استخراج اطلاعات از سایت

دانشجویان: آرمین
استاد راهنما: دکتر کاهانی

زمستان ۹۶

فهرست مطالب

- استخراج اطلاعات
- Web Extraction
- انواع الگوها
- پرداز ش زبان طبیعی
- ارزیابی سیستم استخراج اطلاعات
- استخراج وب با استفاده از درختانDOM
- کتابخانه jsoup
- تلفیق اطلاعات

استخراج اطلاعات

- مشخص کردن قسمت خاصی از اطلاعات در اسناد متنی غیر ساخت یافته یا نیمه ساخت یافته
- تبدیل اطلاعات غیر ساخت یافته در پیکره ای از اسناد و یا صفحات وب به داده های ساخت یافته
- بر روی متون مختلفی اعمال میشود:
  - مقالات روزنامه
  - صفحات وب
  - مقالات علمی
  - پیام های گروهی
  - تبلیغات طبقه بندی شده
  - یادداشت های پزشکی و ..

Web Extraction

- بسیاری از صفحات وب به طور خودکار از یک پایگاه داده زیرین تولید می شوند.
- بنابراین، ساختار HTML صفحات نسبتا مشخص و منظم است. (نیمه ساختار یافته)
- با این حال، خروجی برای مصرف انسان است، نه تفسیر ماشین.
- یک سیستم استخراج اطلاعات برای صفحات تولید شده این امکان را فراهم می کند که یک وب سایت به عنوان یک پایگاه داده ساختاری نمایش داده شود.

Web Extraction

- استخراج برای یک وب سایت نیمه ساختار یافته گاهی اوقات به عنوان یک *wrapper* تعریف می شود.
- فرآیند استخراج از چنین صفحاتی گاهی اوقات به عنوان *web scraping* نیز شناخته می شود.
- سیستم های *web scraping*:
- استفاده از تکنیک های تجزیه DOM،
- بینایی ماشین
- و پردازش زبان طبیعی
- وب سایت های بزرگ برای محافظت از داده های خود از الگوریتم های web scrapers معمولا از الگوریتم های دفاعی استفاده می کنند.

انواع الگوها

- شکاف در الگو معمولا توسط یک زیر رشته از سند پر می شود.
- برخی از شکاف ها ممکن است یک مجموعه ثابت از پرکننده های احتمالی از پیش تعیین شده داشته باشند .
  - مدرک : کارشناسی، کارشناسی ارشد، دکتری
  - رتبه علمی: استاد/ دانشیار/ استادیار
- بعضی از شکاف ها ممکن است چند پرکننده داشته باشند.
  - زبان برنامه نویسی
  - آشنایی با زبانهای خارجی
- بعضی از دامنه ها ممکن است چندین الگو استخراج شده در هر سند را اجازه دهند.
  - لیست مقالات
  - لیست کتابها

پرداز ش زبان طبیعی

- اگر استخراج اطلاعات از صفحات وبی که به طور خودکار ایجاد شده اند صورت گیرد، الگوهای ساده عبارات منظم معمولا کار می کنند.
- اگر استخراج از متن ساده، بدون ساختار، نوشته شده توسط انسان باشد، برخی از ابزارهای NLP ممکن است کمک کننده باشند.
  - Part-of-speech (POS) tagging
    - علامت گذاری هر کلمه به عنوان اسم، فعل، پیش فرض و غیره
  - Syntactic parsing
    - شناسایی عبارات: NP، VP، PP
  - Semantic word categories(به عنوان مثال از WordNet)
    کشتن: کشتن، قتل، ترور، خفه کردن، خفه شدن
- الگوهای استخراج کننده می توانند از POS یا برچسب های عبارات استفاده کنند.

