
ddt Documentation

Release 1.0

Yamuna Krishnamurthy

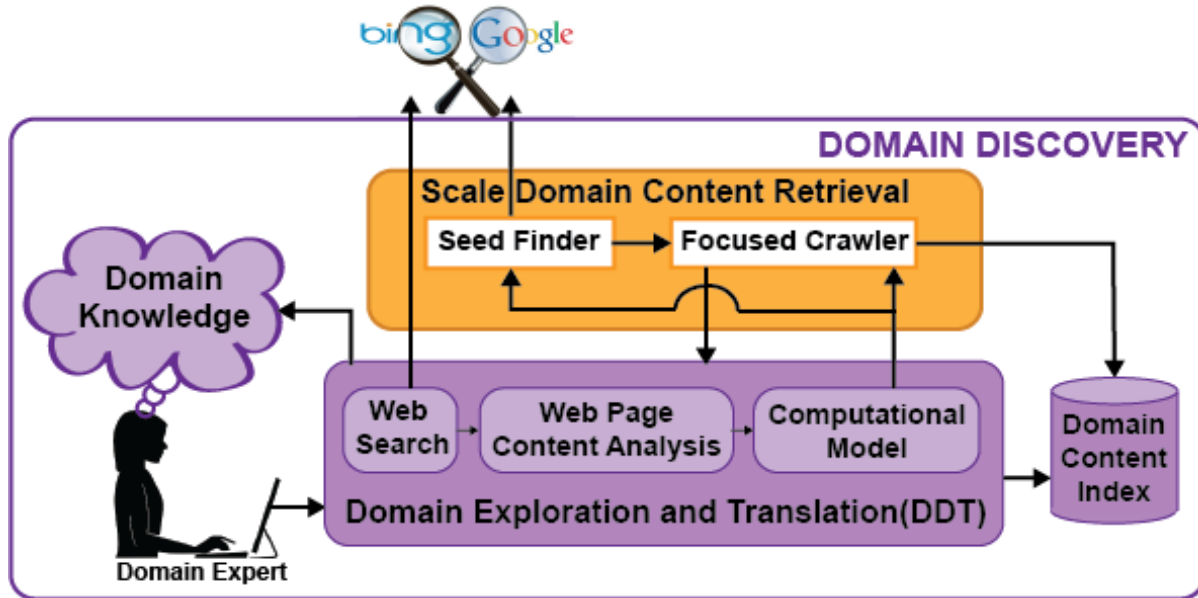
Dec 11, 2021

Contents

1	Contents	3
1.1	Install and Run	3
1.2	Getting Started	6
1.3	How To	19
1.4	Publication	45
1.5	Contact	46
2	Links	47
3	Indices and tables	49

Domain Discovery is the process of acquiring, understanding and exploring data for a specific domain. Some example domains include human trafficking, illegal sale of weapons and micro-cap fraud. Before a user starts the domain discovery process, she has an “idea” of what she is looking for based on prior knowledge. During domain discovery, the user obtains additional knowledge about how the information she is looking for is represented on the Web. This new knowledge of the domain becomes prior knowledge, leading to an iterative process of domain discovery as illustrated in Figure 2. The goals of the domain discovery process are:

- Help users learn about a domain and how (and where) it is represented on the Web.
- Acquire a sufficient number of Web pages that capture the user's notion of the domain so that a computational model can be constructed to automatically recognize relevant content.



The Domain Discovery Tool (DDT) is an interactive system that helps explore and better understand a domain (or topic) as it is represented on the Web. It achieves this by integrating human insights with machine computation (data mining and machine learning) through visualization. DDT allows a domain expert to visualize and analyze pages returned by a search engine or a crawler, and easily provide feedback about relevance. This feedback, in turn, can be used to address two challenges:

- Assist users in the process of domain understanding and discovery, guiding them to construct effective queries to be issued to a search engine to find additional relevant information;
- Provide an easy-to-use interface whereby users can quickly provide feedback regarding the relevance of pages which can then be used to create learning classifiers for the domains of interest; and
- Support the configuration and deployment of focused crawlers that automatically and efficiently search the Web for additional pages on the topic. DDT allows users to quickly select crawling seeds as well as positive and negatives required to create the page classifier required for the focus topic.

