Introduction to

Information Retrieval

فصل دوم: مجموعه واژگان عبارات و لیست پست‌ها

علی قنبری سرخی

گام‌های عمده شاخص وارونه

- اسنادی را که باید شاخص گذاری شوند جمع آوری کنید.
- متن را نشانه گذاری کنید.
- از نظر زبان شناسی نشانه (token) را پیش پردازش کنید.
- اسنادی را که در آنها هر عبارت ذکر شده است، شاخص گذاری کنید.

در این فصل

- چگونگی تعیین واحد پایه یک سند و تعیین دنباله کاراکتری در یک سند (بخش ۲-۱)
- مسائل زبان‌شناسی در مورد نشانه‌گذاری و پردازش زبان‌شناسی، که مجموعه واژگان عبارات را تعیین می‌کند بررسی می‌شود (بخش ۲-۲)
- نشانه‌گذاری، روند تقسیم جریان کارکترها به نشانه‌هاست، و پیش‌پردازش زبان‌شناسی پس از آن به ساخت دسته‌های هم‌ارزی از نشانه‌ها که مجموعه عبارات شاخص‌گذاری شده هستند می‌پردازد.
- شاخص‌گذاری در فصل ۱ و ۴ ارائه می‌شود. سپس به پیاده‌سازی لیست پست‌ها باز می‌گردیم.
- در بخش ۲-۳، یک ساختمان داده برای لیست پست‌ها بسط یافته را خواهیم آزمود که پرس‌وجوی سریع‌تر را پشتیبانی می‌کند.
- در بخش ۲-۴ ساختمان داده پست‌ها را طوری که برای اداره پرس‌وجوهای مجاورت و عبارات مناسب باشد به طریقی که عموما در مدل‌های بولی بسط یافته و روی وب ظاهر می‌شود می‌سازیم.

دستیابی به دنباله کاراکتر به یک سند

- اسناد دیجیتال نوعا بایت‌های یک فایل یا روی سرویس دهنده وب هستند. اولین گام پردازش، تبدیل این دنباله بایت به دنباله خطی از کاراکترها است.
- به طور مثال برای متن انگلیسی ساده با کدگذاری ASCII این موضوع امکان‌پذیر است.
- برای متون نوشتاری پیچیده‌تر (شامل زبان‌های آسیایی) می‌توان از کدگذاری UTF8 استفاده کرد.
- برای تعیین فرمت سند می‌توان از روش‌های زیر استفاده کرد.
  - روش‌های دسته بندی یادگیری ماشین (فصل ۱۳)
  - استفاده از فراداده
  - انتخاب توسط کاربر
- هنگامی که روش کدگذاری تعیین شد دنباله‌ای از کدها به کارکترها تبدیل می‌شود.
- در اسناد مختلف ممکن است کدگذاری‌های دیگری نیز نیاز باشد مانند اسناد XML. چه کد گذاری‌هایی؟ (مثال &map)

انتخاب واحد سند

- به طور کلی برای اسناد طولانی، مسئله دانه‌بندی (تعیین مقیاس یا سطح جزئیات) شاخص گذاری اهمیت زیادی دارد.
- برای مجموعه کتاب‌ها، در نظر گرفتن کل کتاب به عنوان یک سند ایده خوبی نیست. در اینصورت جستجوی عبارت Chinese toys ممکن است ما را به کتابی برساند که china را در فصل اول و toys را در فصل آخر ذکر کرده است. طبیعتا این کتاب هیچ ارتباطی به مسئله ما ندارد.
- یک انتخاب خوب، انتخاب هر فصل یا پاراگراف به عنوان یک واحد سند است.
- چرا نمی‌توان به سطوح پایین‌تر رفت؟

