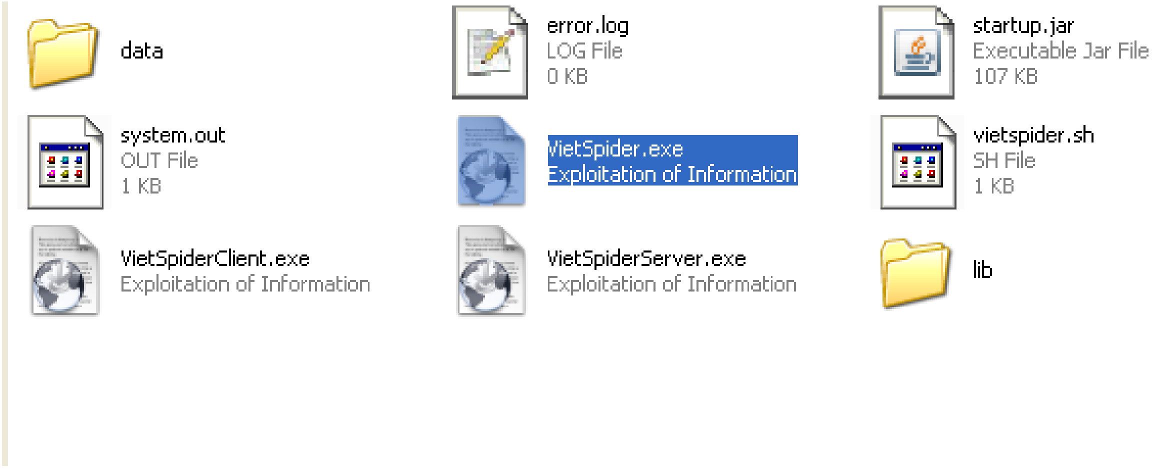
VietSpider Quick Guide – Step by Step – extract data from Android Market

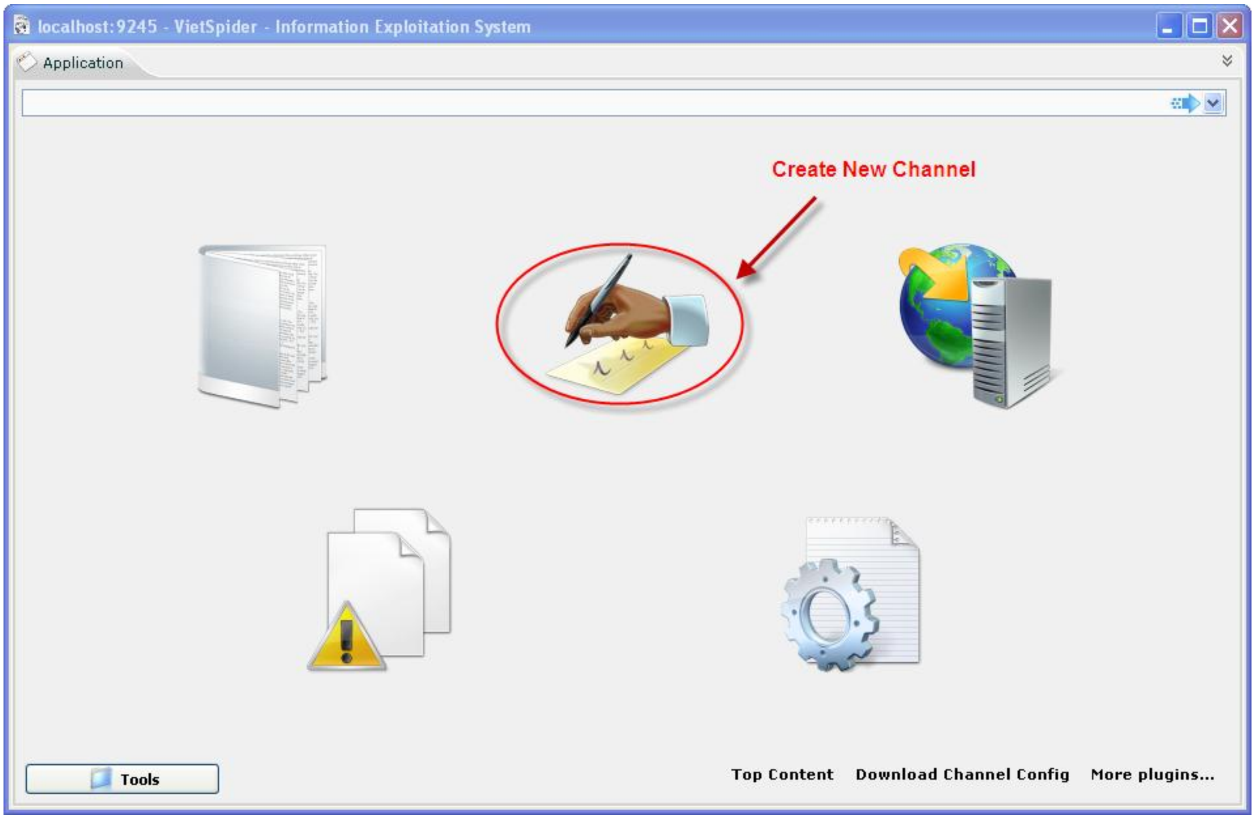
Part 1: Create new channel

1. **Launch VietSpider**



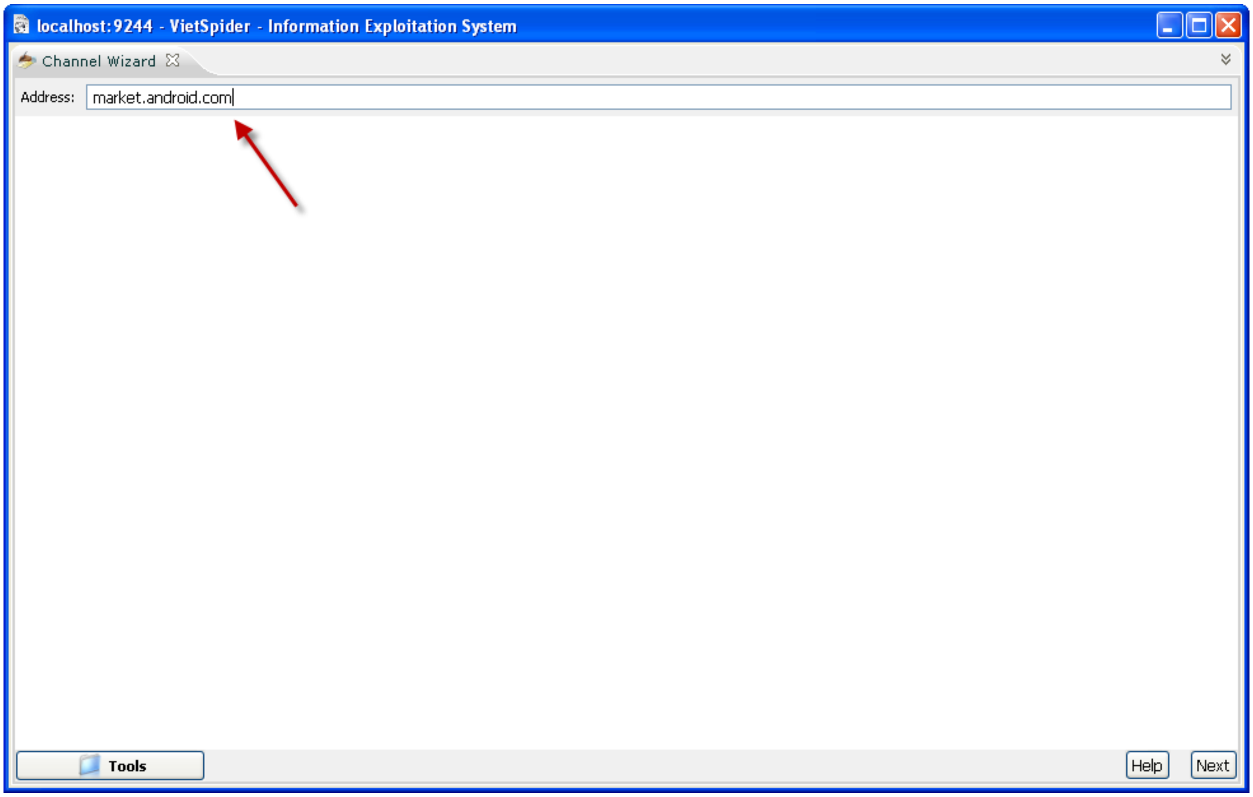
Open VietSpider folder and double click on VietSpider.exe

2. **Open Create Channel Wizard**



Click on Create New Channel icon

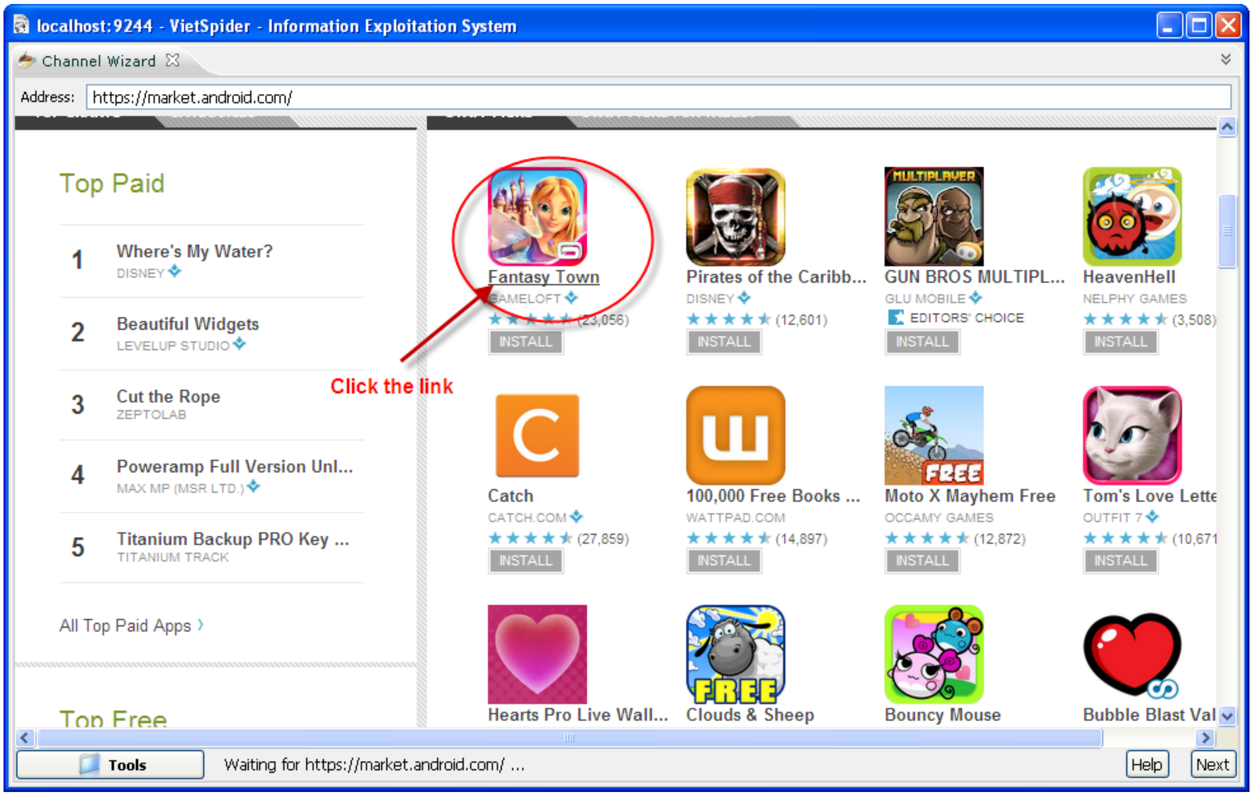
3. **Browse website**


localhost:9244 - VietSpider - Information Exploitation System
Channel Wizard
Address:
Tools
Help
Next

market.android.com

Input your address into the Address text box.