1.1 Install and Run

You can install the system from source or using Docker.

1.1.1 Docker Version

You must have docker installed ([Docker Installation for Mac](#) , [Docker Installation for Ubuntu](#))

Background Mode

You must have docker compose installed to run the background version. For Mac docker-compose is included in the docker installation. For Ubuntu follow instructions under Linux tab in [docker compose install for linux](#)

In order to run the docker version in background download:

To run only DDT (no crawlers): Download `docker-compose.yml`.

To run DDT, deep crawler and focused crawler: Download the following files in the same directory

`docker-compose.yml.ache`. Rename the downloaded **`docker-compose.yml.ache`** to **`docker-compose.yml`**.
`ache.yml`

Now use the following commands to run DDT (and crawlers if applicable):

```
>>> cd {path-to-downloaded-docker-compose.yml}
>>> docker-compose up -d
```

The above commands will start elasticsearch and DDT processes (and crawlers if applicable). The elasticsearch and DDT (and crawler if applicable) data are stored in the directory `{path-to-downloaded-docker-compose.yml}/data`

You can check the output of the DDT tool using:

```
>>> docker logs dd_tool
```

You will see a message **“ENGINE Bus STARTED”** when DDT is running successfully. You can now use DDT.

Use Domain Discovery Tool

To shutdown the processes run:

```
>>> cd {path-to-downloaded-docker-compose.yml}
>>> docker-compose stop
```

Interactive Mode

To run using the interactive docker version download the script `run_docker_ddt` and run it:

```
>>> cd {path-to-downloaded-run_docker_ddt}
>>> chmod a+x run_docker_ddt
>>> ./run_docker_ddt
```

The above script will prompt to enter a directory where you would like to persist all the web pages for the domains you create. You can enter the path to a directory on the host you are running DDT or just press **Enter** to use the default directory which is `{path-to-downloaded-run_docker_ddt}/data`. The data is stored in the [elasticsearch](#) data format (You can later use this directory as the data directory to any elasticsearch). The script will start elasticsearch with the data directory provided.

The script will then start DDT. You will see a message **“ENGINE Bus STARTED”** when DDT is running successfully. You can now use DDT.

Use Domain Discovery Tool

Trouble Shooting

In case you see the following error:

```
>>> ERROR: for elasticsearch Cannot create container for service elasticsearch:
↳ Conflict. The container name "/elastic" is already in use by container
↳ b714e105ccbf3a6d5a718c76c2ce1e5a51ea6f10a5f4997a6e5b12b9c7faf50e. You have to
↳ remove (or rename) that container to be able to reuse that name.
```

run the following command:

```
>>> docker rm elastic
```

In case you see the following error:

```
>>> ERROR: for ddt Cannot create container for service ddt: Conflict. The container
↳ name "/dd_tool" is already in use by container
↳ 326881fda035692aa0a5c03ec808294aaad2f9fd816baa13270d2fe50e7e1e77. You have to
↳ remove (or rename) that container to be able to reuse that name.
```

```
>>> docker rm dd_tool
```


1.1.2 Local development

Building and deploying the Domain Discovery Tool can be done using its Makefile to create a local development environment. The conda build environment is currently only supported on 64-bit OS X and Linux.

Install Conda

First install `conda` (anaconda) for python 2.7.

Install Java

Install `JDK 1.8`.

Install Elasticsearch

Download Elasticsearch 1.6.2 [here](#), extract the file and run Elasticsearch:

```
>>> cd {path-to-installed-Elasticsearch}
>>> ./bin/elasticsearch
```

Install Domain Discovery API

```
>>> git clone https://github.com/ViDA-NYU/domain_discovery_API
>>> cd domain_discovery_API
```

The *make* command builds `dd_api` and downloads/installs its dependencies.

```
>>> make
```

Add `domain_discovery_API` to the environment:

```
>>> export DD_API_HOME="{path-to-cloned-domain_discovery_API-repository}"
```

Clone the DDT repository and enter it:

```
>>> git clone https://github.com/ViDA-NYU/domain_discovery_tool
>>> cd domain_discovery_tool
```