نشانه گذاری

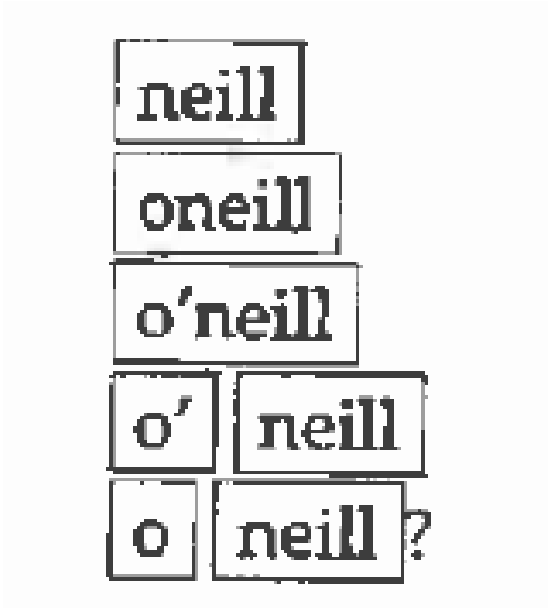
- با داشتن دنباله کاراکتر و واحد سند معین، نشانه‌گذاری شکستن دنباله کاراکتر به بخش‌های کوچک به نام نشانه است. همزمان با این عمل کاراکترهای خاص مانند علامت‌گذاری‌های نگارشی دور ریخته می‌شوند.

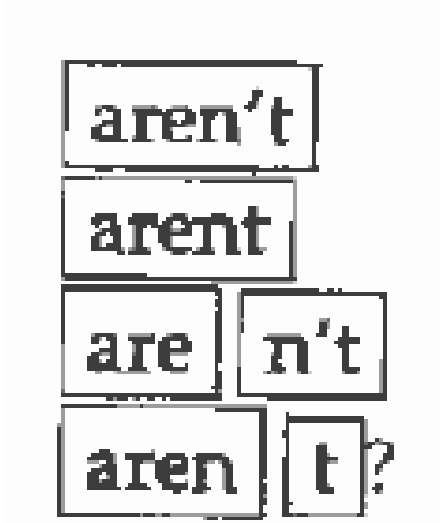
- Input: Friends, Romans, Countrymen, lend me your ears;
- Output: Friends Romans Countrymen lend me your ears
- سوال مهم در اینجا این است که نشانه‌های صحیح برای استفاده کدامند؟
  - در مثال بالا، ساده به نظر می‌رسد. چگونه این کار را انجام می‌دهیم؟

یک مثال دیگر

Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing.

- برای O’Neill و aren’t، کدامیک از نشانه‌گذاری های زیر مطلوب است؟





یک مثال دیگر(ادامه)

- یک استراتژی ساده این است که فقط تمامی کاراکترهای غیر حرف-رقم را جدا کنیم.

aren t o neill

- این روش برای aren’t بد است ولی برای O’neill به نظر انتخاب خوبی است.
- در مورد پرس و جوی o’neill AND capital کدام موارد انطباق دارد؟
- در مورد پرس و جوی neill AND capital کدام موارد انطباق دارد؟
- بنابراین باید نشانه گذاری متن و پرس و جوها یکسان باشد.
- مسائل نشانه گذاری خاص زبان هستند. بنابراین نیاز است که زبان سند شناخته شده باشد.

نشانه‌گذاری (ادامه)

- برای اکثر زبان‌ها نشانه‌های خاص غیر متعارفی وجود دارد که علاقمندیم آنها را به عنوان یک عبارت تشخیص دهیم. مانند C#، C++ و هواپیمای B-52.
- در کامپیوترها دنباله‌های کاراکتری متفاوتی دارد که تمایل داریم به صورت واحد در نظر گرفته شود.
  - آدرس ایمیل مانند alyan.nezhadi@gmail.com
  - آدرس سایت‌ها مانند http://du.ac.ir/
  - آدرس‌های IP مانند 172.16.2.14
- یک راه حل ممکن حذف آنها از نشانه‌های شاخص گذاری است. اینکار باعث محدود کردن مردم در جستجو کردن است که هزینه‌های زیادی دارد. به طور مثال می‌خواهیم به دنبال یک IP در متن بگردیم تا آنرا تغییر دهیم.

نشانه‌گذاری (ادامه)

- در انگلیسی خط تیره برای اهداف زیادی به کار می‌رود.
  - تقسیم صدا در کلمات مانند co-education
  - اتصال اسامی به عنوان یک کلمه مانند Hewlett-Packard
  - نمایش گروه بندی کلمات مانند hold-him-backand-drag-him-away
- در مورد اول باید نشانه گذاری به صورت coedution باشد. در مورد سومی باید کلمات جدا شوند و نهایتا در مورد دوم دقیقا مشخص نیست.
- جداسازی بر اساس فضای خالی می‌تواند آنچه که باید یک نشانه باشد را نیز جدا کند مانند Los Angeles، York University و New York University.