4. **Select a data page**



localhost:9244 - VietSpider - Information Exploitation System
Channel Wizard
Top Paid
1 Where's My Water? DISNEY
2 Beautiful Widgets LEVELUP STUDIO
3 Cut the Rope ZEPTOLAB
4 Poweramp Full Version Unl... MAX MP (MSR LTD.)
5 Titanium Backup PRO Key ... TITANIUM TRACK
All Top Paid Apps ›
Top Free
Fantasy Town GAMELOFT (23,056) INSTALL
Pirates of the Caribb... DISNEY (12,601) INSTALL
GUN BROS MULTIPL... GLU MOBILE EDITORS' CHOICE INSTALL
HeavenHell NELPHY GAMES (3,508) INSTALL
Catch CATCH.COM (27,859) INSTALL
100,000 Free Books ... WATTPAD.COM (14,897) INSTALL
Moto X Mayhem Free OCCAMY GAMES (12,872) INSTALL
Tom's Love Lette OUTFIT 7 (10,671 INSTALL
Hearts Pro Live Wall...
Clouds & Sheep
Bouncy Mouse
Bubble Blast Val
Tools
Help
Next

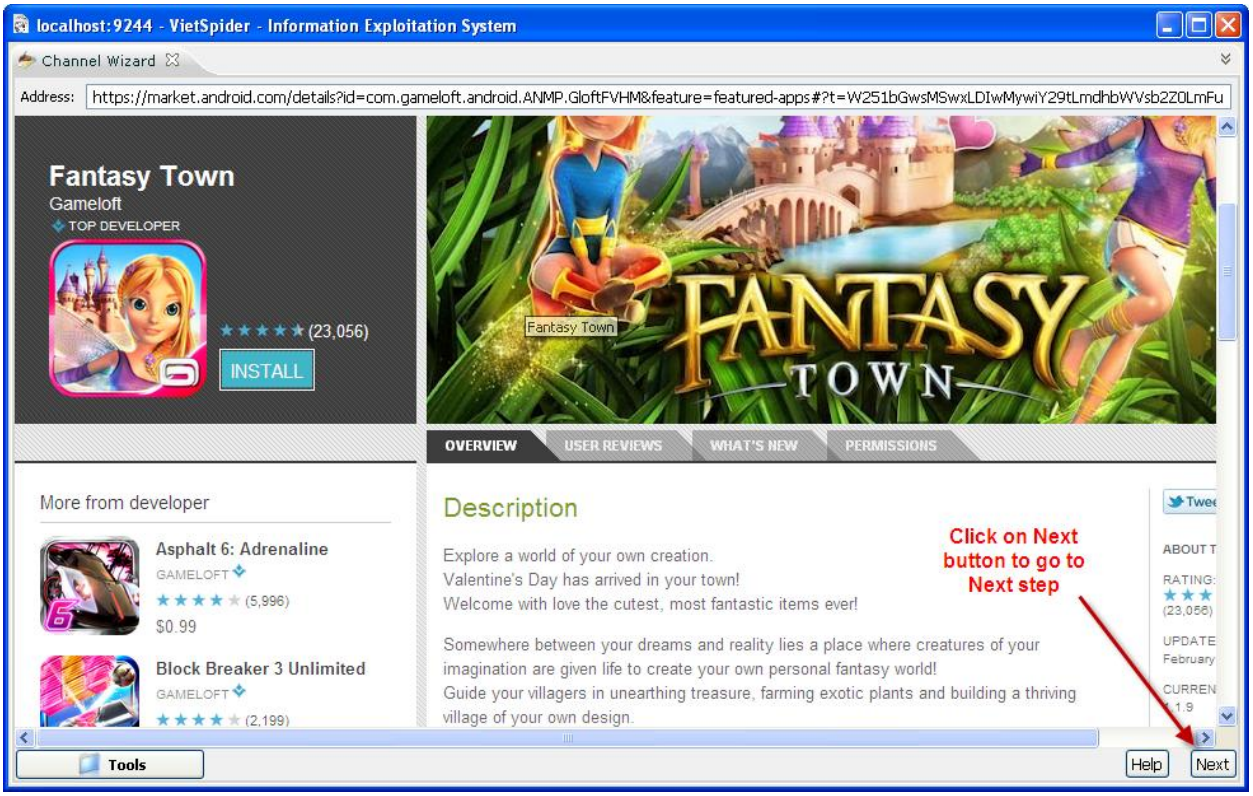
Address: https://market.android.com/

Waiting for https://market.android.com/ ...

Click the link

Select a sample data page. This page will be used for selecting extract data area.

5. Go to Next step



Address: https://market.android.com/details?id=com.gameloft.android.ANMP.GloftFVHM&feature=featured-apps#?t=W251bGwsMSwxLDIwMywiY29tLmdhbWVsb2Z0LmFu

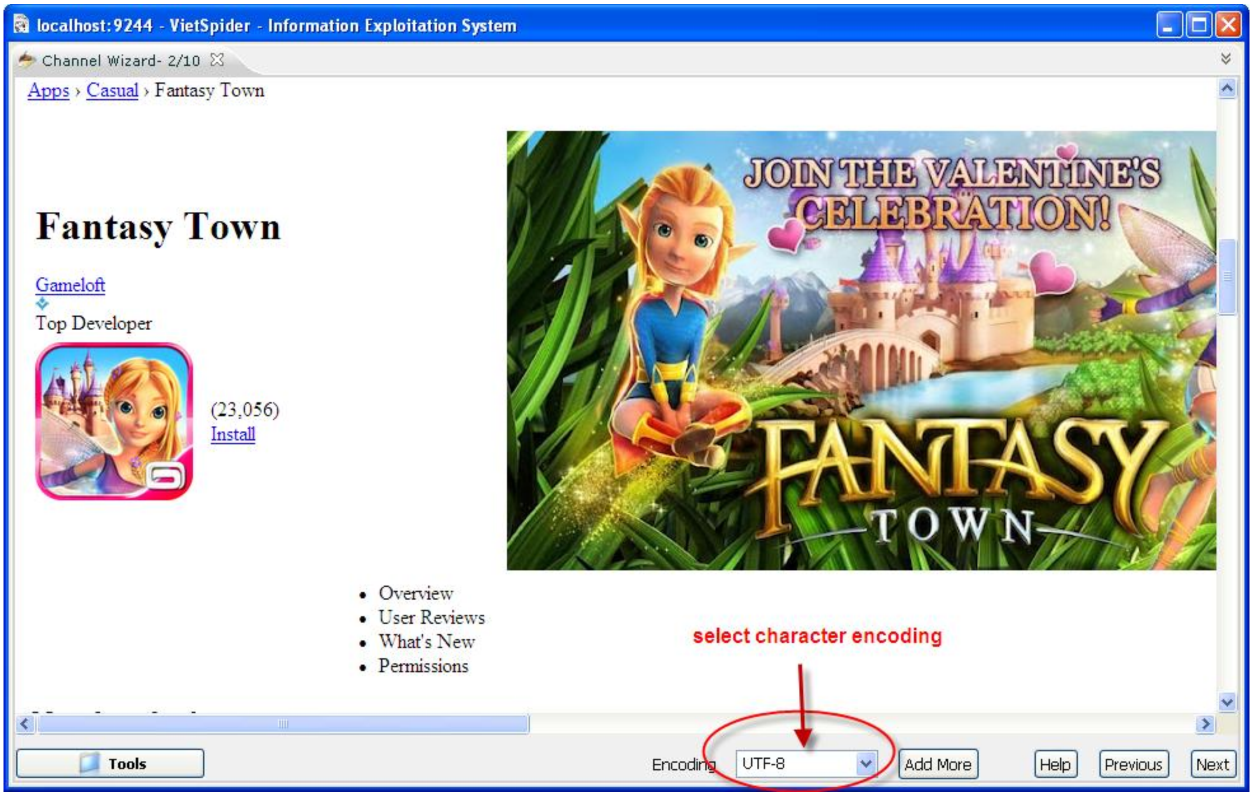
Description

Explore a world of your own creation.
Valentine's Day has arrived in your town!
Welcome with love the cutest, most fantastic items ever!

Somewhere between your dreams and reality lies a place where creatures of your imagination are given life to create your own personal fantasy world!
Guide your villagers in unearthing treasure, farming exotic plants and building a thriving village of your own design.

Click on Next button to go to Next step

6. **Select Character Encoding for this website**

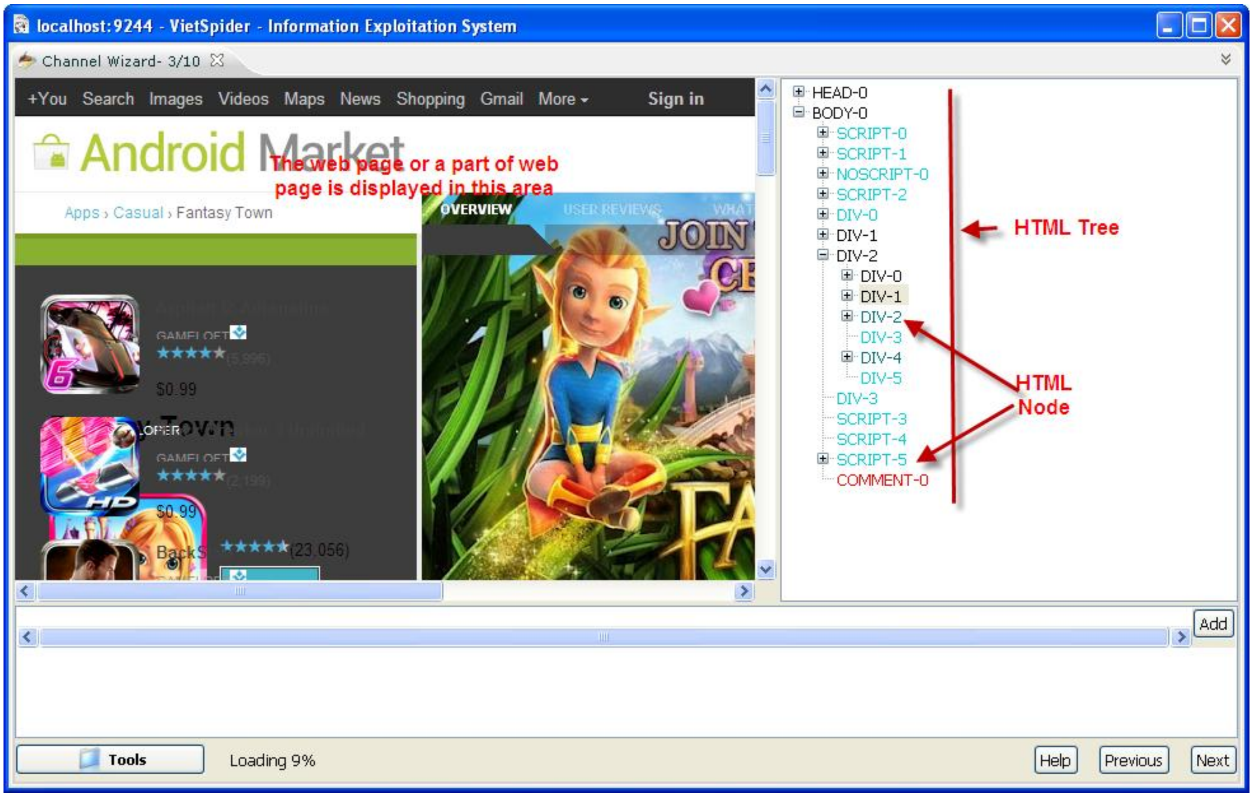


select character encoding

Encoding UTF-8

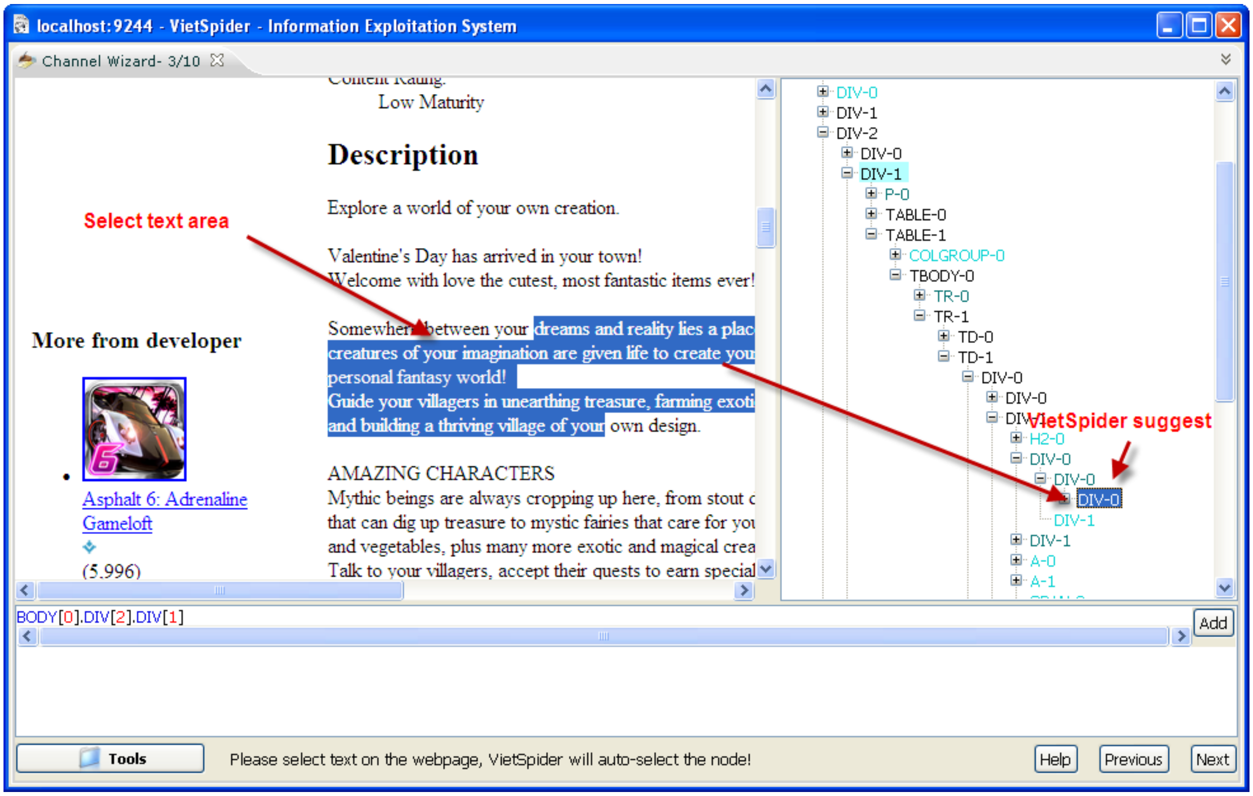
The character encoding selected here will be used to crawl data from this channel. If the web page displays properly with character encoding, please click on Next button to go to next step.

