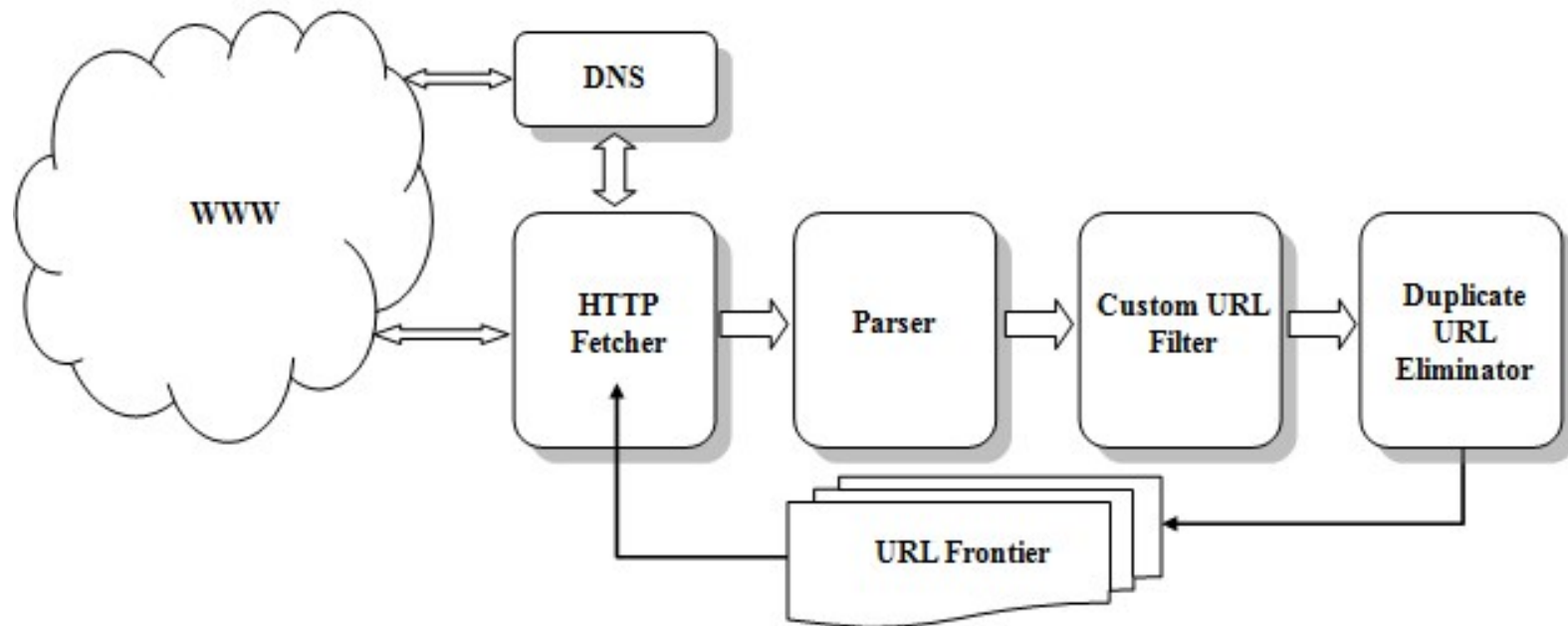


Web crawler

проф. д-р инж. Христо Вълчанов

<http://cs.tu-varna.bg>

Архитектура на Web crawler



Особености

- Налице е изчакване за получаване на отговора.
- Редуцирането на това време – нишки.
- Потенциално засипват сайтовете със заявки за страници (flooding).
- Избягване на това – изчакване между заявките към един и същ web сървър (*politeness policies*).

Пример за нишка на crawler

```
procedure CRAWLERTHREAD(frontier)
  while not frontier.done() do
    website ← frontier.nextSite()
    url ← website.nextURL()
    if website.permitsCrawl(url) then
      text ← retrieveURL(url)
      storeDocument(url, text)
      for each url in parse(text) do
        frontier.addURL(url)
      end for
    end if
    frontier.releaseSite(website)
  end while
end procedure
```

Актуалност на страниците (freshness)

- Използва се специална заявка на протокола HTTP – HEAD.