نشانه‌گذاری (ادامه)

- زبان آلمانی کلمات مرکب را بدون فاصله می‌نویسد. مانند

Computerlinguistik

lebensversicherungsgesellschaft

- سرحد محدودیات در زبان‌های آسیای شرقی(مانند چینی) است که متون بدون هیچ فاصله نوشته می‌شوند.

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

نشانه‌گذاری (ادامه)

- ابهام در قطعه بندی کلمات زبان چینی. دو کاراکتر را می‌توان به عنوان یک کلمه به معنای "راهب" در نظر گرفت. همچنین به عنوان دنباله‌ای دو کلمه معنای "و" یا "هنوز" در نظر گرفت.

12

حذف عبارات متعارف : کلمات توقف

- گاهی اوقات کلمات بسیار عام که ظاهرا ارزش اندکی در کمک به انتخاب اسناد منطبق با نیاز کاربر دارند، باید از مجموعه واژگان مستثنی شوند. این کلمات، کلمات توقف نامیده می‌شوند.

a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with

13

حذف عبارات متعارف : کلمات توقف (ادامه)

- استراتژی عمومی برای تعیین لیست توقف این است که عبارات بر اساس فراوانی مجموعه (تعداد کل دفعاتی که یک عبارت در مجموعه اسناد ظاهر می‌شود) مرتب می‌شود. سپس عبارات با بیشترین فراوانی (اغلب به صورت دستی انتخاب می‌شود) به عنوان لیست توقف انتخاب می‌شود.
- استفاده از این لیست باعث کاهش تعداد پست‌ها می‌شود.
- در بسیاری از موارد که کلمات توقف شاخص نمی‌شوند، آسیب اندک است. چه زمانی آسیب شدید است؟

14

حذف عبارات متعارف : کلمات توقف (ادامه)

- در موارد زیر حذف کلمات توقف آسیب زیادی به جستجو می‌زند.

President of United States

flights to London

As we may think

To be or not to be

- سیستم‌های بازیابی اطلاعات دارای لیست‌های توقف بزرگ (۳۰۰ – ۲۰۰ عبارت)، لیست‌های توقف کوچک (۱۲ – ۷ عبارت) و یا بدون لیست توقف هستند.
- موتورهای جستجوی وب از لیست توقف استفاده نمی‌کنند.

15

نرمالسازی (دسته کردن هم ارزی عبارات)

- موارد بسیاری وجود دارد که دو دنباله کاراکتر کاملا یکسان نیستند ولی شما تمایل دارید که تطبیق رخ دهد. مانند U.S.A. و USA
- نرمالسازی نشانه‌ها روند استاندارد سازی نشانه‌ها است بطوریکه تطبیق، علیرغم تفاوت‌های صوری در دنباله کاراکترها، رخ دهد.
- متداولترین روش این است که دسته‌های هم ارزی به طور ضمنی ایجاد کنیم. برای نمونه اگر نشانه‌های anti-virus و antivirus هر دو به عبارت antivirus نگاشت شود(هم در متن سند و هم در پرس و جوها)، جستجو برای یک عبارت، اسنادی که شامل دیگری است را شامل می‌شود.

16

نرمالسازی (ادامه)

- در بسیاری از زبان‌ها اعراب‌ها در سیستم نگارش مرسوم بوده و صداهای مختلف را متمایز می‌کنند.
- سوال مهم معمولا زبان شناسی نیست. سوال اصلی این است که کاربران چگونه این کلمات را جستجو می‌کنند.
- در این موارد بهترین راهکار حذف اعراب کلمات است.