7. Define data areas on the data page



Click on HTML Node to select data area that you want VietSpider extracts information.

8. Auto select html node



Content Rating:
Low Maturity

Description

Explore a world of your own creation.

Valentine's Day has arrived in your town!
Welcome with love the cutest, most fantastic items ever!

Somewhere between your dreams and reality lies a plac
creatures of your imagination are given life to create you
personal fantasy world!
Guide your villagers in unearthing treasure, farming exoti
and building a thriving village of your own design.

AMAZING CHARACTERS
Mythic beings are always cropping up here, from stout c
that can dig up treasure to mystic fairies that care for you
and vegetables, plus many more exotic and magical crea
Talk to your villagers, accept their quests to earn special

More from developer

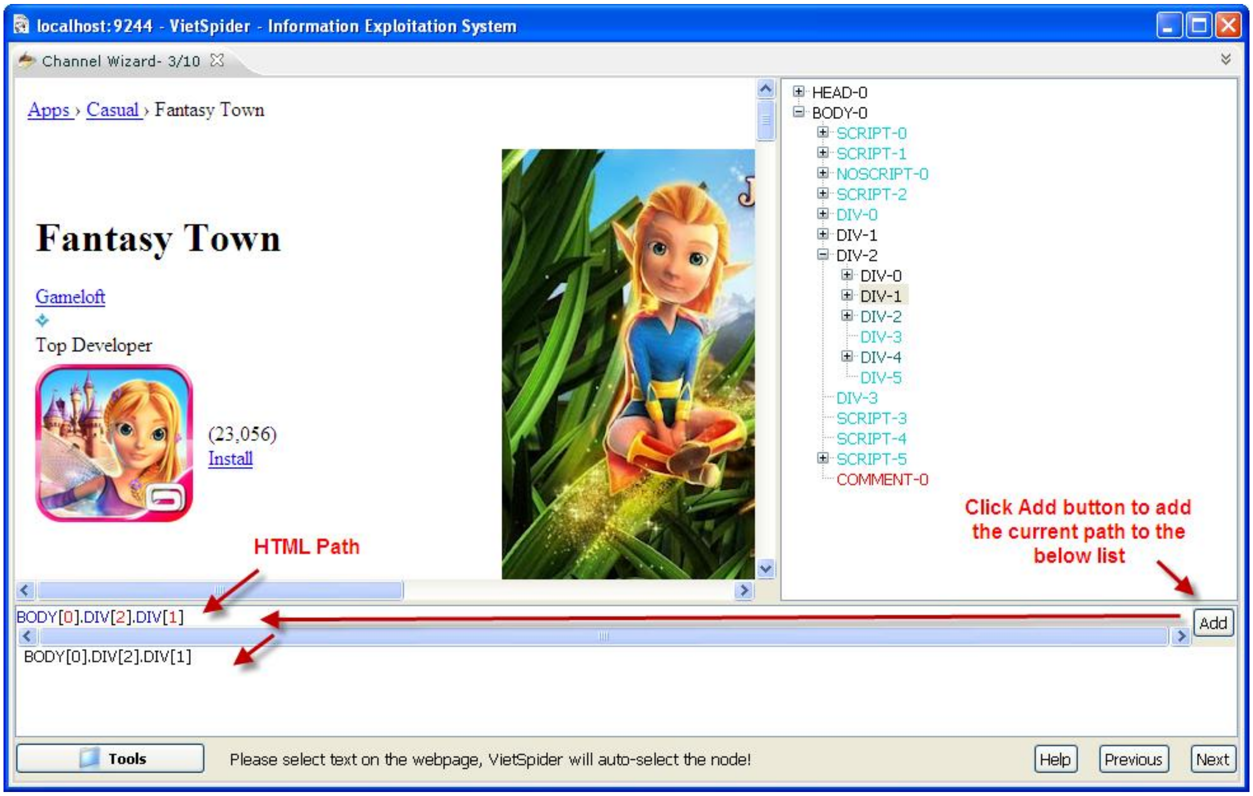
Asphalt 6: Adrenaline
Gameloft
(5,996)

BODY[0].DIV[2].DIV[1]

Please select text on the webpage, VietSpider will auto-select the node!

Please block a passage of text, VietSpider will suggest the node that contains the equivalent data.

9. Add extract data node(s)



HTML Path

Click Add button to add the current path to the below list

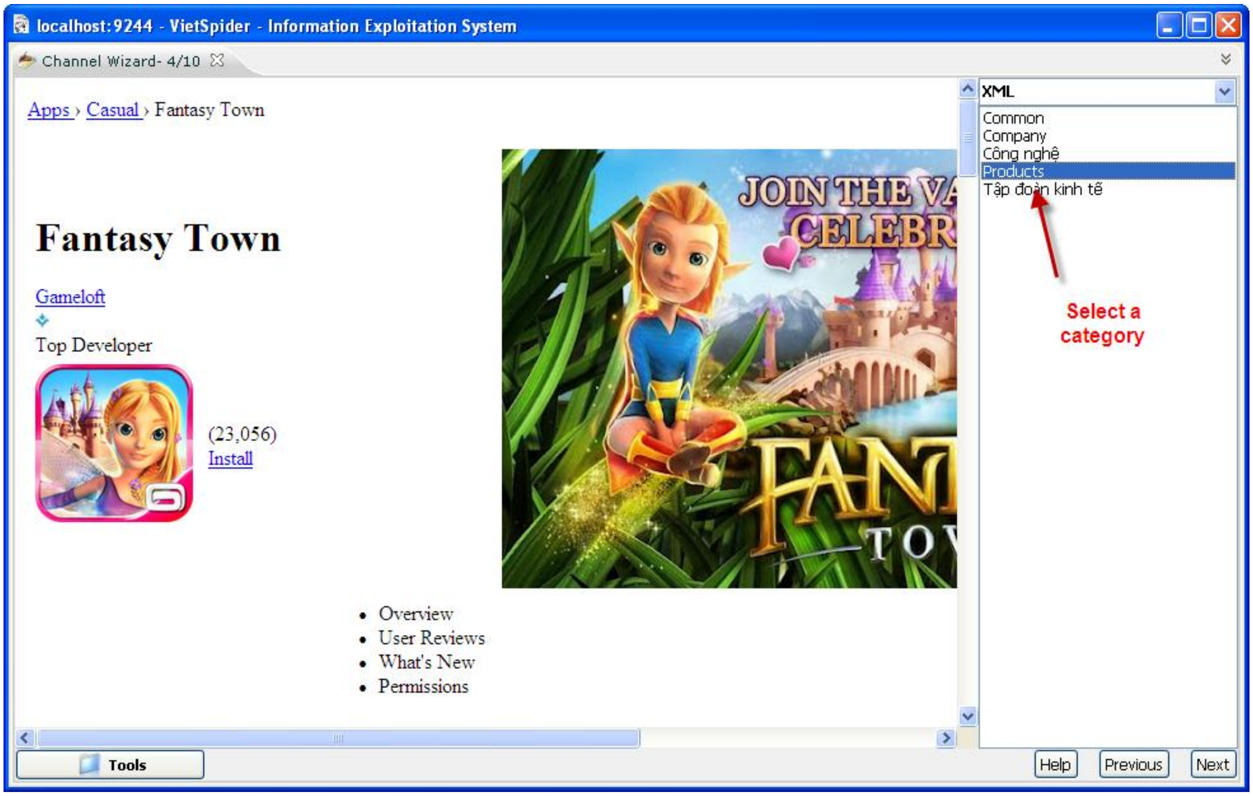
BODY[0].DIV[2].DIV[1]

BODY[0].DIV[2].DIV[1]

Please select text on the webpage, VietSpider will auto-select the node!

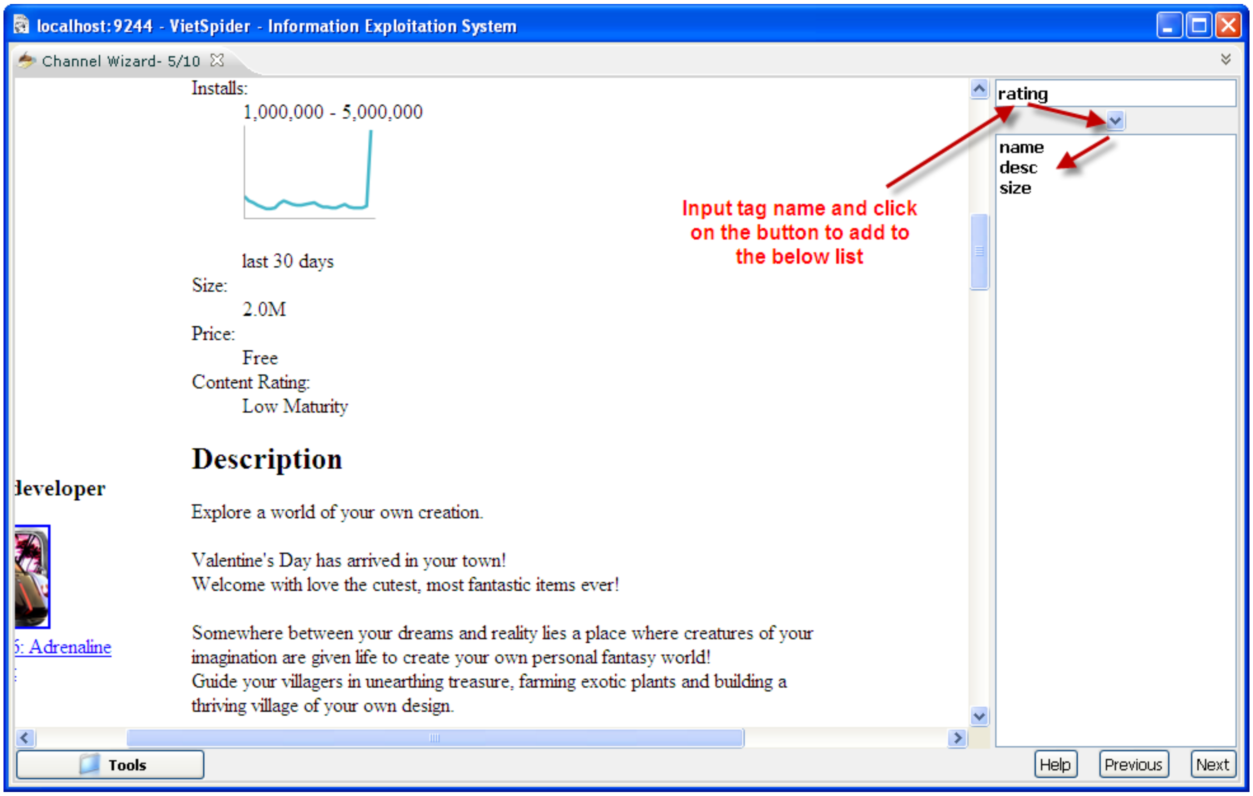
Click Add button to add the current path (from selected HTML Node) to the HTML Path list. The list contains path(s) to data area(s) that you want to extract. Click Next button to go to next step.

10. Select category for channel.



Select a category in the list and click Next button.

11. **Define XML elements of data document.**



Installs:
1,000,000 - 5,000,000
last 30 days
Size:
2.0M
Price:
Free
Content Rating:
Low Maturity

Description

Explore a world of your own creation.

Valentine's Day has arrived in your town!
Welcome with love the cutest, most fantastic items ever!

Somewhere between your dreams and reality lies a place where creatures of your imagination are given life to create your own personal fantasy world!
Guide your villagers in unearthing treasure, farming exotic plants and building a thriving village of your own design.

