Introduction to

Information Retrieval

معرفی بازیابی اطلاعات

علی قنبری

مرجع درس

- **An Introduction to Information Retrieval**, Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Cambridge University Press, Cambridge, England.

ارزیابی

- میانترم ۵ نمره
- پایانترم ۱۲ نمره
- تمرین ۳ نمره
- پروژه ؟ نمره
- حضور و غیاب ۱ نمره

Ali.ghanbari289@gmail.com

Introduction to

Information Retrieval

فصل اول: بازیابی بولی

علی قنبری

تعریف بازیابی اطلاعات

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

بازیابی اطلاعات یافتن مواد (معمولا اسناد) از یک ماهیت بدون ساختار (معمولا متن) است که یک نیاز اطلاعاتی را داخل مجموعه‌های بزرگ (که معمولا در کامپیوتر ذخیره می‌شوند) برآورده می‌کند.

5

مثالی از بازیابی اطلاعات

- مجموعه کتب شکسپیر شامل کتاب‌های قطوری است که بسیاری از افراد آن را دارند. فرض کنید می‌خواهید تعیین کنید که کدام نمایش شکسپیر شامل کلمات Brutus و Caesar است و شامل Calpurnia نیست.
- یک راه این است که از ابتدای کتاب شروع کنید و خط به خط متن را بخوانید و هر نمایشی که شامل اسم‌های Brutus و Caesar بود، ثبت کرده و نمایشی که اسم Calpurnia را داشت، مستثنی کنید. ساده‌ترین صورت بازیابی اسناد برای کامپیوتر، این نوع اسکن خطی در میان اسناد است.

مثالی از بازیابی اطلاعات

- راه جلوگیری از اسکن خطی متون برای هر پرس و جو این است که اسناد را از پیش شاخص گذاری کنیم.

- فرض کنید که برای هر سند(در اینجا نمایش‌های شکسپیر) ثبت می‌کنیم که آیا شامل هر یک از کلماتی است که شکسپیر به کار برده است یا خیر(حدود ۳۲۰۰۰ کلمه مختلف شکسپیر به کار برده است).

- نتیجه عمل بالا یک ماتریس تلاقی دودویی عبارت – سند است.

مثالی از ماتریس تلاقی

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
. . .						

یک به معنای وجود کلمه در نمایشنامه است و صفر به معنای عدم وجود کلمه در نمایشنامه است.

بازیابی بولی

- مدل بولی، یکی از ساده ترین مدل‌های بازیابی اطلاعات است.
- Query ها عبارات بولی مانند CAESAR AND BRUTUS هستند.
- موتورهای جستجو همه اسنادی را برمیگرداند که عبارت بولین query برای آنها درست است.

- بسته به اینکه به سطر یا ستون ماتریس نگاه کنیم می‌توانیم برداری برای هر عبارت داشته باشیم که اسنادی را که در آن پدیدار می‌شوند، نشان می‌دهد. یا آنکه برداری برای هر سند داشته باشیم که عباراتی را که در آن ظاهر شده اند، نشان می‌دهد.

9

بردار وقوع (Incidence vectors)

- ✓ برای پاسخ به پرسش BRUTUS AND CAESAR AND NOT CALPURNIA
  - بردارهای متناظر با BRUTUS, CAESAR ,CALPURNIA را برمی‌داریم.
  - مکمل بردار CALPURNIA را محاسبه می‌کنیم.
  - عملگر and بیتی روی سه بردار انجام می‌دهیم.

110100 AND 110111 AND 101111 = 100100

10

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
. . .						
result:	1	0	0	1	0	0

11

پاسخ به پرسش

Anthony and Cleopatra, Act III, Scene ii

Agrippa [Aside to Domitius Enobarbus]: Why, Enobarbus,
When Antony found Julius Caesar dead,
He cried almost to roaring; and he wept
When at Philippi he found Brutus slain.

Hamlet, Act III, Scene ii

Lord Polonius: I did enact Julius Caesar: I was killed i'
the Capitol; Brutus killed me.