17

نرمالسازی (ادامه)

- یک استراتژی رایج برای غیر حساس کردن به حروف کوچک و بزرگ، تبدیل تمامی حروف به حروف کوچک است.
- ایده بالا، اغلب خوب است. به طور مثال، این روش اجازه می‌دهد Automobile در ابتدای جمله‌ها با پرس و جوی automobile تطبیق کند. یا آنکه کاربر برای دسترسی به اطلاعات ماشین Ferrari می‌تواند ferrari جستجو کند که راحت‌تر است.
- آیا این روش همیشه خوب است؟

18

نرمالسازی (ادامه)

- بسیاری از اسامی خاص از اسامی عام مشتق شده اند و تنها راه تشخیص آنها حروف آنها است.
  - شرکت‌ها و سازمان‌ها مانند General Motors
  - اسامی افراد مانند Bush
- در زبان انگلیسی بهتر است تنها برخی حروف را کوچک کنیم. به طور مثال کلماتی که در ابتدای جمله هستند.
- کلماتی که در میان جمله هستند و با حروف بزرگ نوشته شده‌اند، نباید به حروف کوچک تبدیل شود.

19

نرمالسازی (ادامه)

- تاریخ‌ها، زمان‌ها و موارد مشابه در فرمت‌های گوناگون چالش‌های زیادی دارد.
- شما می‌خواهد تاریخ 3/12/91 و Mar. 12. 1991 یکسان در نظر گرفته شود.
- پردازش صحیح در اینجا پیچیده است.
- این تاریخ در آمریکا برابر Mar. 12. 1991 و در اروپا معادل 3. Dec. 1991 است.

20

ریشه گیری و مدخل گیری

- به دلایل دستورات زبانی صورت‌های مختلفی از کلماتی مانند organize، organizes و organizing استفاده می‌شود.
- هدف ریشه گیری و مدخل گیری کاهش صورت‌های صرفی و نحوی کلمه است تا به یک صورت پایه متعارف رسید.
  - *am, are, is → be*
  - *car, cars, car's, cars' → car*
- نگاشت جمله زیر را در نظر بگیرید.
  - *the boy's cars are different colors → the boy car be different color*

21

ریشه گیری و مدخل گیری (ادامه)

- ریشه‌گیری معمولا به فرایند مکاشفه‌ای خاصی اشاره دارد که اکثر اوقات انتهای کلمات را به امید رسیدن به ریشه کلمات قطع می‌کند. همچنین گاهی شامل حذف ضمیمه‌های نحوی است.
- مدخل گیری معمولا به انجام صحیح اموری در مورد کاربرد یک واژه و تحلیل ریخت شناسی کلمات اشاره دارد و معمولا به حذف انتهای صرفی کلمات و بازگشت به صورت لغت‌نامه‌ای کلمه می‌پردازد.
- اگر با نشانه saw مواجه شویم، ریشه گیری ممکن است s را برگرداند در حالیکه مدخل گیری see یا saw را بر می‌گرداند.

22

الگوریتم ریشه گیر Porter

- عمومی ترین الگوریتم ریشه گیری در زبان انگلیسی، الگوریتم ریشه گیر Porter است.
- این الگوریتم شامل ۵ مرحله کاهش کلمه است که به ترتیب به کار می‌رود.
- به طور مثال مرحله اول، شامل قوانین زیر است.

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

23

الگوریتم ریشه گیر Porter

- عمومی ترین الگوریتم ریشه گیری در زبان انگلیسی، الگوریتم ریشه گیر Porter است.
- این الگوریتم شامل ۵ مرحله کاهش کلمه است که به ترتیب به کار می‌رود.
- به طور مثال مرحله اول، شامل قوانین زیر است.

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

24

الگوریتم ریشه گیر Porter (ادامه)

- بسیاری از قوانین جدیدتر مفهوم اندازه کلمه را در نظر می‌گیرد.

Rule	Example
$(m > 1)EMENT \rightarrow$	$repalacement \rightarrow repalc$

- کلمه cement به c تغییر نمی‌یابد.