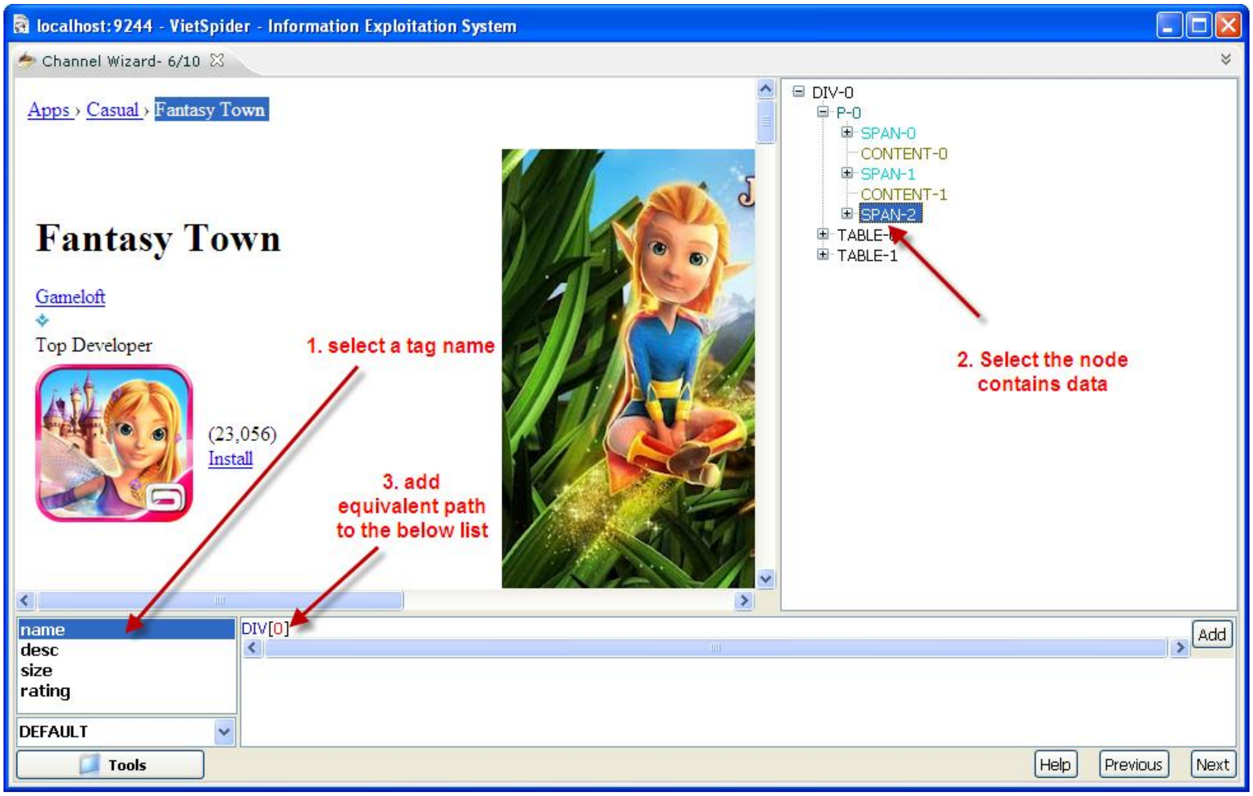
rating

name
desc
size

Input tag name and click on the button to add to the below list

You must define XML schame for extracted data document. Please define XML elements by input tag name and add it to the list. Click on Next button when you done.

12. Define data area for XML elements



1. select a tag name

2. Select the node contains data

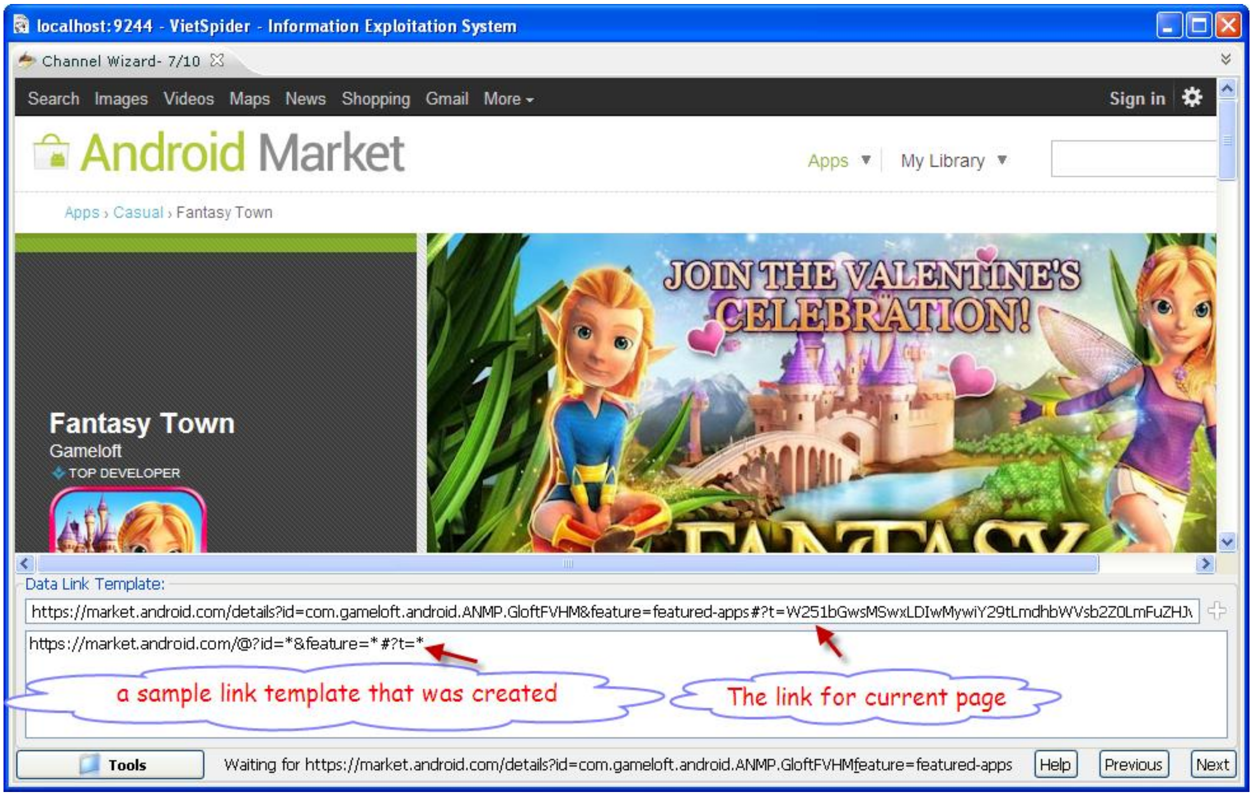
3. add equivalent path to the below list

Define data area for XML element, you must select a tag name -> select the node(s) that contains data area then add equivalent path to the list.

There are 4 data type for element:

- Default – VietSpider will auto detect data for equivalent element. That could be CDATA or Text
- Text – the element contains text value.
- CDATA – the element contains HTML tag.
- File - data type is used for image tag or download file link.

13. Define data link template.



localhost:9244 - VietSpider - Information Exploitation System
Channel Wizard- 7/10
Search Images Videos Maps News Shopping Gmail More
Sign in
Android Market
Apps
My Library
Apps › Casual › Fantasy Town
Fantasy Town
Gameloft
TOP DEVELOPER
JOIN THE VALENTINE'S CELEBRATION!
FANTASY
a sample link template that was created
The link for current page
Tools
Help
Previous
Next

Data Link Template:

https://market.android.com/details?id=com.gameloft.android.ANMP.GloftFVHM&feature=featured-apps#?t=W251bGwsMSwxLDIwMywiY29tLmdhbWVsb2Z0LmFuZHJ

https://market.android.com/@?id=*&feature=*#?t=*

Waiting for https://market.android.com/details?id=com.gameloft.android.ANMP.GloftFVHMfeature=featured-apps

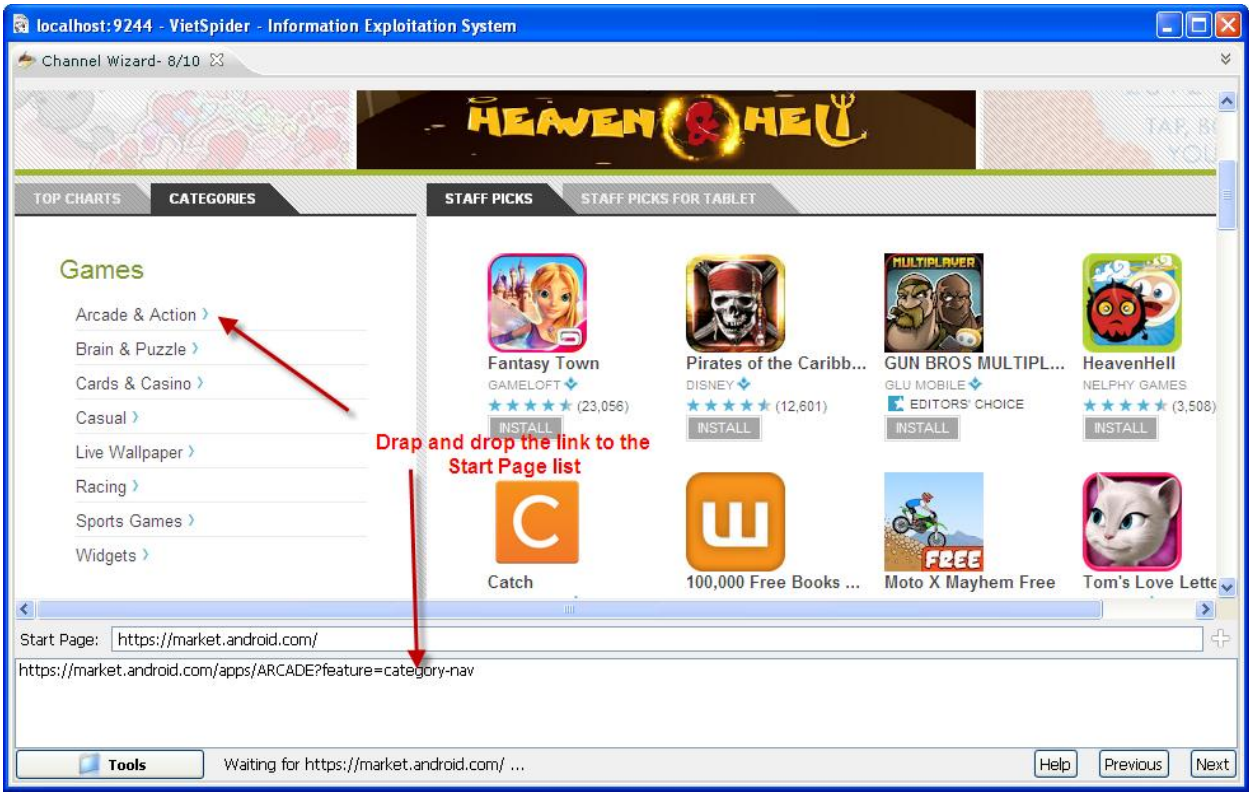
Data Link Template: this is template that will be used for VietSpider filters data link(s) from link collection when the Crawler crawl data.

Summary of regular-expression constructs

Character – Matches

- *   - any character
- @   - letter character
- $ - digit character

14. Add the Start page collection.



Drap and drop the link to the Start Page list

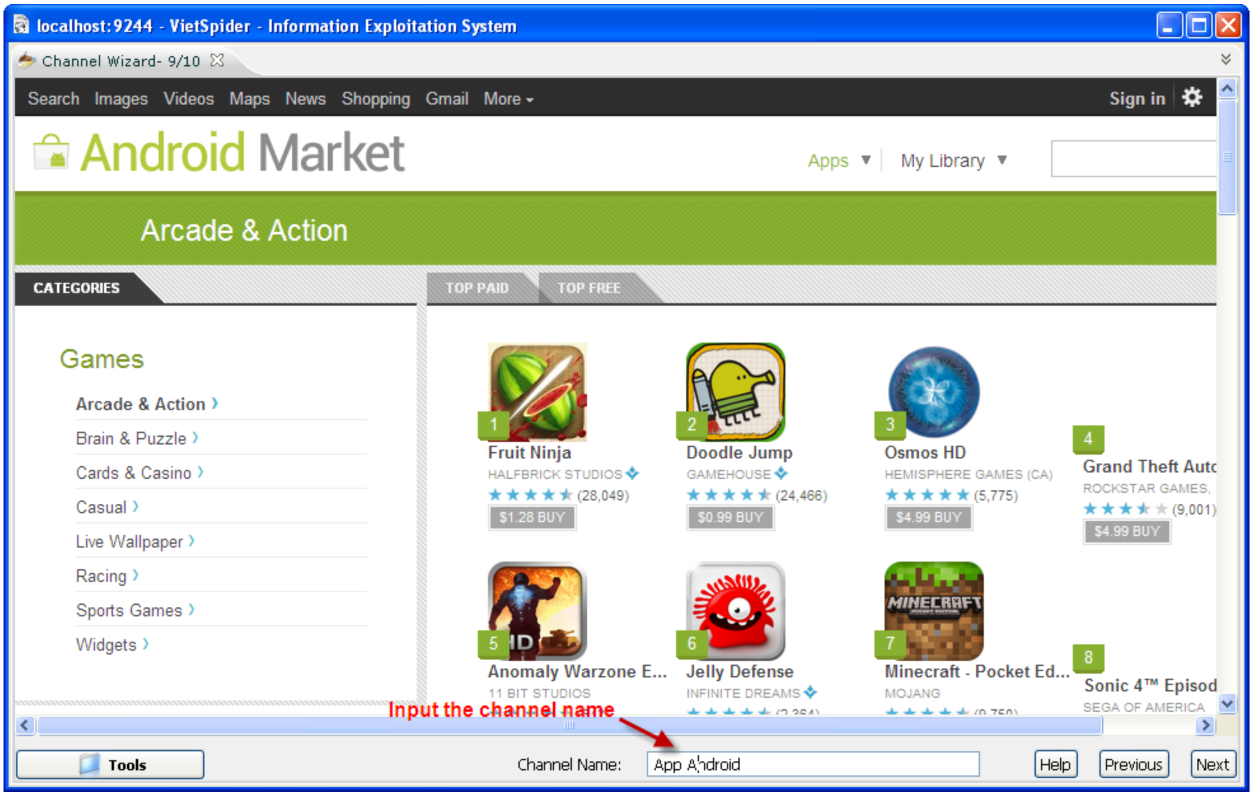
Start Page: https://market.android.com/

https://market.android.com/apps/ARCADE?feature=category-nav

Waiting for https://market.android.com/ ...