12

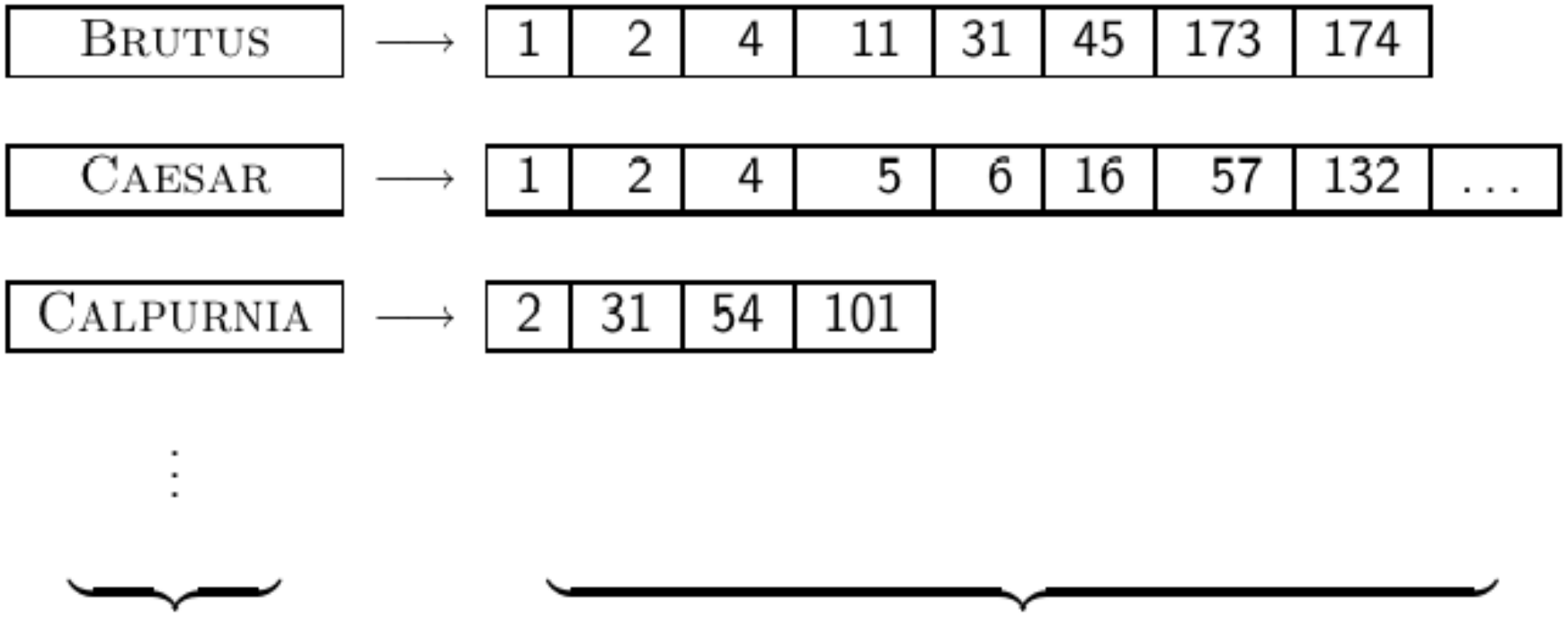
مجموعه‌های بزرگ

- تعداد $N = 10^6$ متن با حدود 1000 کلمه(۲ تا ۳ صفحه کتاب) را در نظر بگیرید.
- اگر هر کلمه ۶ بایت حافظه بخواهد، مجموعه اسناد حدود ۶ گیگابایت خواهد بود.
- فرض کنیم 500’000 کلمه مجزا در این اسناد وجود داشته باشد.
- ماتریس تلاقی شامل $500'000 * 10^6$ عنصر خواهد بود.
- می‌توان نشان داد که حداقل 99.8% عناصر این ماتریس صفر است.
- این ماتریس **تنک** است.
- نمایش بهتر از ماتریس تلاقی آن است که فقط عناصر یک را ذخیره کنیم.

13

شاخص معکوس

برای هر کلمه، لیستی از همه اسنادی که شامل آن کلمه هستند را ذخیره می‌کنیم.



14

ساخت شاخص معکوس

1. جمع‌آوری اسنادی که باید شاخص گذاری شوند.

Friends, Romans, countrymen. | So let it be with Caesar | . . .

2. تبدیل هر سند به لیستی از نشانه‌ها

Friends | Romans | countrymen | So | . . .