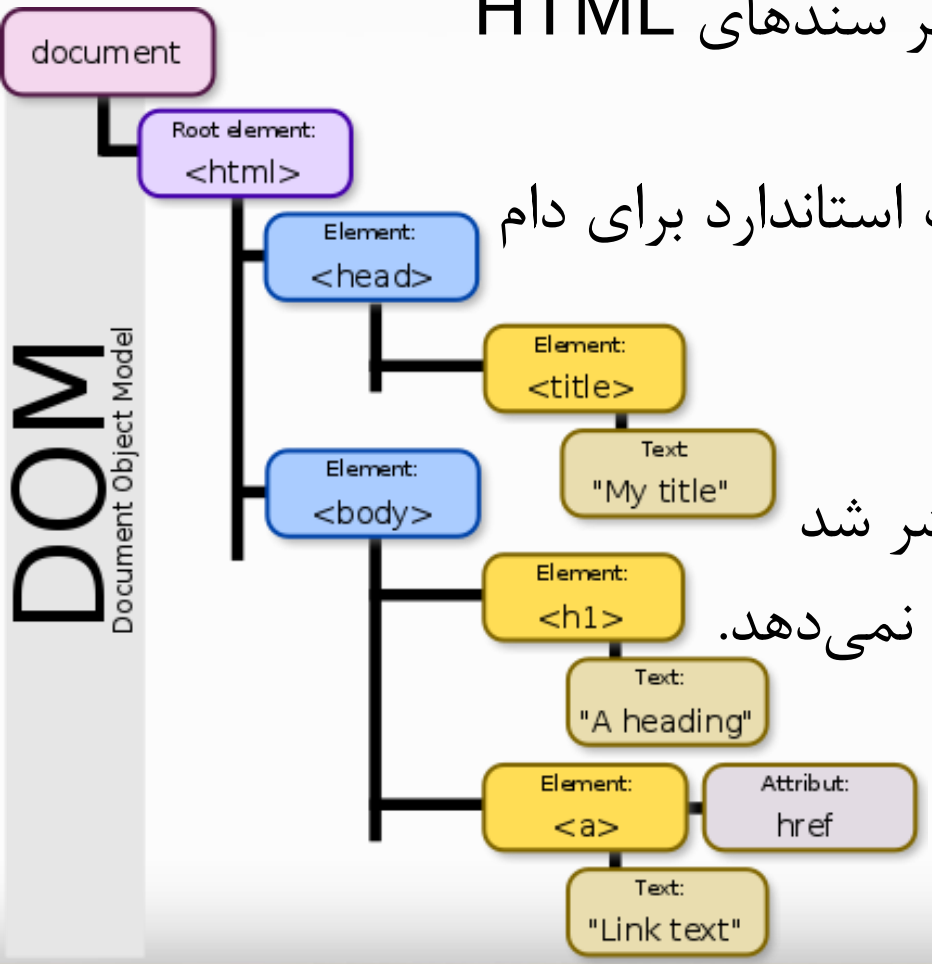
ارزیابی سیستم استخراج اطلاعات

- ارزیابی عملکرد سیستم باید روی داده های آزمایشی مستقل که بصورت دستی برچسب خورده اند و در طول توسعه سیستم استفاده نمی شوند، انجام گیرد
- برای هر سند آزمون موارد زیر را اندازه گیری کنید :
  - N : تعداد کل استخراج درست در الگو
  - E: تعداد کل جفت اسلات / مقدار استخراج شده توسط سیستم
  - C: تعداد جفت های اسلات / مقدار استخراج شده که صحیح هستند
- محاسبه مقدار میانگین معیارهای سازگار با IR:

- Recall = *C*/*N*
- Precision = *C*/*E*
- F-Measure = Harmonic mean of recall and precision