The Start Page(s): the list of links that VietSpider will use to visit and find data link when crawl data. You can drap and drop the link from the browser.

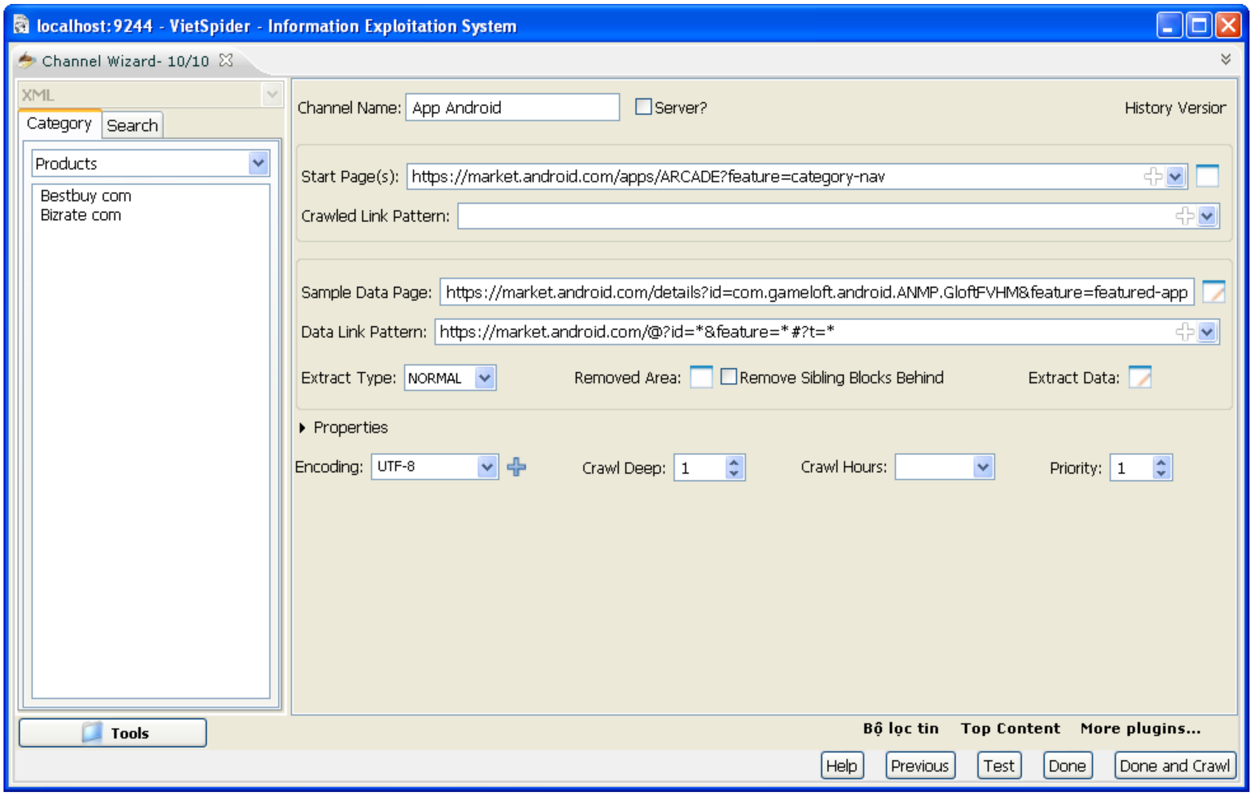
15. Define channel name



Channel Name: App Android

Input the channel name and click Next button to go to the last step.

16. Finish and save the channel



Channel Name: App Android ☐ Server?

Start Page(s): https://market.android.com/apps/ARCADE?feature=category-nav

Crawled Link Pattern:

Sample Data Page: https://market.android.com/details?id=com.gameloft.android.ANMP.GloftFVHM&feature=featured-app

Data Link Pattern: https://market.android.com/@?id=*&feature=*#?t=*

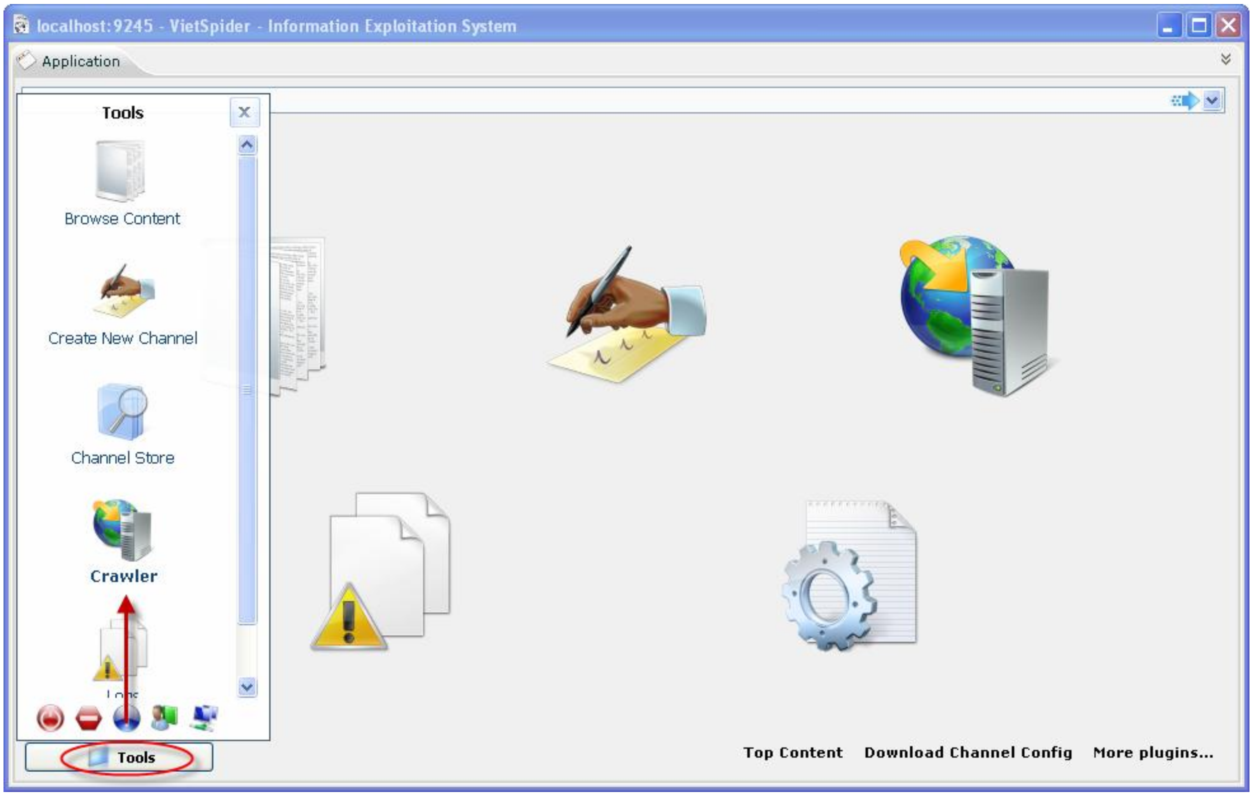
Extract Type: NORMAL Removed Area: ☐ Remove Sibling Blocks Behind Extract Data:

Encoding: UTF-8 Crawl Deep: 1 Crawl Hours: Priority: 1

If you want to test data extracting, you can copy another data link from the website and paste it to the Sample Data Page text box and click on Test button. Click Done button to save the channel to the Channel store.

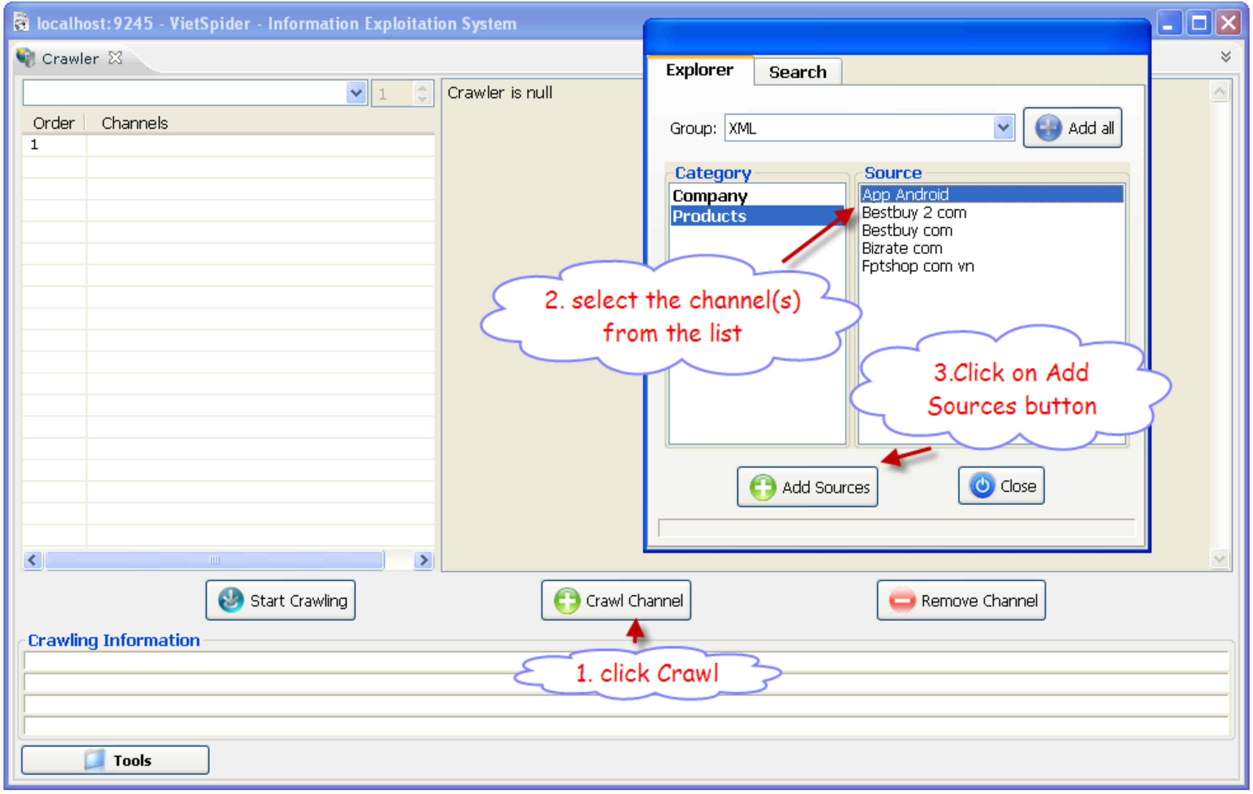
Part 2: Crawl data by the configured channel.

17. **Open Crawler**



From VietSpider user interface, click Tools button -> click on Crawler icon to open Crawler.

18. Add channel(s) to crawling list.



1. click Crawl

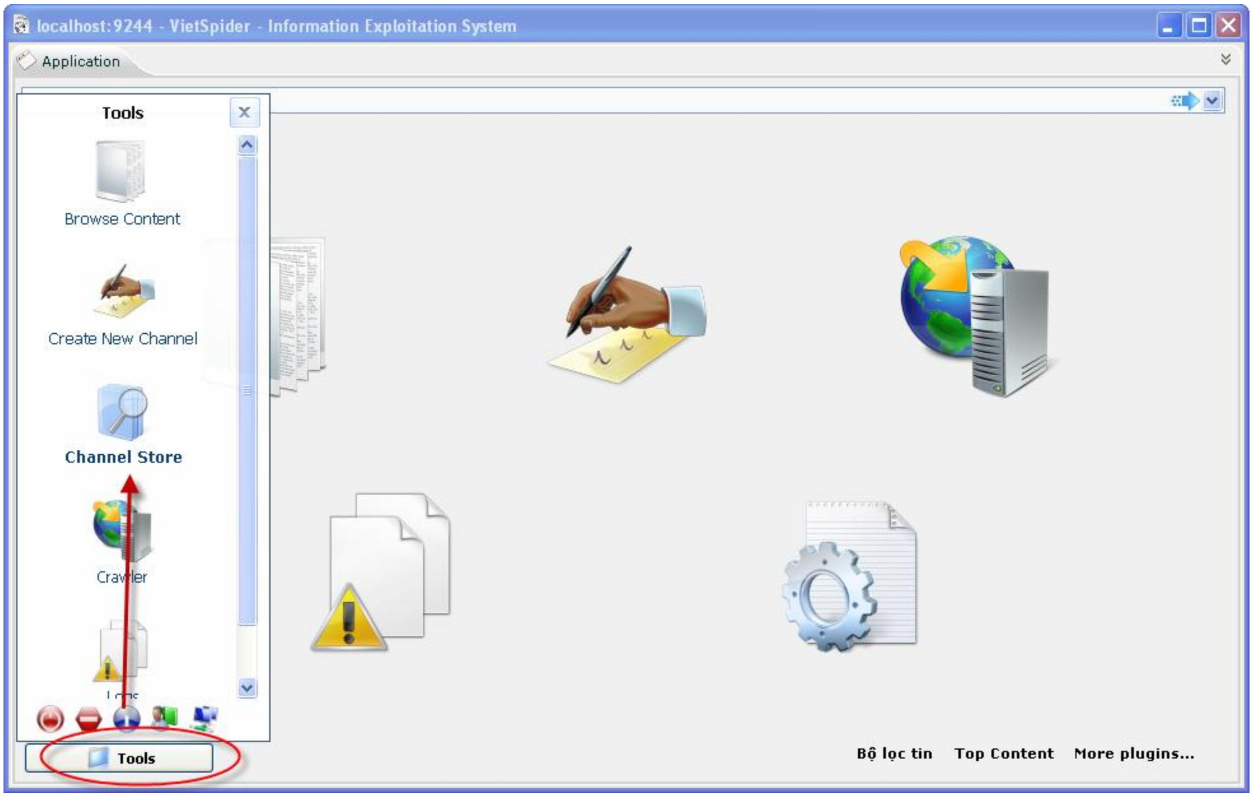
2. select the channel(s) from the list

3.Click on Add Sources button

Click on Crawl Channel button and add the channel(s) that you want to download and extract data. Click Start Crawling button for starting crawling data from website(s).