3. پیش پردازش، تولید لیستی از نشانه‌های نرمال شده

friend | roman | countryman | so | . . .

4. شاخص گذاری اسنادی که هر عبارت در آن رخ می‌دهد. (به کمک ایجاد شاخص وارونه شامل لغت نامه و پست‌ها)

مرحله دوم و سوم

Doc 1. I did enact Julius Caesar: I was killed i' the Capitol; Brutus killed me.
Doc 2. So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious:

⟹

Doc 1. i did enact julius caesar i was killed i' the capitol brutus killed me
Doc 2. so let it be with caesar the noble brutus hath told you caesar was ambitious

16

مرحله چهارم – گام ۱

Doc 1. i did enact julius caesar i was killed i' the capitol brutus killed me
Doc 2. so let it be with caesar the noble brutus hath told you caesar was ambitious

term	docID
i	1
did	1
enact	1
julius	1
caesar	1
i	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

- درون یک مجموعه سند، فرض می‌کنیم که هر سند یک شماره سریال منحصر به فرد به نام شناسه سند دارد.
- در طول ساخت شاخص به راحتی می‌توانیم یک عدد صحیح متوالی را برای هر سند جدید، هنگامی که اولین بار با آن مواجه شدیم، اختصاص دهیم.

17

مرحله چهارم – گام ۲ (مرتب سازی بر اساس حروف الفبا)

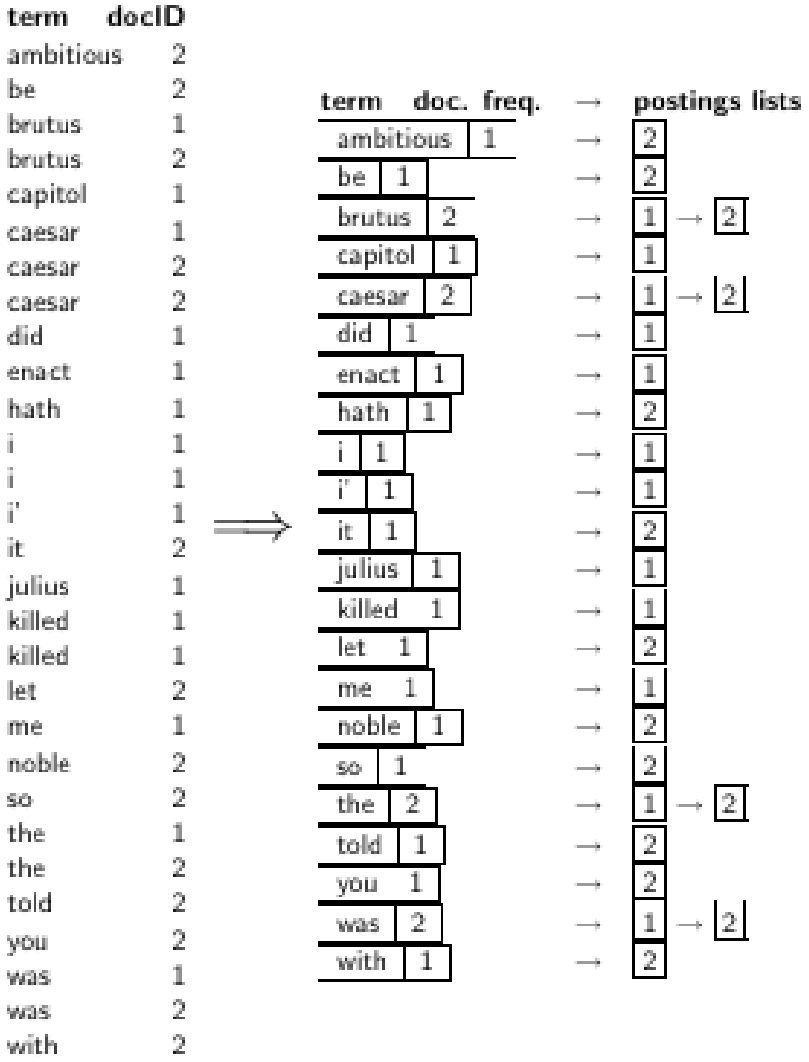
term	docID
i	1
did	1
enact	1
julius	1
caesar	1
i	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

⟹

term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
i	1
i	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

18

مرحله چهارم – گام ۳(ساخت لیست پست‌ها و تعداد تکرار آن‌ها)



19

مثال

- شاخص وارونه‌ای برای مجموعه اسناد زیر رسم کنید.
- سند ۱:

new home sales top forecasts

- سند ۲:

home sales rise in july

- سند ۳:

increase in home sales in july

- سند ۴:

july new home sales rise

فرض کنید همه نشانه‌های به کار رفته در اسناد، نرمال هستند.