Client request: HEAD /csinfo/people.html HTTP/1.1
Host: www.cs.umass.edu

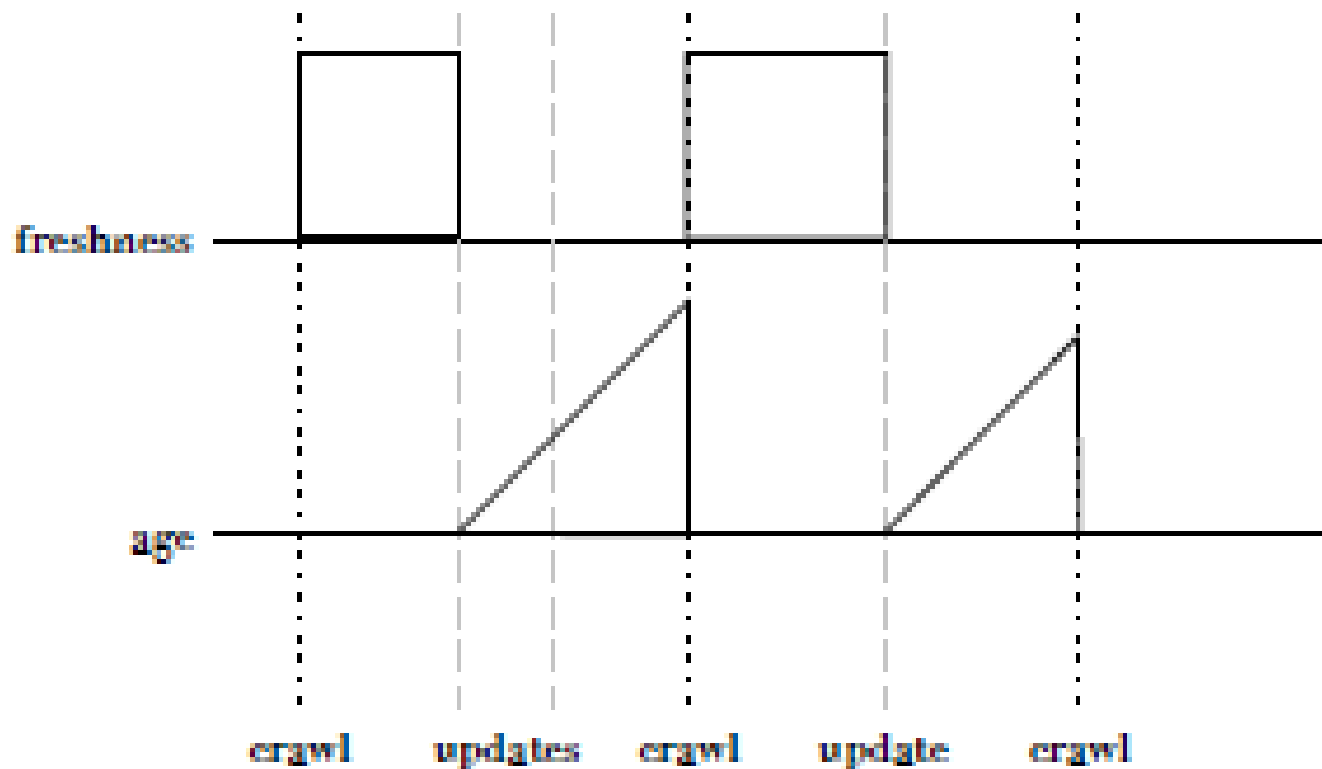
HTTP/1.1 200 OK
Date: Thu, 03 Apr 2008 05:17:54 GMT
Server: Apache/2.0.52 (CentOS)
Last-Modified: Fri, 04 Jan 2008 15:28:39 GMT
Server response: ETag: "239c33-2576-2a2837c0"
Accept-Ranges: bytes
Content-Length: 9590
Connection: close
Content-Type: text/html; charset=ISO-8859-1

Актуалност на страниците (freshness)

- Не е възможно постоянно да се проверяват всички страници.
 - Използва се *Freshness* – частта на страниците, които са обновени.
 - Метрика - Ако има по-ново копие на страницата, тя е обновена (fresh).
- Не всички страници се променят често.