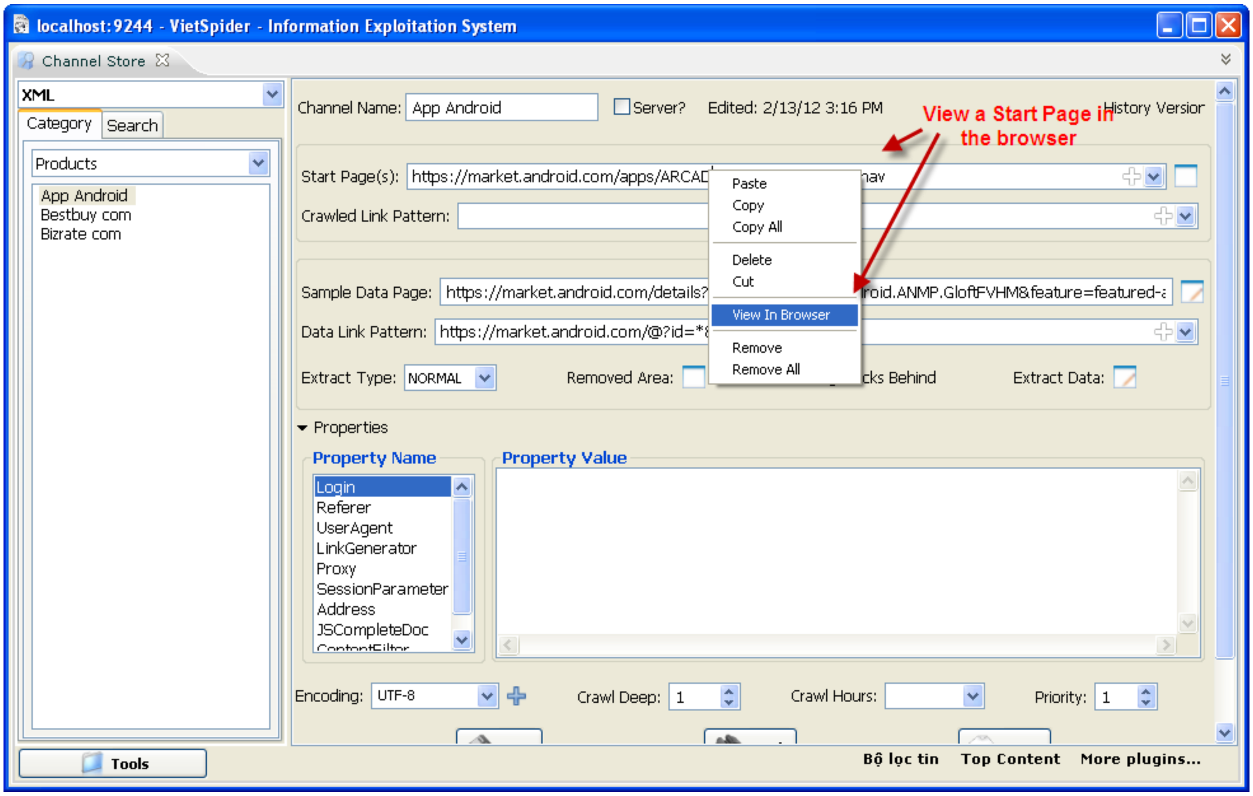
Part 3: Repair the channel.

19. Open the Channel Store



Click Tools button then click on Channel Store icon.

20. View website in the browser.



Channel Name: App Android □ Server? Edited: 2/13/12 3:16 PM

Start Page(s): https://market.android.com/apps/ARCAD...nav

Sample Data Page: https://market.android.com/details?...roid.ANMP.GloftFVHM&feature=featured-a

Data Link Pattern: https://market.android.com/@?id=*...

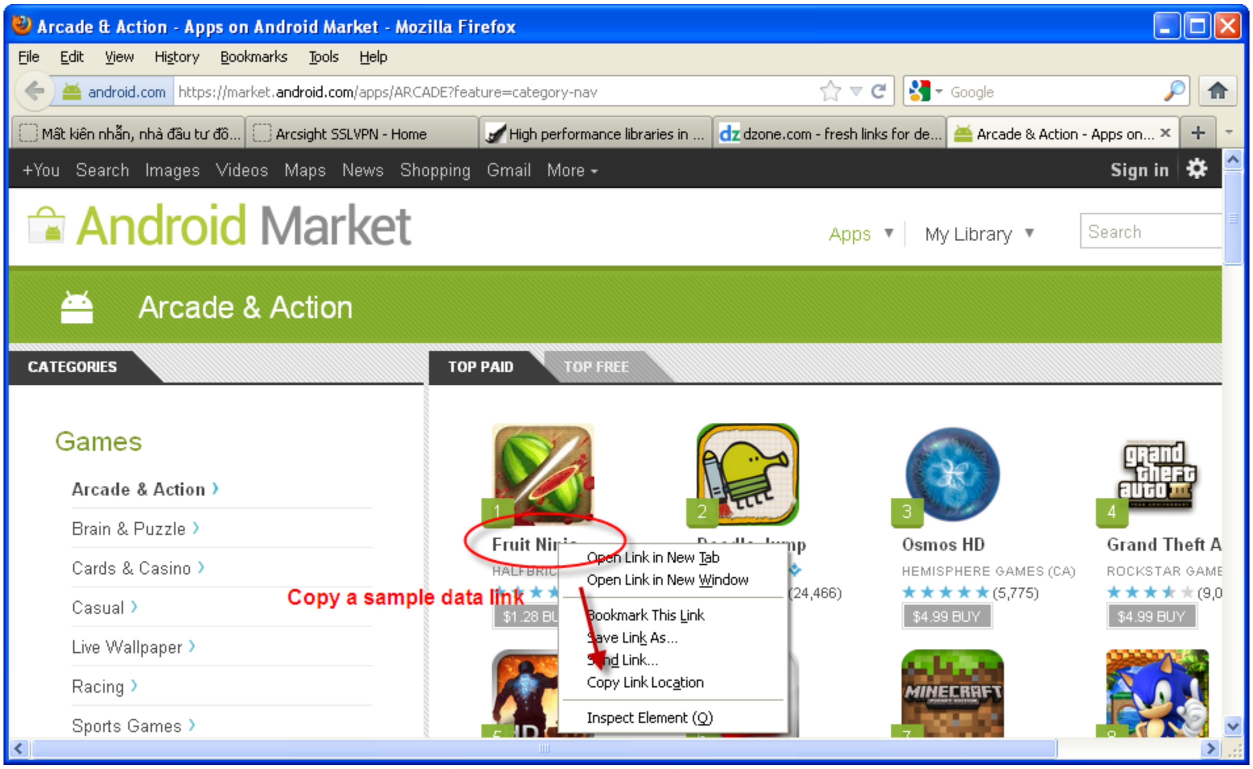
Extract Type: NORMAL

Encoding: UTF-8 Crawl Deep: 1 Priority: 1

View a Start Page in the browser

Select the channel that you want repair it. From Start Page(s) text box, right click then select View In Browser.

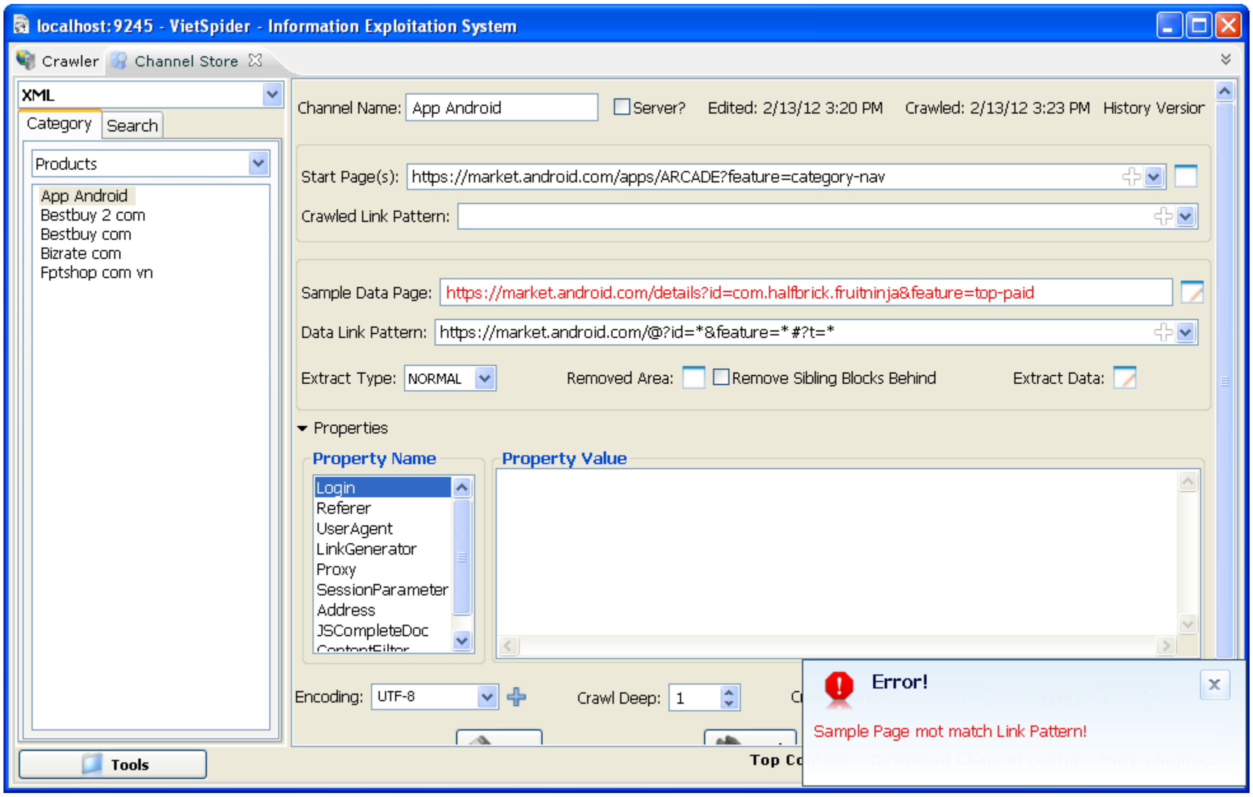
21. Copy a sample data link



Copy a sample data link

Copy a sample data link for testing defined extract data area.

