1. Bài Toán Phân Tích Giá Cả Thị Trường



1. Mục đích

* Trong thời đại thương trường là chiến trường và việc giá cả của 1 sản phẩm liên tục thay đổi mang tính quyết định lớn cho chiến lược kinh doanh của 1 doanh nghiệp thì việc cần nắm bắt rõ giá cả biến động của các loại hàng sản phẩm là điều hết sức cần thiết.

1. Yêu cầu

* Thu thập dữ liệu từ các trang website.
* Phân loại dữ liệu.
* So sánh đối chiếu với sản phẩm tương ứng về giá, khuyến mại để kịp thời thay đổi chiến lược kinh doanh cho từng mặt hàng.
* Đưa ra biểu đồ xu hướng về giá của 1 mẫu sản phẩm?

1. Giải pháp
2. Thu thập dữ liệu

* Có nhiều phương pháp để scrape dữ liệu từ các trang web khác:
* **Python**
  + [Beautiful Soup](http://www.crummy.com/software/BeautifulSoup/)
  + [lxml](http://codespeak.net/lxml/)
  + [HTQL](http://htql.net/)
  + [Scrapy](http://scrapy.org/)
  + [Mechanize](http://wwwsearch.sourceforge.net/mechanize/)
* **Ruby**
  + [Nokogiri](http://nokogiri.org/)
  + [Hpricot](https://github.com/hpricot/hpricot/)
  + [Mechanize](https://github.com/tenderlove/mechanize)
  + [scrAPI](http://rubyforge.org/projects/scrapi/)
  + [scRUBYt!](http://scrubyt.org/)
  + [wombat](https://github.com/felipecsl/wombat)
  + [Watir](http://watir.com/)
* **.NET**
  + [Html Agility Pack](http://htmlagilitypack.codeplex.com/)
  + [WatiN](http://watin.org/)
* **Perl**
  + [WWW::Mechanize](http://search.cpan.org/dist/WWW-Mechanize/)
  + [Web-Scraper](http://search.cpan.org/dist/Web-Scraper/)
* **Java**
  + [Tag Soup](http://home.ccil.org/~cowan/XML/tagsoup/)
  + [HtmlUnit](http://htmlunit.sourceforge.net/)
  + [Web-Harvest](http://web-harvest.sourceforge.net/)
  + [jARVEST](http://sing.ei.uvigo.es/jarvest)
  + [jsoup](http://jsoup.org/) (open suorce free) not support https
  + [Jericho HTML Parser](http://jericho.htmlparser.net/docs/index.html)
  + Selenium
  + http://jaunt-api.com/
* **JavaScript**
  + [node.io](https://github.com/chriso/node.io)
  + [phantomjs](http://phantomjs.org/) (casperJS…)
* **PHP**
  + [htmlSQL](https://github.com/hxseven/htmlSQL)
* **Most of them**
  + [Screen-Scraper](http://www.screen-scraper.com/)
* Mỗi phương pháp đều có ưu và nhược điểm riêng. Việc cần biết nhiều phương pháp sẽ hỗ trợ tốt cho việc thu thập dữ liệu.
* Selenium là phương pháp an toàn và ổn định do việc áp dụng cơ chế automation test để lấy dữ liệu về. Tức là nó dùng thông qua 1 trình duyện driver và mô tả lại gần giống như tương tác người dùng nên việc bị chặn lấy dữ liệu từ các trang web là rất ít.

1. Phân loại dữ liệu(Text Mining and Classification)

* FullText search là một kỹ thuật hỗ trợ tốt cho việc phân loại dữ liệu dựa trên việc áp dụng tìm kiếm trên các từ khóa (keyword).
* Việc áp dụng fulltext search đơn giản, nhanh chóng và hiệu quả trên tập dữ liệu lớn.

1. Kiến trúc hệ thống

App Scrape Data

Database

Analyst data

(FullText Search)

App View

…

call

SQL Server

**Technical Scrape**

1. Selenium

2. CasperJS

3. [jsoup](http://jsoup.org/)

….