Възраст на страниците (Age)

- По-добра метрика



Age метрика

- Очаквана възраст на страница t дена след последното и изтегляне (λ пъти за 1 ден)

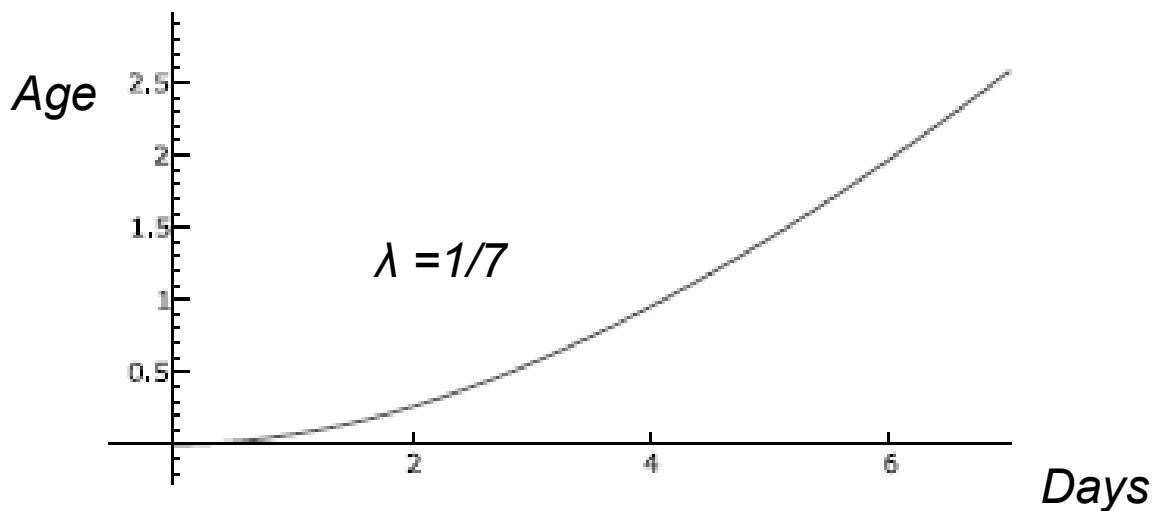
$$\text{Age}(\lambda, t) = \int_0^t P(\text{page changed at time } x)(t - x)dx$$

- Възрастта: $(t - x)$ – изтеглена във време t , но сменена в момент x .

Age метрика

- Промяната на страниците следват Поасоново разпределение.

$$\text{Age}(\lambda, t) = \int_0^t \lambda e^{-\lambda x} (t - x) dx$$



Тематичен crawling

- Фокусирано търсене по дадена тема (vertical crawling)
 - По-малки колекции;
 - Изтеглят се само страници, свързани с дадена тема;
 - Разчита се на факта, че страница от дадена тема ще съдържа линкове към страници от същата тема.
 - Необходимост от средства за определяне дали страница е по дадена тема - класификатори

Deep Web

- Трудни за извличане web сайтове:
 - Частни сайтове;
 - Резултати от форми;
 - Скриптови страници.

Sitemaps

- Съдържат списъци от URL и данни за тях.
- Създават се от администраторите на сайтове.
- Указват на crawler как да намери страници на сайт

Sitemaps

```
<url>
  <loc>http://www.company.com/</loc>
  <lastmod>2008-01-15</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.7</priority>
</url>
<url>
  <loc>http://www.company.com/items?item=truck</loc>
  <changefreq>weekly</changefreq>
</url>
<url>
  <loc>http://www.company.com/items?item=bicycle</loc>
  <changefreq>daily</changefreq>
</url>
</urlset>
```

Преобразуване

- Текстът се съхранява в множество различни несъвместими формати (raw, PDF, MSWord, HTML).
- Преобразуване в консистентен формат (тагов) - HTML, XML.

Прекодиране

- Различни схеми на кодиране на символите на азбуката.
- Китайски език – над 40000 символа, използвани често – около 3000.
- Много езици имат много различни схеми на кодиране (Chinese-Japanes-Korean).
- Не може да има много езици в един файл.

Прекодиране

- Unicode.
- UTF-8 – 1 байт за English и 4 байта за традиционен китайски.
- UTF-32 – 4 байта за всеки символ.
- Много приложения използват UTF-32 за представяне на текст (позволява бързо търсене) и UTF-8 за съхраняване на диск (по-малко памет).