25

الگوریتم ریشه گیر Porter (ادامه)

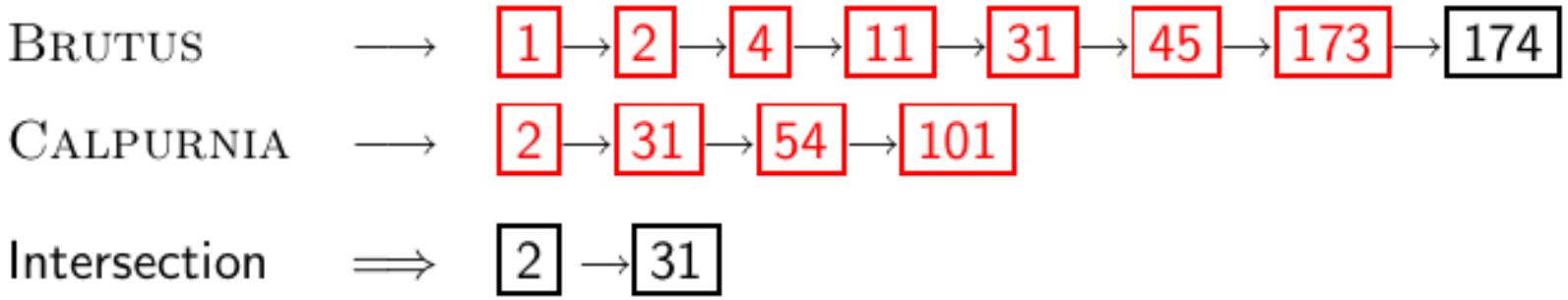
- اگر چه این روش به برخی پرس و جو ها کمک می‌کند ولی به برخی دیگر آسیب می‌زند.

Operate operating operates operation operational

- همه این کلمات را به oper کاهش می‌یابد.

26

اشتراک لیست پست‌ها

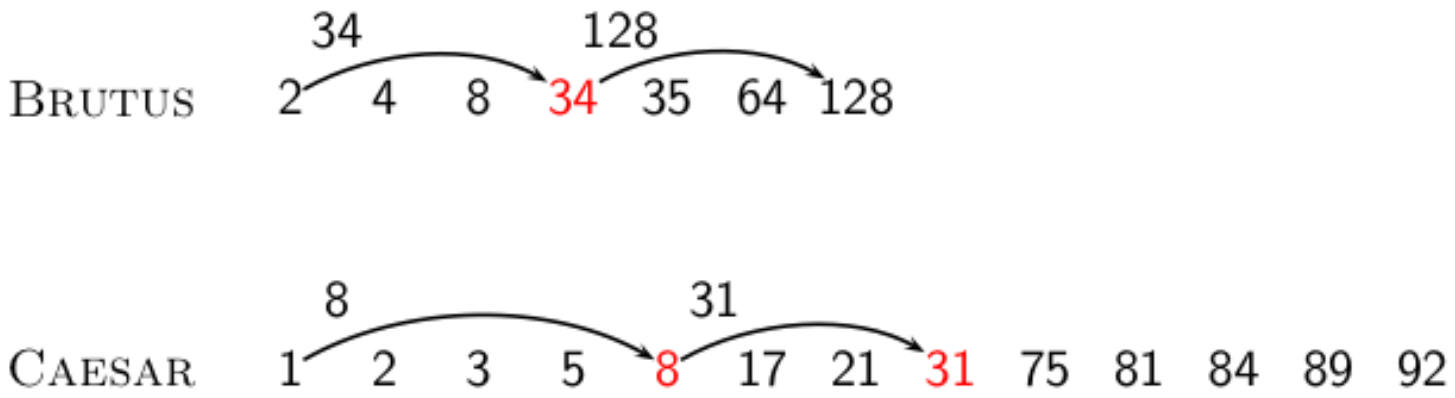


- پیچیدگی الگوریتم پایه اشتراک لیست پست‌ها (فصل ۱) به صورت $O(m+n)$ است.

