בגרורה
- נפצל עלה רק כאשר מספר הרשומות **בעלה** גדול מקיבול **העלה**

הערות:

- האזור העיקרי של עלה וגרורותיו יימצאו באותו הגליל
- כדי לחסוך מקום אפשר לשתף גרורות

חיפוש רשומה עם מפתח X

1. בצע חיפוש בעץ B ומצא את כתובת העלה L אשר מכיל את x
2. אם כתובת הגזרה הראשונה של L היא a ו-L מכיל g גזרות, כתובת הגזרה של x היא $s = a + (h(x) \bmod g)$
3. אם x לא נמצא ו-s מלא, המשך את החיפוש בגרורה של s



לרוב, גישת דיסק בודדת

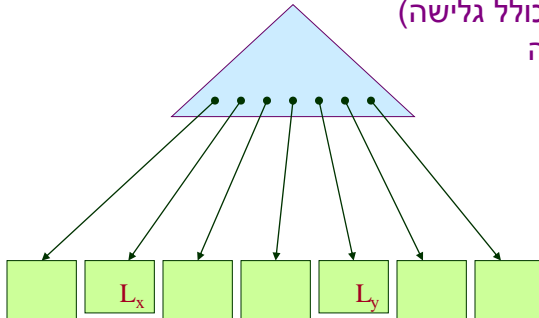
הכנסת רשומה X

1. חפש את x בעץ B, ותגיע לגזרה s של עלה L
2. אם נותר מקום עבור x, הכנס וסיים
3. אחרת — הוסף את x לגרורה של s
4. אם מספר רשומות < גודל עלה: פצל את העלה – רשומות L תחולקנה בין שני עלים: מצא רשומה שערכה הוא החציון, פצל לפי רשומה זו, והוסף עלה לאינדקס (לעץ B).

חיפוש טווח $[x, y]$

1. יהי L_x העלה של x ו- L_y העלה של y
2. עבור על כל העלים $L_x \leq L \leq L_y$ לפי הסדר
עבור כל עלה:

קרא את כל הסקטורים (כולל גלישה)
מיין את כל רשומות העלה



קובץ הפוך (Inverted File) אינדקס הפוך (Inverted Index)

- מבנה עזר לחיפוש טקסט עבור מילות מפתח:
- נתונות מילות מפתח, מצא את המסמכים הרלוונטיים



TF-IDF

(Term Frequency-Inverse Document Frequency)

• השפעות על חיפוש:

- (1) ככל שמילת חיפוש מופיעה יותר פעמים במסמך כך המסמך יותר רלוונטי
- (2) מילים שכיחות משפיעות פחות ממילים נדירות



משקל TF-IDF

• שיטה אחת מני רבות לדירוג מסמכים כתשובה לשאלת חיפוש

– term frequency – שכיחות מילה במסמך

– המילה cat מופיעה 3 פעמים במסמך בן 100 מילים

$$TF = 3/100 = 0.03$$

– inverse document frequency – ההפכי של שכיחות המילה בכלל

המסמכים

– המילה cat מופיעה ב-1,000 מסמכים מתוך 10,000,000

$$IDF = \log(10,000,000/1,000) = 4$$

• TF-IDF weight – המשקל שמקבל המסמך בחיפוש של המילה

$$TF-IDF = TF * IDF = 0.12$$



חישוב TF-IDF

• האינפורמציה הדרושה לחישוב:

- מספר המסמכים הכללי
 - מספר המלים בכל מסמך
 - מספר המסמכים בהם מופיעה המילה X
 - מספר המופעים של המילה X בכל מסמך
- ↔ קל לחישוב ולאחסון באינדקס רגיל
- ↔ מורכב לחישוב, מאוחסן באינדקס הפוך

דוגמה לאינדקס הפוך

• טקסט :

1 6 12 16 18 25 29 36 40 45 54 58 66 70

That house has a garden. The garden has many flowers. The flowers are beautiful.

• קובץ הפוך (אינדקס הפוך)

Vocabulary	Occurrences
beautiful	70
flowers	45, 58
garden	18, 29
house	6

קבצים הפוכים עבור TF-IDF

Vocabulary entry

k	d_k	
-----	-------	--

Posting file entry

doc_1	f_{1k}	doc_2	f_{2k}	...
---------	----------	---------	----------	-----

d_k : document frequency of term k

doc_i : i -th document that contains term k

f_{ik} : term frequency of term k in document i

cat	1000	
-----	------	--



doc_1	3	...	doc_{1000}	f_{1000k}
---------	---	-----	--------------	-------------

הקובץ ההופכי

docs	$t1$	$t2$	$t3$
D1	1	0	1
D2	1	0	0
D3	0	1	1
D4	1	0	0
D5	1	1	1
D6	1	1	0
D7	0	1	0
D8	0	1	0
D9	0	0	1
D10	0	1	1