Отстраняване на „шум“

- Web страници с информация, която не съответства на основното съдържание на страницата.
- Излишната информация – „шум“ (noise).
- Как да се открият блоковете с тази информация във web страница.

Пример за „шум“

The image is a screenshot of the CNN.com website from June 5, 2008. The main headline is "Aquarium plays whale shark matchmaker" under the "SCIENCE & SPACE" category. The article text describes how the Georgia Aquarium facilitated a meeting between two female whale sharks, Alice and Trixie, transported from Taiwan. A red rectangular box highlights the main body of the article, from the sub-headline "Two females flown 8,000 miles for double date in Atlanta" down to the paragraph about the aquarium's 6.2-million-gallon tank. An arrow points from the text "Content block" to this red box. The page includes a left sidebar with navigation links, a top search bar, and a right sidebar with a "Save up to 75% on Last-Minute Cruises" advertisement and an "E-MAIL ALERTS" sign-up form.

CNN.com Member Center: [Sign In](#) / [Register](#) International Edition

SEARCH 117 1077 CNN.COM Search

Home Page World U.S. Weather Business Sports Analysis Politics LAW Technology Science & Space Health Entertainment Crime Travel Education Special Reports Video Audio LReports

UPDATES: 11:00 AM
TALKSHOW

SERVICES: E-mail RSS Podcasts Mobile CNN Pipeline SEARCH

117 1077 1077 1077

SEARCH

SCIENCE & SPACE

Aquarium plays whale shark matchmaker

Two females flown 8,000 miles for double date in Atlanta

Monday, June 5, 2008, Posted 8:28 p.m. EDT (21:28 GMT)

ATLANTA, Georgia (CNN) — Ralph and Norton, meet Alice and Trixie.

The Georgia Aquarium's two male whale sharks got some female companionship on Saturday, when they were joined by two females transported to Atlanta from Taipei, Taiwan.

Researchers are hoping the sharks will mate.

The females — 11 feet and 14 feet long — were flown more than 8,000 miles by UPS, which reconfigured a company E-747 freighter with advanced marine life support systems to carry them. (Watch what I did to get the sharks together — 1:52)

The pilot said they treated the massive but like first-class passengers.

"As we were doing the descent, we asked to start down a little sooner to make a nice shallow descent, to not make things too uncomfortable back there for the whale sharks," UPS pilot Capt. Bob Crum said.

The plane's center of balance was carefully planned, according to a statement from the aquarium, and veterinarians accompanied the sharks.

The delivery company also brought the two males to Atlanta, where researchers can study the whale sharks' behavior, breeding and development.

The whale sharks — named after the male characters in the 1950s sitcom "The Honeymooners" — were delivered to the aquarium in special transportation containers.

The Georgia Aquarium, which opened in November, is the world's largest aquarium. It was a \$250 million gift to Georgia from Bernie Marcus, co-founder of The Home Depot and his wife, Bill, through the Marcus Foundation.

It is the only aquarium outside of Asia to showcase whale sharks, which are the largest fish on Earth.

The aquarium's 6.2-million-gallon "Ocean Voyager" tank can hold up to six whale sharks. (Watch what I did to get the sharks together — 1:52)

YOUR E-MAIL ALERTS

Atlanta (Georgia)

Taiwan

ACTIVATE or Create Your Own

Manage Alerts | What is This?

Save up to 75% on Last-Minute Cruises.

Vacations To Go.com

Best Price Guarantee

GO!

Story Tools

Print | Email | Facebook | Twitter | Digg | StumbleUpon | Reddit | LinkedIn | Delicious | Technorati | RSS