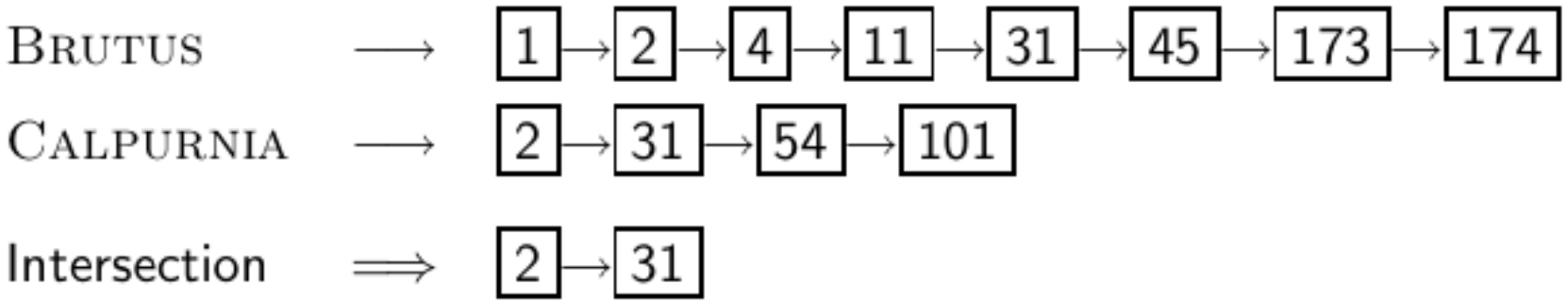
یک مثال از پرس و جوی عطفی ساده به کمک شاخص معکوس

- پرس و جوی BRUTUS AND CALPURNIA را در نظر بگیرید.
- برای پیدا کردن پاسخ پرس و جوی بالا به کمک شاخص معکوس:

۱. محل BRUTUS را در دیکشنری پیدا می‌کنیم.
۲. پست‌های آنرا بازیابی می‌کنیم.
۳. محل CALPURNIA را در دیکشنری پیدا می‌کنیم.
۴. پست‌های آنرا بازیابی می‌کنیم.
۵. اشتراک دو لیست پست را محاسبه می‌کنیم.

21

یک مثال از پرس و جوی عطفی ساده به کمک شاخص معکوس(ادامه)



22

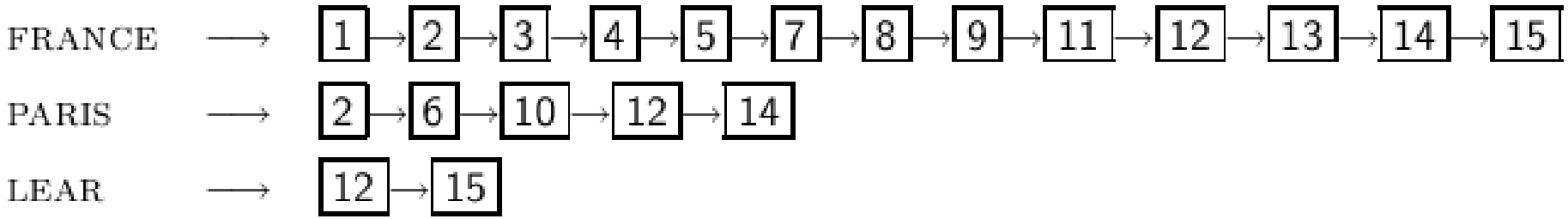
الگوریتم اشتراک دو لیست پست

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

این الگوریتم فقط زمانی درست است که لیست پست‌ها مرتب باشند.

23

مثال دیگر



نتیجه پرس و جوی ((paris AND NOT france) OR lear را محاسبه نمایید.

24