Term.	D1	D2	D3	D4	D5	D6	D7	...
$t1$	1	1	0	1	1	1	0	
$t2$	0	0	1	0	1	1	1	
$t3$	1	0	1	0	1	0	0	

יצירת הקובץ ההופכי

לכל מילה, שומרים את
מזהה המסמך המכיל
אותה

Doc 1

Now is the time
for all good men
to come to the aid
of their country

Doc 2

It was a dark and
stormy night in
the country
manor. The time
was past midnight



Term	Doc #
now	1
is	1
the	1
time	1
for	1
all	1
good	1
men	1
to	1
come	1
to	1
the	1
aid	1
of	1
their	1
country	1
it	2
was	2
a	2
dark	2
and	2
stormy	2
night	2
in	2
the	2
country	2
manor	2
the	2
time	2
was	2
past	2
midnight	2

מפעילים מיון

Term	Doc #	Term	Doc #
now	1	a	2
is	1	aid	1
the	1	all	1
time	1	and	2
for	1	come	1
all	1	country	1
good	1	country	2
men	1	dark	2
to	1	for	1
come	1	good	1
to	1	in	2
the	1	is	1
aid	1	it	2
of	1	manor	2
their	1	men	1
country	1	midnight	2
it	2	night	2
was	2	now	1
a	2	of	1
dark	2	past	2
and	2	stormy	2
stormy	2	the	1
night	2	the	1
in	2	the	2
the	2	the	2
country	2	their	1
manor	2	time	1
the	2	time	2
time	2	to	1
was	2	to	1
past	2	was	2
midnight	2	was	2

ממזגים רשומות
כפולות ושומרים
תדירות הופעות

Term	Doc #	Term	Doc #	Freq
a	2	a	2	1
aid	1	aid	1	1
all	1	all	1	1
and	2	and	2	1
come	1	come	1	1
country	1	country	1	1
country	2	country	2	1
dark	2	dark	2	1
for	1	for	1	1
good	1	good	1	1
in	2	in	2	1
is	1	is	1	1
it	2	it	2	1
manor	2	manor	2	1
men	1	men	1	1
midnight	2	midnight	2	1
night	2	night	2	1
now	1	now	1	1
of	1	of	1	1
past	2	past	2	1
stormy	2	stormy	2	1
the	1	the	1	2
the	1	the	2	2
the	2	the	1	2
their	1	their	2	1
time	1	time	1	1
time	2	time	2	1
to	1	to	1	2
to	1	to	2	2
was	2	was	2	2
was	2	was	2	2

Term	Doc #	Freq
a	2	1
aid	1	1
all	1	1
and	2	1
come	1	1
country	1	1
country	2	1
dark	2	1
for	1	1
good	1	1
in	2	1
is	1	1
it	2	1
manor	2	1
men	1	1
midnight	2	1
night	2	1
now	1	1
of	1	1
past	2	1
stormy	2	1
the	1	2
the	2	2
their	1	1
time	1	1
time	2	1
to	1	2
was	2	2

Dictionary

Postings

Term	N docs	Tot Freq	Doc #	Freq
a	1	1	2	1
aid	1	1	1	1
all	1	1	1	1
and	1	1	2	1
come	1	1	1	1
country	2	2	1	1
country	2	2	2	1
dark	1	1	2	1
for	1	1	1	1
good	1	1	1	1
in	1	1	2	1
is	1	1	1	1
it	1	1	2	1
manor	1	1	2	1
men	1	1	2	1
midnight	1	1	1	1
midnight	1	1	2	1
night	1	1	2	1
now	1	1	1	1
of	1	1	1	1
past	1	1	2	1
stormy	1	1	2	1
the	2	4	1	2
the	2	4	2	2
their	1	1	1	1
time	2	2	1	1
time	2	2	2	1
to	1	2	1	2
was	1	2	2	2

Dictionary

Term	N docs	Tot Freq
a	1	1
aid	1	1
all	1	1
and	1	1
come	1	1
country	2	2
dark	1	1
for	1	1
good	1	1
in	1	1
is	1	1
it	1	1
manor	1	1
men	1	1
midnight	1	1
night	1	1
now	1	1
of	1	1
past	1	1
stormy	1	1
the	2	4
their	1	1
time	2	2
to	1	2
was	1	2

Postings

Doc #	Freq
2	1
1	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	2
2	2
1	1
2	1
1	2
2	2

חיפוש של

"time" AND "dark"

שני המסמכים מכילים
את המילה "time" אך
רק מסמך מספר 2 מכיל
את המילה "dark"