- آیا می‌توان الگوریتم بهتری پیشنهاد داد؟

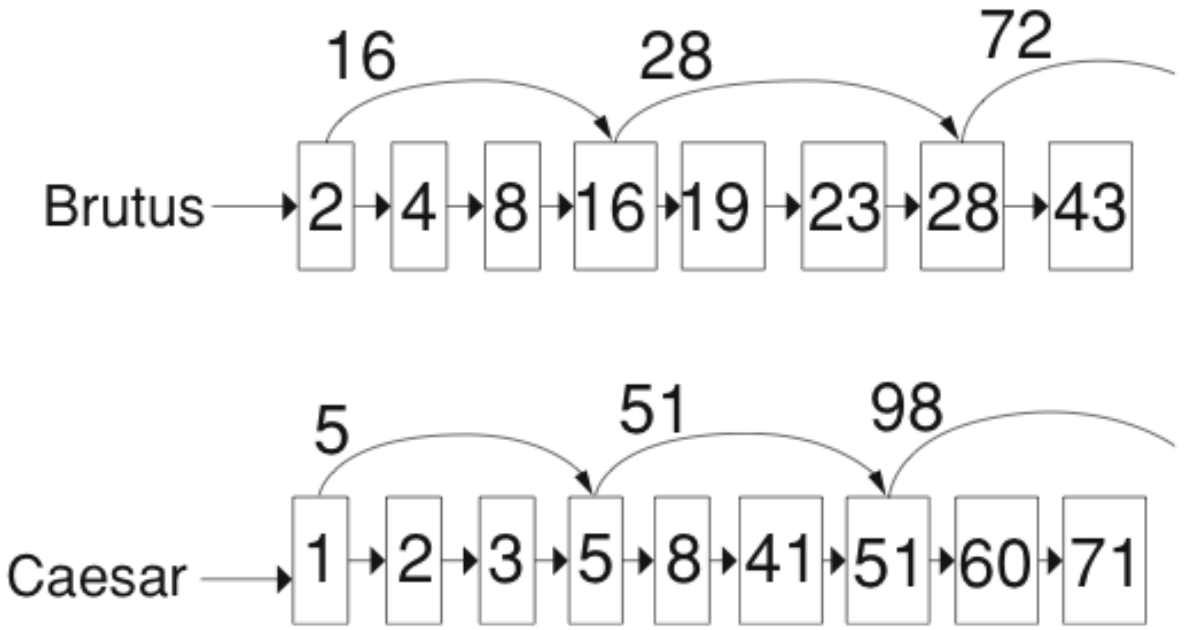
27

اشتراک سریع‌تر لیست پست‌ها



28

اشتراک سریع‌تر لیست پست‌ها (ادامه)



29

اشتراک سریع‌تر لیست پست‌ها (ادامه)

- یک راه برای افزایش سرعت، استفاده از لیست پرش است.

- اشاره‌گرهای پرش میانبرهای کارآمدی هستند که اجازه می‌دهند از پردازش بخش‌هایی از لیست پست‌ها که در نتایج جستجوها موثر نیستند، جلوگیری شود.

30

اشتراک سریع‌تر لیست پست‌ها (ادامه)

```
INTERSECTWITHSKIPS(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 if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```


31

پست‌های موقعیتی و پرس و جو های اصطلاح

- ما می‌خواهیم که پرس و جویی مانند Stanford University را با فرض اینکه اصطلاح است، ایجاد کنیم. در این پرس و جو نباید جمله زیر تطبیق یابد.

the inventor Stanford Ovshinsky never went to university

- بیشتر موتورهای جستجوی اخیرگیومه نقل قول مستقیم را بر پرس و جوی اصطلاح پیشتیبانی می‌کند.

32

پست‌های موقعیتی و پرس و جو های اصطلاح (ادامه)

- یک راهکار برای اداره اصطلاحات این است که هر جفت متوالی در سند را یک اصطلاح در نظر بگیریم. برای مثال متن Friends, Romans, Countrymen می‌تواند یک شاخص دو کلمه‌ای را تشکیل دهد.

Friends romans

Romans countrymen

- پردازش پرس و جو های دوکلمه‌ای عملی و سریع است.

33

پست‌های موقعیتی و پرس و جو های اصطلاح (ادامه)

- اصطلاحات طولانی می‌تواند با کوتاهتر شدن، پردازش شوند.

- پرس و جوی Stanford university palo alto را می‌توان به صورت زیر پردازش کرد.

“stanford university” AND “university palo” AND “palo alto”

- این پرس و جو می‌تواند در عمل به خوبی کار کند اما ممکن است مثبت‌های کاذب را نیز نتیجه دهد. چه زمانی این روش می‌تواند اشتباه پاسخ دهد؟

34