Subscribe to Time for \$5.00

SPACE

Section Page | Video

Automatically prepare for third spacewalk

Business chooses Putin's successor

TOP STORIES

Home Page | Video | Most Popular

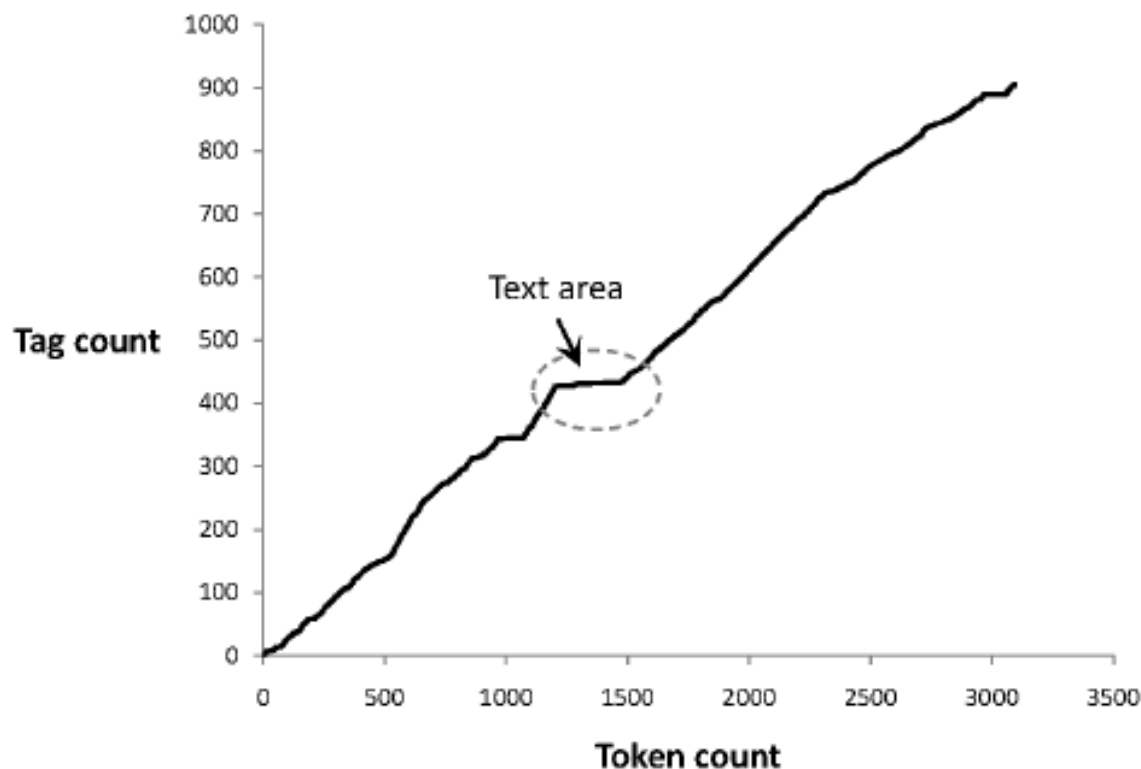
Business chooses Putin's successor

Automatically prepare for third spacewalk

Business chooses Putin's successor

Откриване на „шум“

Натрупване на тагове в страницата – основното съдържание на страницата съответства на равнинна част в средата на разпределението