22. Test and repair Data Link Pattern



localhost:9245 - VietSpider - Information Exploitation System

Channel Name: App Android □Server? Edited: 2/13/12 3:20 PM Crawled: 2/13/12 3:23 PM History Versior

Start Page(s): https://market.android.com/apps/ARCADE?feature=category-nav

Crawled Link Pattern:

Sample Data Page: https://market.android.com/details?id=com.halfbrick.fruitninja&feature=top-paid

Data Link Pattern: https://market.android.com/@?id=*&feature=*#?t=*

Extract Type: NORMAL Removed Area: □Remove Sibling Blocks Behind Extract Data:

Encoding: UTF-8 Crawl Deep: 1

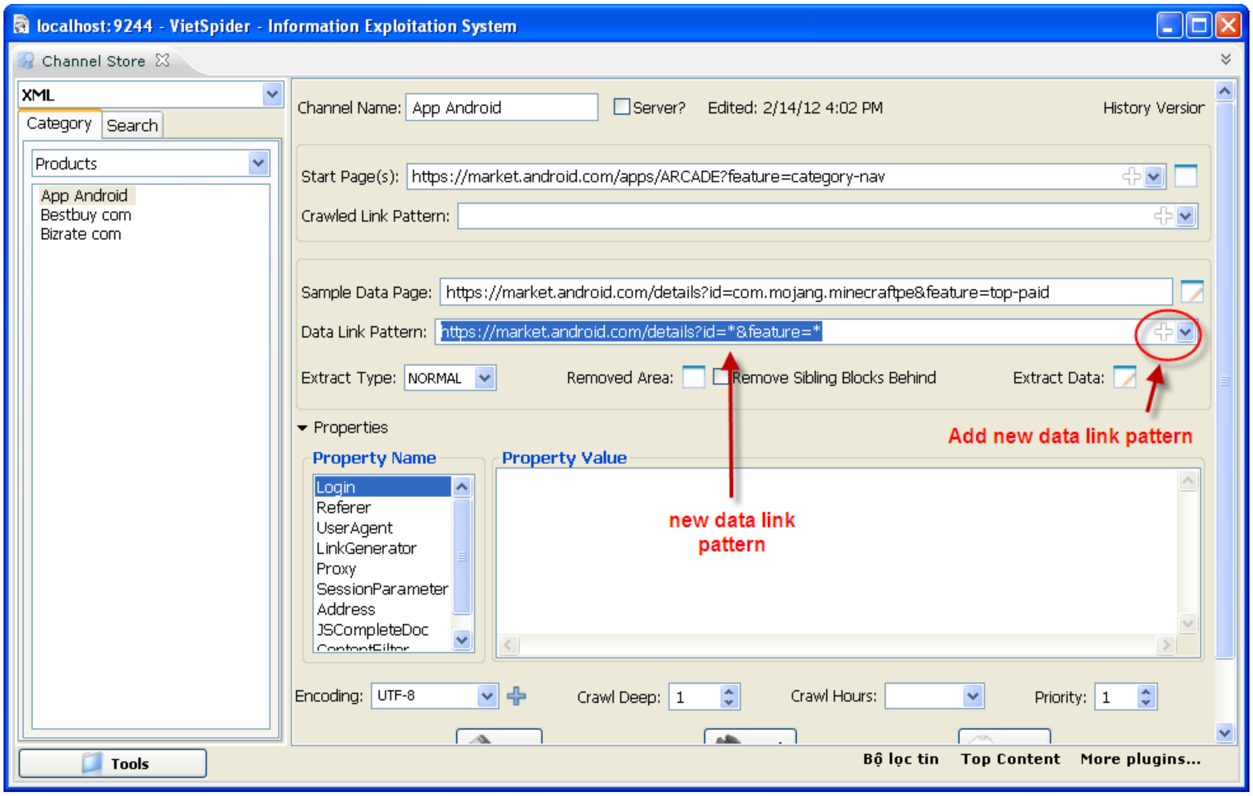
Error!

Sample Page mot match Link Pattern!

Paste the link to Sample Data Page text box, you can see the error message from VietSpider. Please add the below patterns to Data Link Pattern to fix it.

https://market.android.com/details?id=*&feature=*

23. Add New Data Link Pattern



Channel Name: App Android □ Server? Edited: 2/14/12 4:02 PM History Versior

Start Page(s): https://market.android.com/apps/ARCADE?feature=category-nav

Crawled Link Pattern:

Sample Data Page: https://market.android.com/details?id=com.mojang.minecraftpe&feature=top-paid

Data Link Pattern: https://market.android.com/details?id=*&feature=*

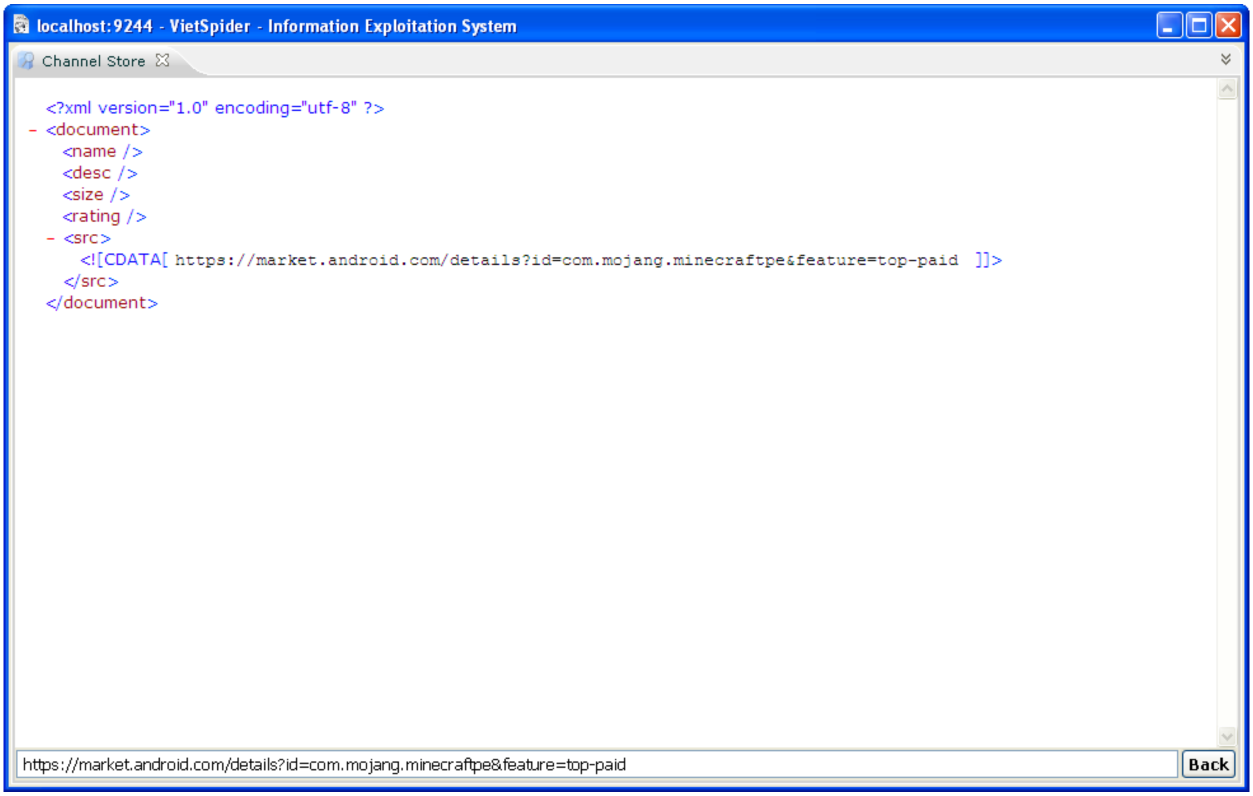
Extract Type: NORMAL Removed Area: □ Remove Sibling Blocks Behind Extract Data:

Encoding: UTF-8 Crawl Deep: 1 Crawl Hours: Priority: 1

Copy and Paste new data link pattern to the text box and click Add icon at the end text box.

Click Test button

24. Test extracted data.

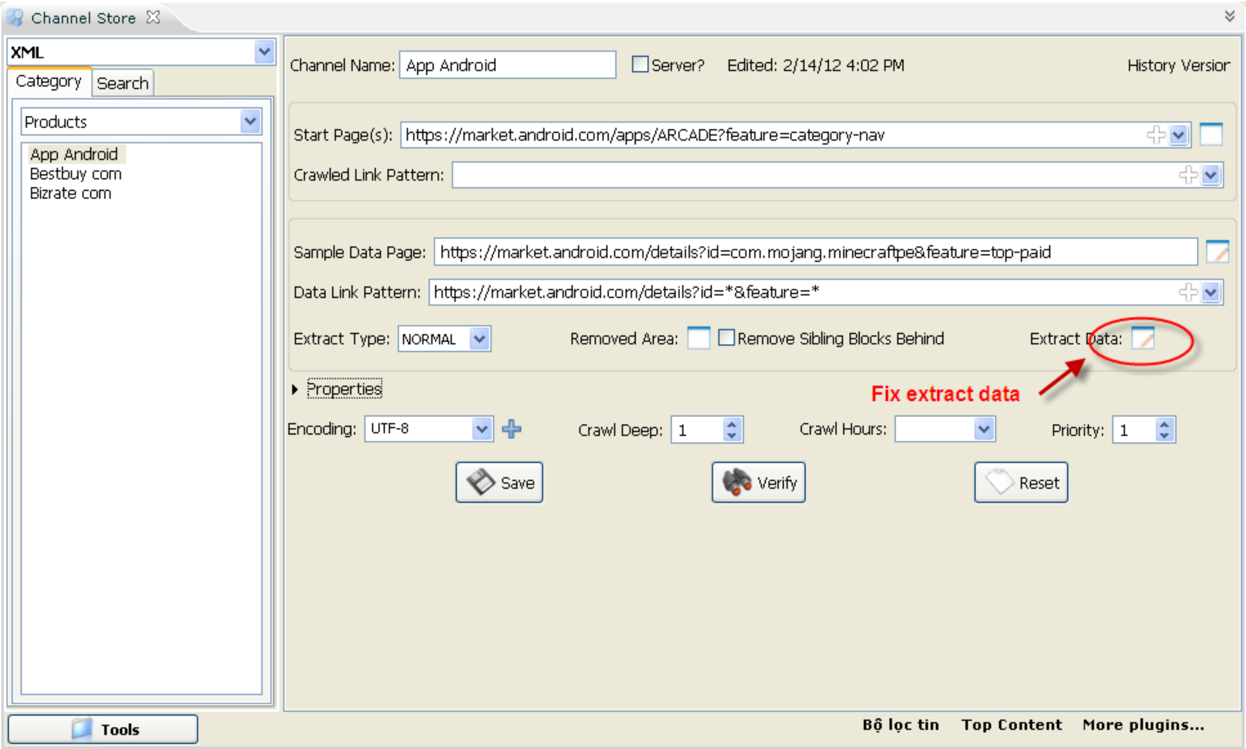

localhost:9244 - VietSpider - Information Exploitation System
Channel Store
Back

```
  <?xml version="1.0" encoding="utf-8" ?>
- <document>
    <name />
    <desc />
    <size />
    <rating />
  - <src>
      <![CDATA[ https://market.android.com/details?id=com.mojang.minecraftpe&feature=top-paid  ]]>
    </src>
  </document>
```

https://market.android.com/details?id=com.mojang.minecraftpe&feature=top-paid

When you click on Test button, you can see extracted data document is empty.

25. Fix extract data for XML element.



Channel Name: App Android □ Server? Edited: 2/14/12 4:02 PM

Start Page(s): https://market.android.com/apps/ARCADE?feature=category-nav

Crawled Link Pattern:

Sample Data Page: https://market.android.com/details?id=com.mojang.minecraftpe&feature=top-paid

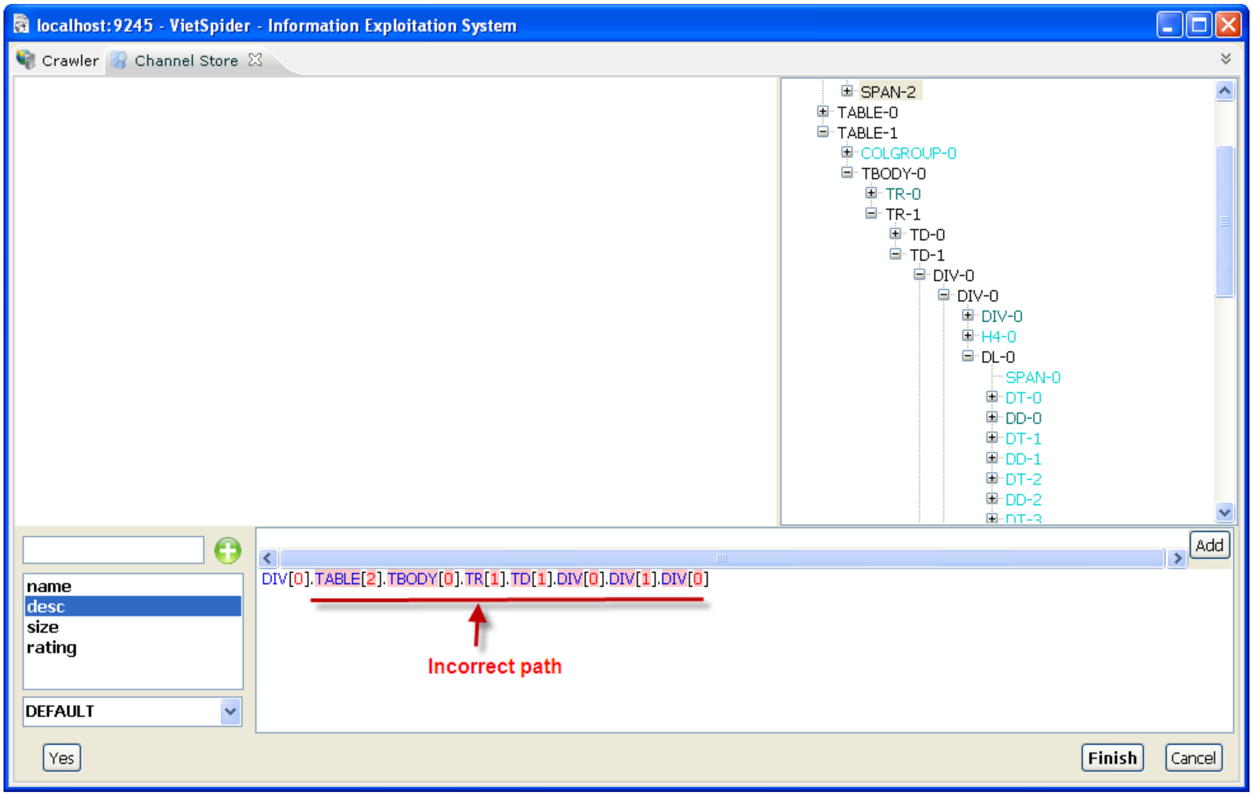
Data Link Pattern: https://market.android.com/details?id=*&feature=*

Extract Type: NORMAL Removed Area: □ Remove Sibling Blocks Behind Extract Data:

Encoding: UTF-8 Crawl Deep: 1 Crawl Hours: Priority: 1

Please review extract data and extract region, click icon.