Use the *make* command to build `ddt` and download/install its dependencies.

```
>>> make
```

After a successful installation, you can activate the DDT development environment:

```
>>> source activate ddt
```

(from the top-level *domain_discovery_tool* directory) execute:

```
>>> ./bin/ddt-dev
```

Use Domain Discovery Tool

1.2 Getting Started

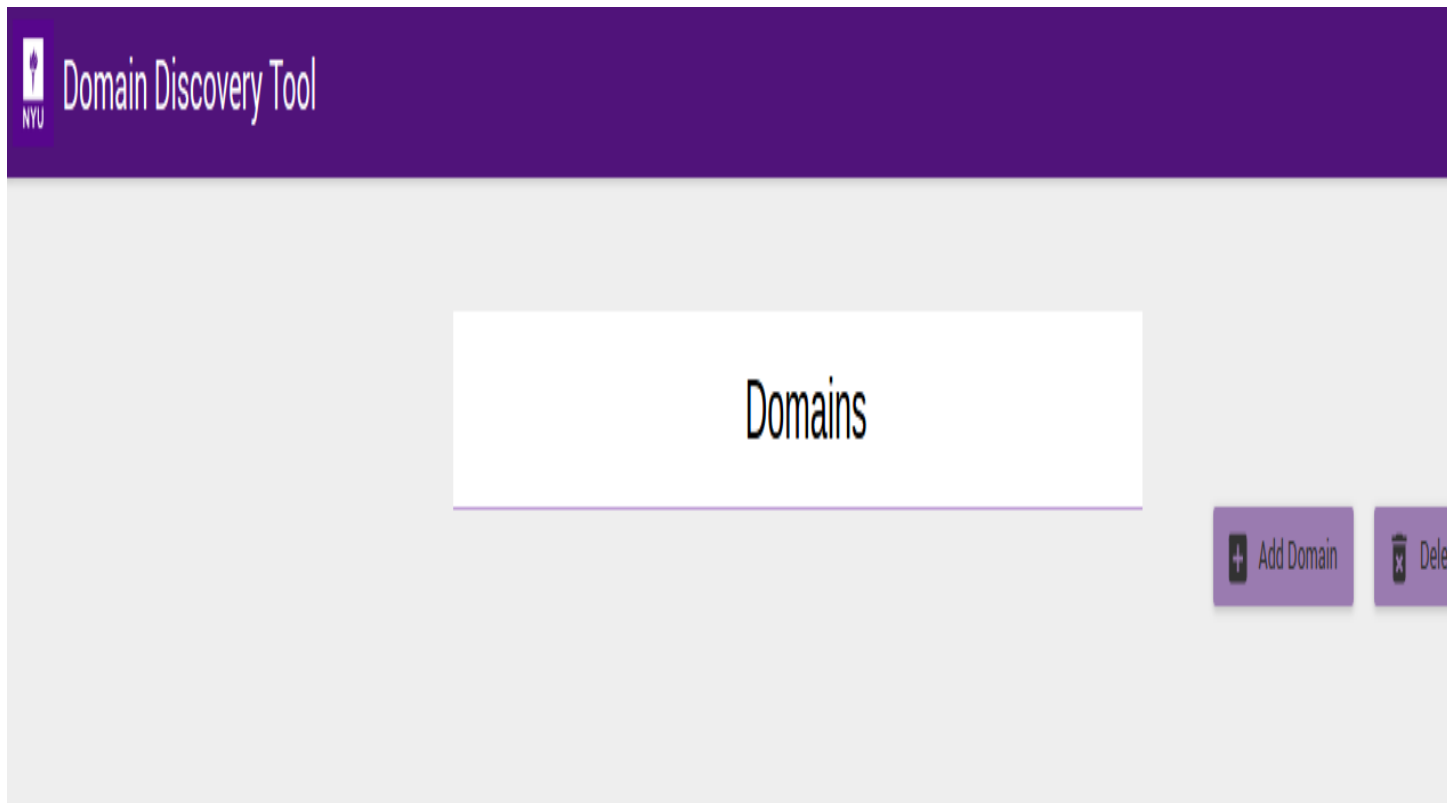
1.2.1 Building Domain Index

Creating a domain specific index involves:

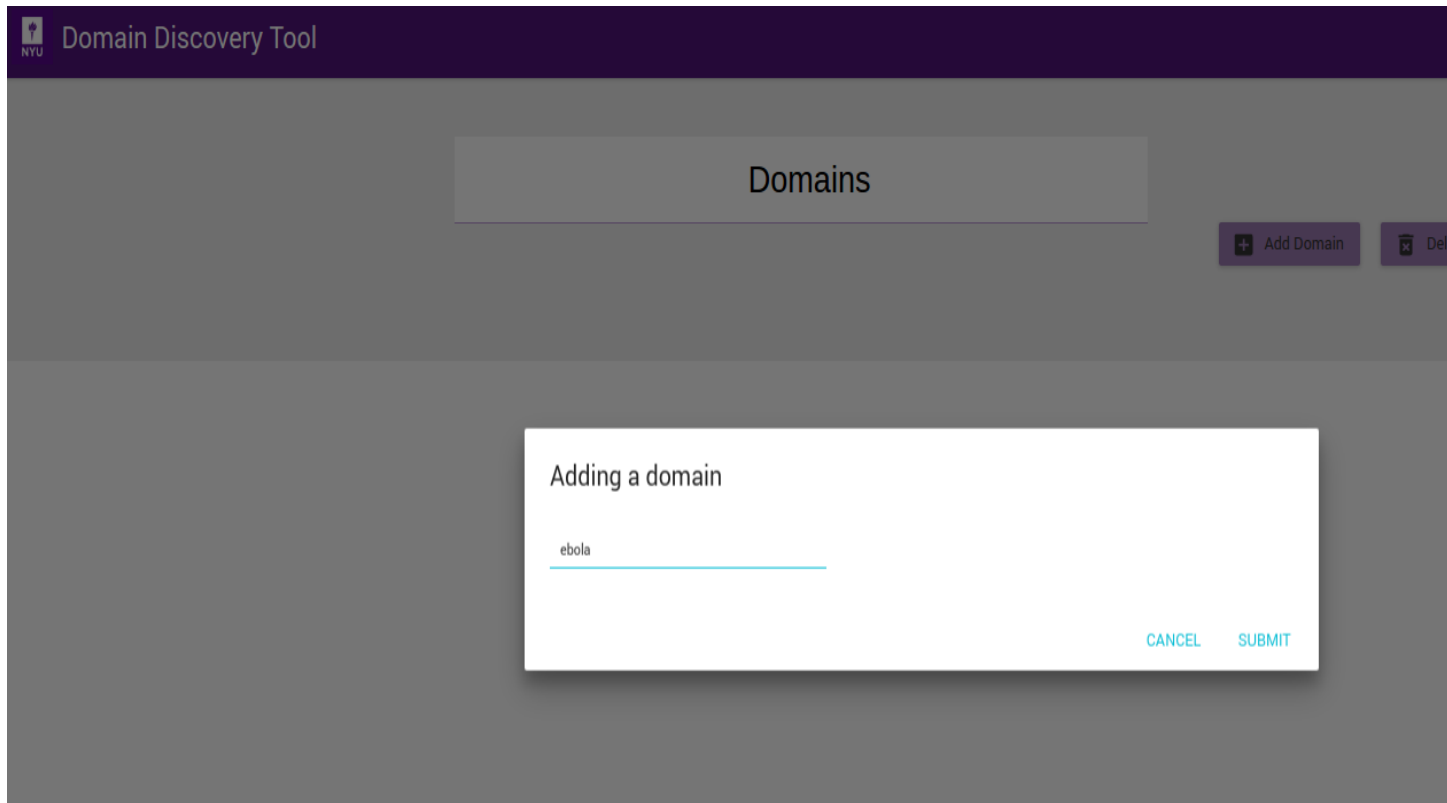
- Uploading known relevant URL domains from which you would like to collect all URLs belonging to that domain. This is called a deep crawl. Follow steps 1,2 and 4, below, for this.
- Create a domain model that can be used for a focused crawl (broad crawl). For this follow all the steps 1-4 below.