مدل شیءگرای سند DOM

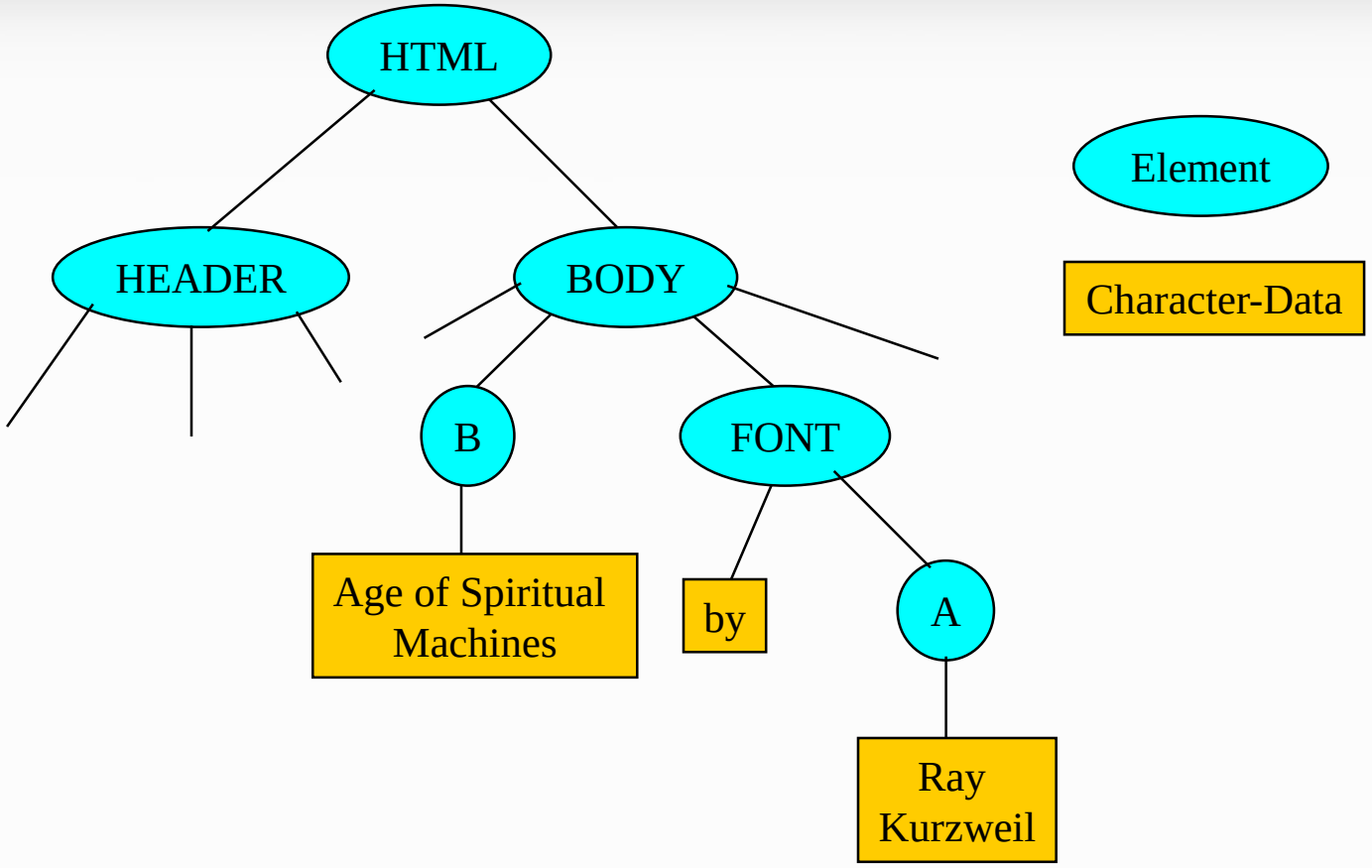
- Document Object Model
- گونه‌های مختلف دام توسط مرورگرهای وب برای پردازش عناصر سندهای HTML پیاده‌سازی می‌شدند.
- کنسرسیوم وب جهان‌شمول وادار شد که با یک سری مشخصات استاندارد برای دام پیشگام شود
- (از این رو آن را W3CDOMنیز می‌گویند).
- آخرین ورژن استاندارد DOM Level 4 در سال ۲۰۱۵ منتشر شد
- دام هیچ محدودیتی روی ساختار داده‌های دربرگیرنده سند قرار نمی‌دهد.
- یک سند خوش‌ساختار می‌تواند به کمک دام شکل درخت‌گونه به خودش بگیرد.



استخراج وب با استفاده از درختانDOM

- با اولین پردازش صفحات وب و تبدیل به درختان DOM ممکن است استخراج اطلاعات صورت گیرد.
- سپس الگوهای استخراج می تواند به عنوان مسیرهایی از ریشه درخت DOM به گره حاوی متن مشخص شود.
- ممکن است هنوز عبارت منظم برای شناسایی بخش مناسب نهایی گره نیاز باشد.

نمونه ای از استخراج درخت DOM



Title: HTML→BODY→B→CharacterData

Author: HTML→ BODY→FONT→A→ CharacterData

کتابخانه jsoup

- jsoup یک کتابخانه مبتنی بر جاوا است
- برای پردازش متن HTML به کار می رود
- این API بسیار مناسب برای استخراج و تغییر داده ها، با استفاده از بهترین روش های DOM، CSS و جی کوئری است.
- یک نمونه کد ساده برای گرفتن یک Url و تبدیل آن به DOM

```
URL url = new URL("http://gosmarter.net?query=cars");
Document doc = Jsoup.parse(url, 3000);
```

کتابخانه jsoup

- بازیابی یک جدول یا div را از درخت DOM

```
Iterator<Element> productList =
doc.select("div[class=productList]").iterator();
assertNotNull(productList.hasNext);
while (productList.hasNext()) {
//Do some processing
}
```

- استخراج نشانی اینترنتی تصویر

```
Element productLink = product.select("a").first();
String href = productLink.attr("abs:href");
```

تلفیق اطلاعات

- پاسخ دادن به سوالات خاصی با استفاده از وب نیاز به یکپارچه سازی اطلاعات از وب سایت های مختلف دارد.
- یکپارچه سازی اطلاعات مربوط به روش های خودکار ادغام است.
- نیاز به wrappers برای استخراج اطلاعات خاص از صفحات وب از سایت های خاص با دقت بالا دارد.
- هرwrapped سایت را به عنوان یک جدول پایگاه داده پردازش می کند و از طریق یک زبان پرس و جو پایگاه داده (مانند SQL) به پرسشهای پیچیده پاسخ می دهد.