26. Update incorrect path

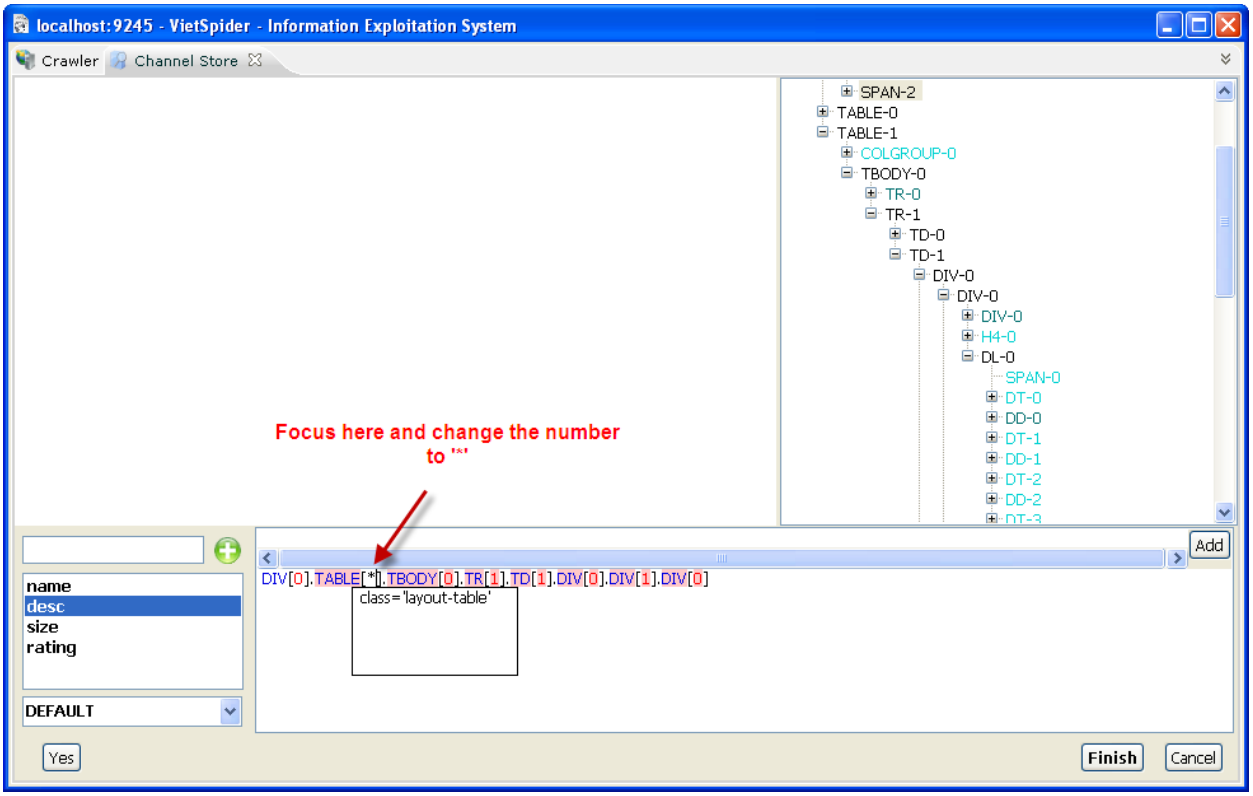


DIV[0].TABLE[2].TBODY[0].TR[1].TD[1].DIV[0].DIV[1].DIV[0]

Incorrect path

name
desc
size
rating

DEFAULT



Focus here and change the number to '*'

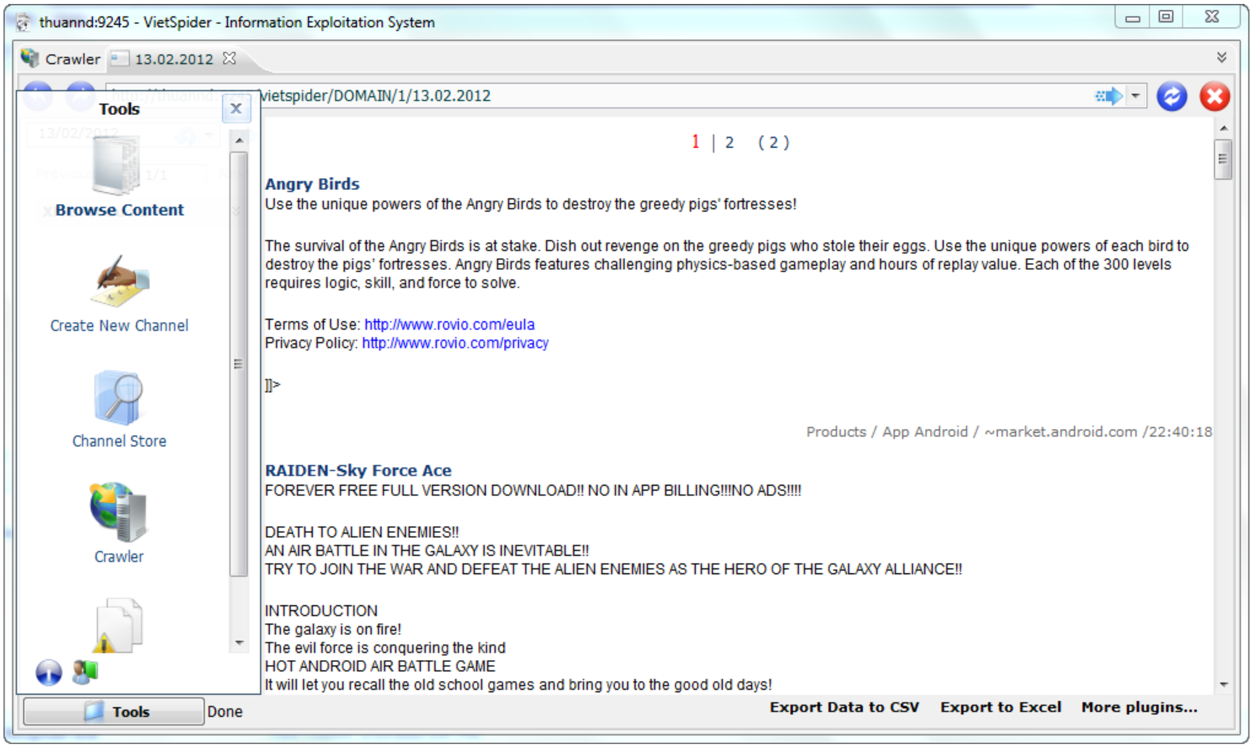
DIV[0].TABLE[*].TBODY[0].TR[1].TD[1].DIV[0].DIV[1].DIV[0]

class='layout-table'

Review the path of XML elements and the path of extract region, update when you see the path is incorrect. Save the channel and re-crawl data when you done.

Part 4: Browse the crawled Data

Click Tools -> Browse Content



1 | 2 (2)

Angry Birds
Use the unique powers of the Angry Birds to destroy the greedy pigs' fortresses!

The survival of the Angry Birds is at stake. Dish out revenge on the greedy pigs who stole their eggs. Use the unique powers of each bird to destroy the pigs' fortresses. Angry Birds features challenging physics-based gameplay and hours of replay value. Each of the 300 levels requires logic, skill, and force to solve.

Terms of Use: http://www.rovio.com/eula
Privacy Policy: http://www.rovio.com/privacy

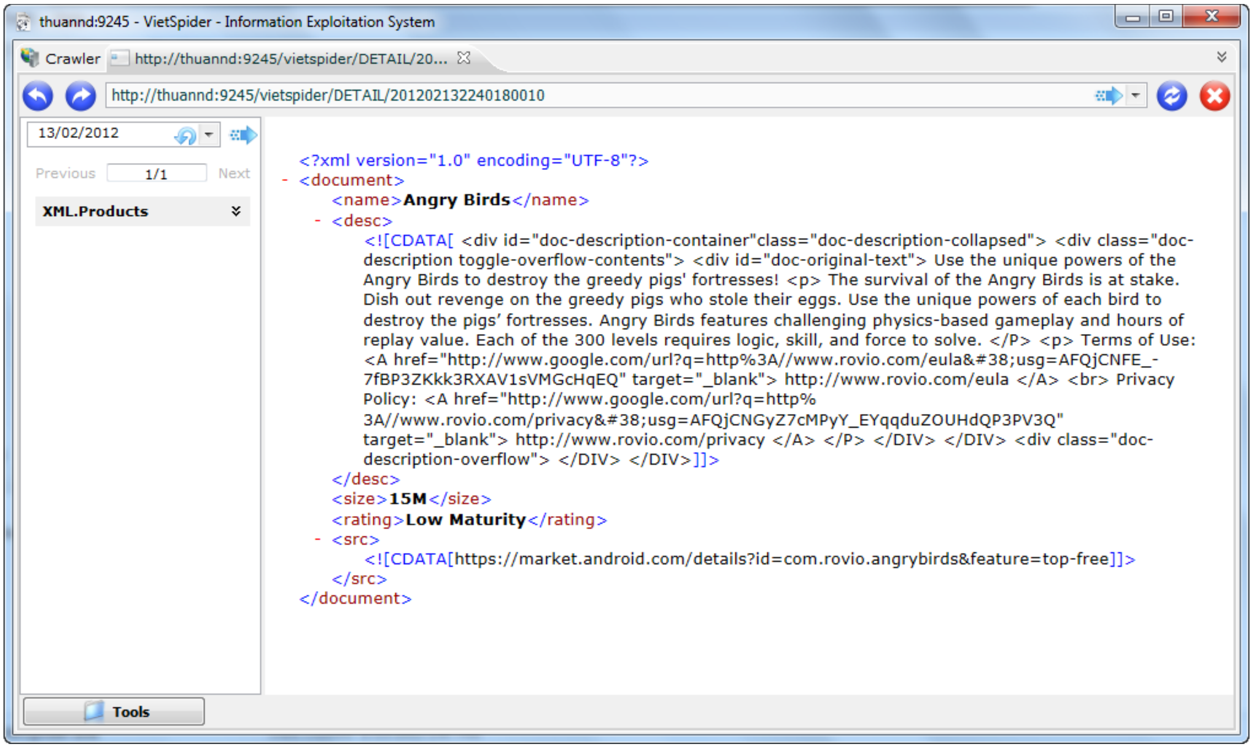
]]>

Products / App Android / ~market.android.com /22:40:18

RAIDEN-Sky Force Ace
FOREVER FREE FULL VERSION DOWNLOAD!! NO IN APP BILLING!!!NO ADS!!!!

DEATH TO ALIEN ENEMIES!!
AN AIR BATTLE IN THE GALAXY IS INEVITABLE!!
TRY TO JOIN THE WAR AND DEFEAT THE ALIEN ENEMIES AS THE HERO OF THE GALAXY ALLIANCE!!

INTRODUCTION
The galaxy is on fire!
The evil force is conquering the kind
HOT ANDROID AIR BATTLE GAME
It will let you recall the old school games and bring you to the good old days!


thuannd:9245 - VietSpider - Information Exploitation System
Crawler
http://thuannd:9245/vietspider/DETAIL/20...
http://thuannd:9245/vietspider/DETAIL/201202132240180010
13/02/2012
Previous
1/1
Next
XML.Products
Tools

```
  <?xml version="1.0" encoding="UTF-8"?>
- <document>
    <name>Angry Birds</name>
  - <desc>
      <![CDATA[ <div id="doc-description-container"class="doc-description-collapsed"> <div class="doc-
      description toggle-overflow-contents"> <div id="doc-original-text"> Use the unique powers of the
      Angry Birds to destroy the greedy pigs' fortresses! <p> The survival of the Angry Birds is at stake.
      Dish out revenge on the greedy pigs who stole their eggs. Use the unique powers of each bird to
      destroy the pigs’ fortresses. Angry Birds features challenging physics-based gameplay and hours of
      replay value. Each of the 300 levels requires logic, skill, and force to solve. </P> <p> Terms of Use:
      <A href="http://www.google.com/url?q=http%3A//www.rovio.com/eula&#38;usg=AFQjCNFE_-
      7fBP3ZKkk3RXAV1sVMGcHqEQ" target="_blank"> http://www.rovio.com/eula </A> <br> Privacy
      Policy: <A href="http://www.google.com/url?q=http%
      3A//www.rovio.com/privacy&#38;usg=AFQjCNGyZ7cMPyY_EYqqduZOUHdQP3PV3Q"
      target="_blank"> http://www.rovio.com/privacy </A> </P> </DIV> </DIV> <div class="doc-
      description-overflow"> </DIV> </DIV>]]>
    </desc>
    <size>15M</size>
    <rating>Low Maturity</rating>
  - <src>
      <![CDATA[https://market.android.com/details?id=com.rovio.angrybirds&feature=top-free]]>
    </src>
  </document>
```