Step 1

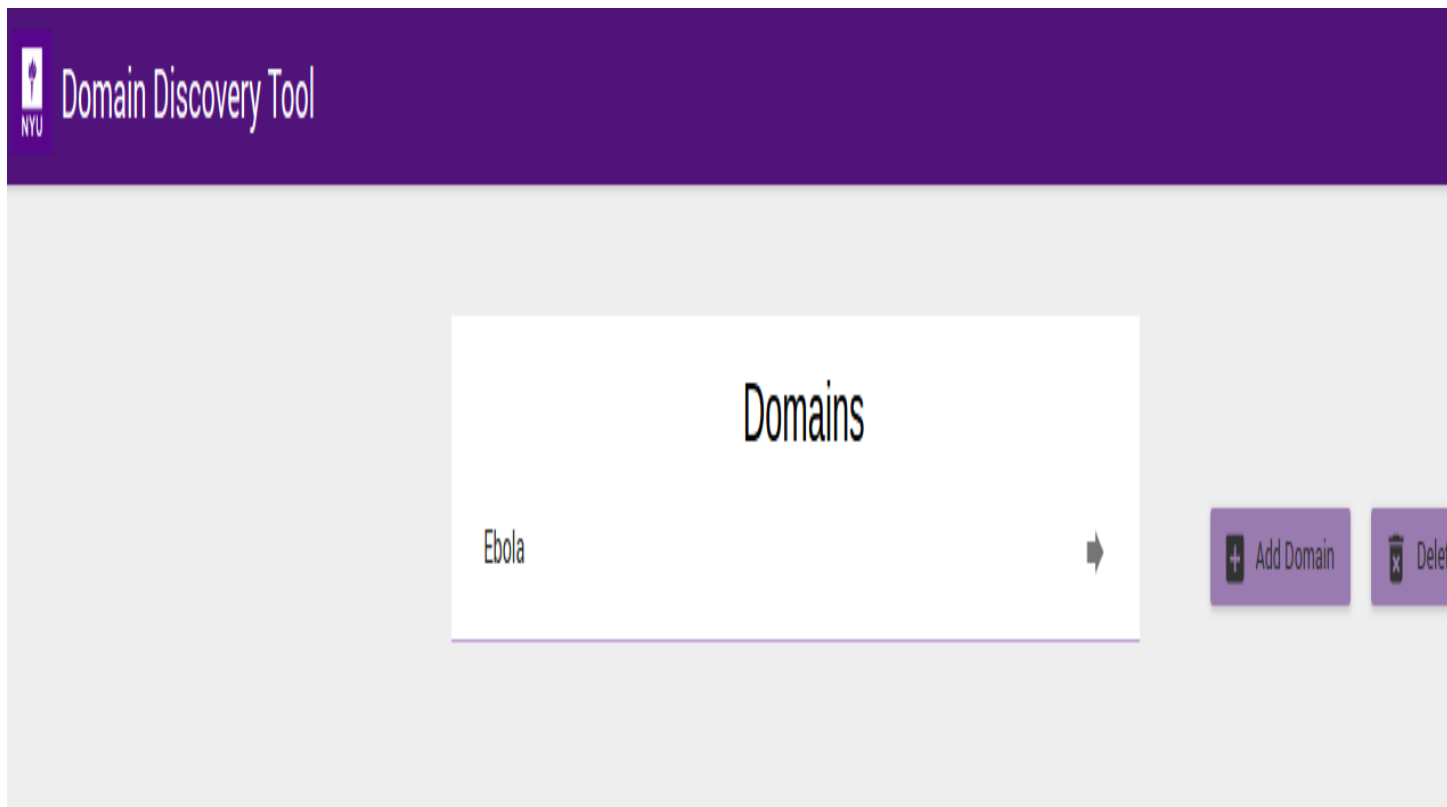
Create Domain



Begin by adding a domain on the Domains page (initial page), shown in the figure above, by clicking on the **Add Domain** button. Domain maintains context of domain discovery.



On the **Adding a domain** dialog shown in figure above, enter the name of the domain you would like to create, for example **Ebola**, and click on **Submit** button. You should now see the new domain you added in the list of domains as shown below.



Once domain is added click on domain name in the list of domains to collect, analyse and annotate web pages.

