Using web scraped data to verify Egyptian consumer price indices

Faculty of Computer and Information Sciences, Ain Shams University, Egypt

Prepared by
Mina Essam Kamel Gerges
Supervisor Prof. Dr. El-sayed M. El-Horbaty

Central Agency for Public Mobilization And Statistics

Data collection and integration

INTRODUCTION

Recently, In Egypt many companies have been published several websites for e-commerce and one of these is souq.com owned by Amazon, Inc. which made scraping data more available and in general appeared what is called: Web scrapers which are software tools for extracting data from web pages. The growth of online markets over recent years means that many products and associated prices information can be found online and possible to be scrapable.

The consumer price index is one of the official statistics which estimate constructed using the prices of a sample of representative items whose prices are collected periodically; so it's one of the best examples in this sense: by replacing the scraping of e-commerce websites and websites which publish the currently prices of products to automatically collect prices for some products and services rather than physical visiting to stores to manually collect the prices. This offers a range of great benefits including: Reducing data collection costs, increasing the frequency of collection and products in the basket, and improving our understanding of price behavior. This paper introduces a developed generic tool that automatically collects online prices, as "Scraped Data", based on multiple Search Engines to crawler newest prices and e-commerce websites. The developed tool aiming to aid in data collection reduction costs process depend on big data analytics.

Methodology

The proposed system architecture consists of five main modules. Namely, Data pre-processing based on seed URL list for online markets, Data crawling contains two main functions are searching by Google search engine and fetching the new links, Data scraping and its main function is extraction information from the downloaded HTML pages, Text processing contains Tokenize Text and normalize it, and Data structuring.

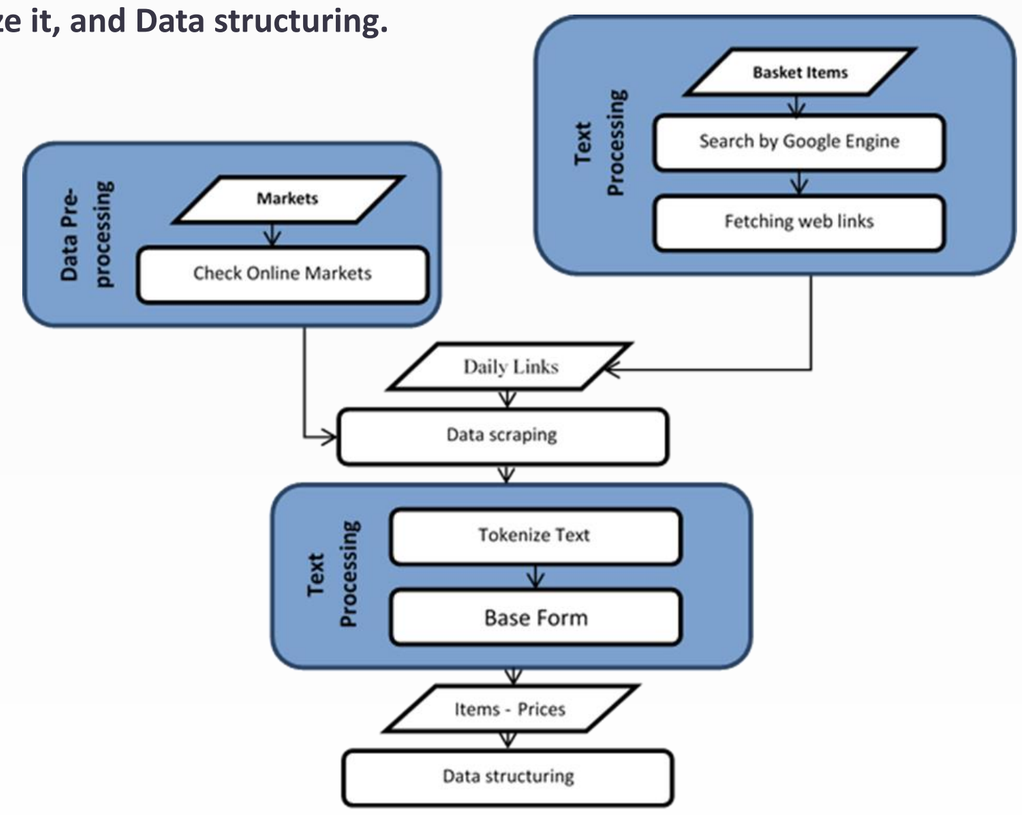

Fig. the proposed Web scraping System architecture

1 Data Pre-processing

preparing the dataset of e-commerce websites and the other websites which publish or show the prices periodically. Online prices are collected, based on both the website offers and the items' services regarding different regions, from different markets' pages

2 Data Crawling

Using of web search engines especially Google SE that use web crawling or spidering software to update their web content or indices of others sites' web content; and by passing new queries for these search engines, we enclose the current date for daily price and web crawlers copy pages for processing by a search engine which indexes the downloaded pages

3 Data Scraping & Text Processing

Extracted information; In each page are arranged in the form of tree nodes annotated as "HTML Tags", where different tags have different both meaning and content; then extracting specific text from all of scraped streams. Separating the collected text into a datasets of tokens then get Specific words are observed such as "item name", "units" …etc. In Arabic; Same words can be written in different forms, accordingly, a conversion process is mandatory in order to get a unique form for each word. Such unique form is considered as the base form "normalization" for that word

4 Data Structuring

The prices which scraped from the websites stored in one structured dataset. And the following information has been extracted and organized in a structured tables

Results

- Lack of prices for some products with a limited shelf life such as vegetables, fruits, fresh meat and fresh fish.
- Increasing the number of items which used in the basket especially the electronics products such as the mobiles, computers, home appliances and grocery.
- Decreasing the period between the collection process and show the size of the changes of these items through the weekdays and whole of month.
- By disappearing some items, increasing some else and decreasing the period of collection give great different values compared to the current approach of Egyptian CPI construction

Conclusion

1. The full use of data extracted from the web may lead to a fundamental change in the understanding of price changes in general.
2. Not all items differ significantly from the current system "manual data collection and static basket items". Further investigations may be considered to the disappearing some items, increasing some else, decreasing the period of collection or whether the price behavior of these elements differs significantly from the items with large variations.
3. The research can be used to verify the accuracy of the manual system until the tool will be completed and can be answer the causes of these results.
4. This paper introduces a development of a generic automated tool for scarping and structuring the online data and enhances the process by passing new queries for search engines. This tool offers great benefits in, not only, reducing data collection costs, but also it allows for increasing the number of items without adding any cost; we validated the collected data by that collected by NSO team manually; the tool can allow for increase the frequency of collection, e.g. collect the data daily; discover new items that did not exist in the basket; and aim to understand behaviour of prices.

REFERENCES…

[1] Hoekstra, R., ten Bosh, O., Harteveld, F. (2012). "Automated data collection from web sources for official statistics: First experiences." Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 28(3-4).

[2] Robert Griffioen, Jan de Haan and Leon Willenborg, Collecting clothing data from the Internet, Division PIM Department of Methodology (Apr. 2014).

[3] Alberto Cavallo, SCRAPED DATA AND STICKY PRICES, NATIONAL BUREAU OF ECONOMIC RESEARCH (Aug. 2015).

[4] Robert Breton, Gareth Clews, Liz Metcalfe, Natasha Milliken, Christopher Payne, Joe Winton and Ainslie Woods, Research indices using web scraped data, Office for National Statistics (Sep. 2015)