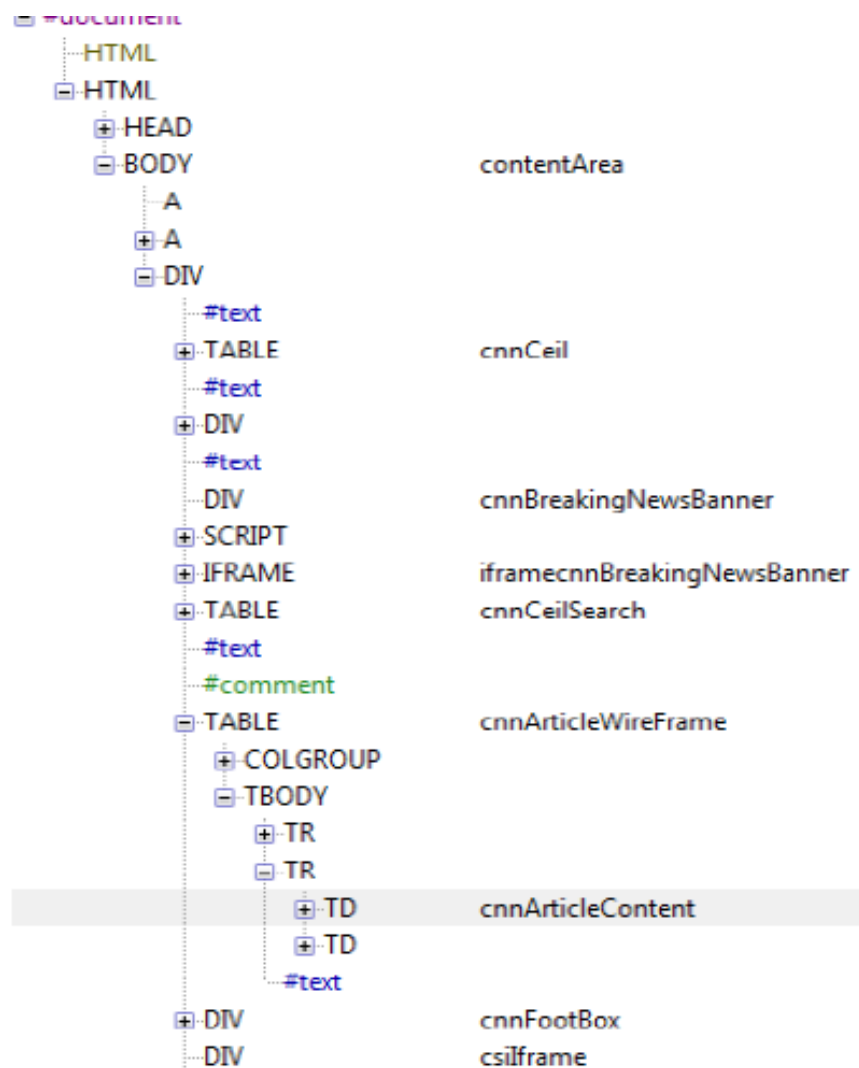
Откриване на „шум“

- Web страницата се представя като последователност от битове, където $b_n=1$ указва, че n -ят token е таг.
- Оптимизационна задача за откриване на стойности на i и j при които се максимизира броят на таговете под i и над j и броят на не-таговете токени между i и j .

$$\sum_{n=0}^{i-1} b_n + \sum_{n=i}^j (1 - b_n) + \sum_{n=j+1}^{N-1} b_n$$

Откриване на „шум“

- Изграждане на DOM структура и откриване на съдържанието



Обработка на текст

- Парсване на текста (*parsing*) – разпознаване на съдържанието и структурата на документите.
- Tokenizing (lexical analysis) – формиране на думи.
- Съдържание на документ:
 - Думи (tokens);
 - Метаданни (автор, тагове).
- Синтактичен анализ

Проблеми

- Кратки думи могат да бъдат важни в някои запитвания (*xp*, *pt*, *world war II*).
- Има думи с и без тирета:
 - В някои случаи не са необходими (*e-bay*, *active-x*, *cd-rom*).
 - В други са част от думата или разделител (*mazda rx-7*, *e-cards*, *t-mobile*)

Проблеми

- Специални символи, важни за тагове, URL и код.
- Главни букви, имащи различно значение (*Bush*, *Apple*).
- Апострофите могат да са част от дума, притежание или грешка (*can't*, *80's*, *o'donnell*).

Проблеми

- Числата имат важно значение (*porsche 911, top 10 courses, windows 10*).
- Точките могат да се появяват в числа, аббревиатури, URL, край на изречение (*I.B.M., Ph.D. , cs.tu-varna.bg*)

Stemming

- Алгоритмични
- Базирани на речници

- Suffix-s stemmer

cakes -> cake

dogs -> dog

Но: *century – centuries ?*

Porter stemmer

- Последователност от стъпки
- На всяка стъпка се премахват или заменят окончания.
- Заменя sses с ss (*stresses* -> *stress*).

Грешки при действие

<i>False positives</i>	<i>False negatives</i>
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

Krovetz stemmer

- Хибриден подход.
- Използва речник за определяне дали думата е валидна.
- Помощни списъци с деривационни суфикси.
- Предимство – корените в повечето случай са пълни думи.

Stemmers

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:

document describ market strategi carri compani agricultur chemic report predict market share chemic
report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share
stimul demand price cut volum sale

Krovetz stemmer:

document describe marketing strategy carry company agriculture chemical report prediction market
share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer
predict sale stimulate demand price cut volume sale

Други езици

kitab	<i>a book</i>
kitabı	<i>my book</i>
alkıtab	<i>the book</i>
kitabıki	<i>your book (f)</i>
kitabıka	<i>your book (m)</i>
kitabıhu	<i>his book</i>
kataba	<i>to write</i>
maktaba	<i>library, bookstore</i>
maktab	<i>office</i>

ktb

Въпроси?