Dictionary

Term	N docs	Tot Freq
a	1	1
aid	1	1
all	1	1
and	1	1
come	1	1
country	2	2
dark	1	1
for	1	1
good	1	1
in	1	1
is	1	1
it	1	1
manor	1	1
men	1	1
midnight	1	1
night	1	1
now	1	1
of	1	1
past	1	1
stormy	1	1
the	2	4
their	1	1
time	2	2
to	1	2
was	1	2

Postings

Doc #	Freq
2	1
1	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	1
2	1
1	2
2	2
1	1
2	1
1	2
2	2

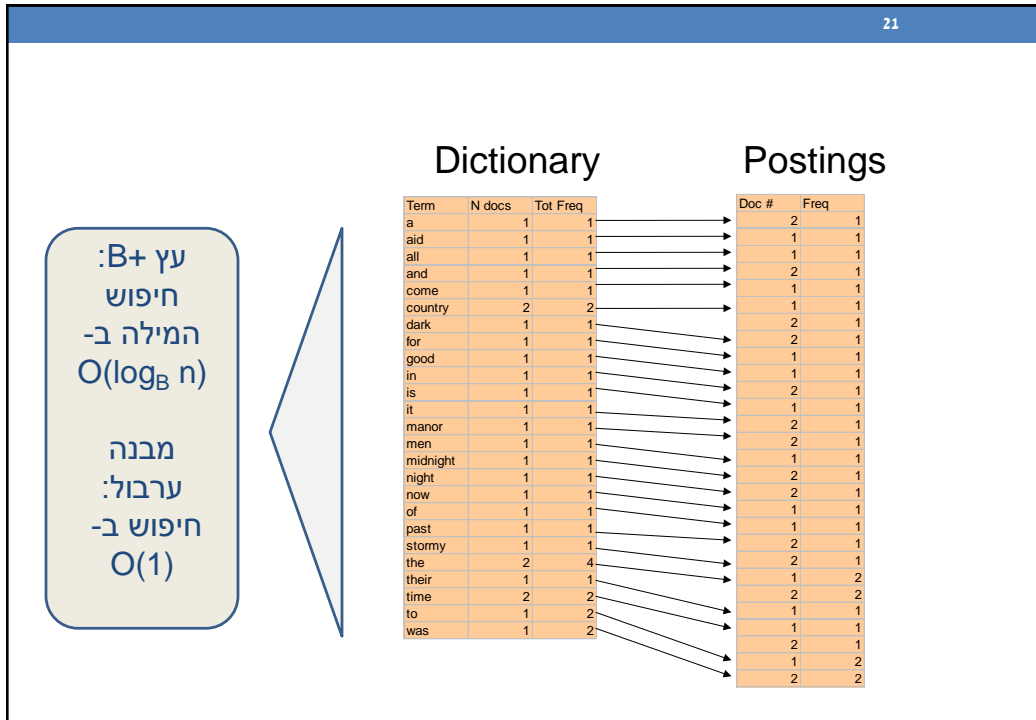
חישוב TF-IDF עבור

"time" AND "dark"

• אחת השיטות: בחיפוש מספר
מלים סוכמים את המשקל של כולן

$$Doc_1: \underbrace{\frac{1}{16} \times \log \frac{2}{2}}_{\text{time}} + \underbrace{\frac{0}{16} \times \log \frac{2}{1}}_{\text{dark}} = 0$$

$$Doc_2: \underbrace{\frac{1}{16} \times \log \frac{2}{2}}_{\text{time}} + \underbrace{\frac{1}{16} \times \log \frac{2}{1}}_{\text{dark}} = \log 2$$



חישוב בעזרת MapReduce

• שלב ראשון: חישוב Term frequency

- Map: (docname, contents) → ((word, docname), 1)
- Reduce: ((word, docname), list(1)) → list((word, docname), f)

• f הוא מספר המופעים של word ב-document

• תזכורת: לפני ה-Reduce() מתבצע מיון

• על פי מה כדאי לבצע את ה-Shuffle (חלוקה ל-reducers)?

חישוב בעזרת MapReduce

• שלב שני: חישוב Document frequency

- Map: $((\text{word}, \text{docname}), f) \rightarrow (\text{word}, (\text{docname}, f, 1))$
- Reduce: $(\text{word}, \text{list}(\text{docname}, f, 1)) \rightarrow \text{list}((\text{word}, \text{docname}), (f, d))$

• d הוא מספר המסמכים בהם מופיעה word

• למה f מופיע בקלט של שתי הפונקציות?

• את תוצאת ה-Reduce() ניתן להכניס ישירות למבנה האינדקס

• ניתן לשמור את מספר המלים במסמך במבנה נפרד או בתוך האינדקס ההפוך

סיכום ביניים