Step 2

Acquire Data

Continuing with our example of the **Ebola** domain, we show here the methods of uploading data. Expand the Search tab on the left panel. You can add data to the domain in the following ways:

Upload URLs

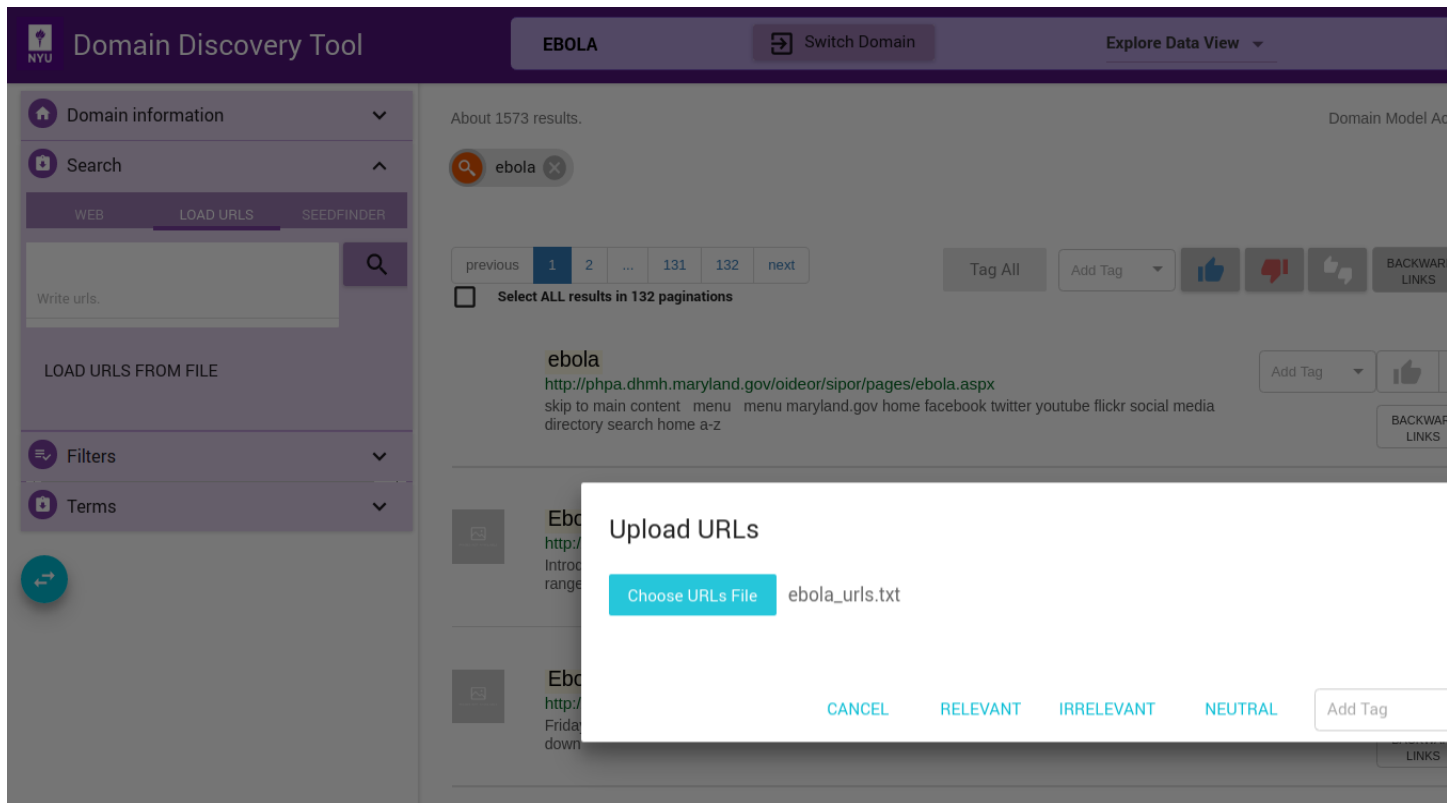
If you have a set of URLs of sites you already know, you can add them from the **LOAD** tab. You can upload the list of URLs in the text box, one fully qualified URL per line, as shown in figure below:

The screenshot shows the Domain Discovery Tool (DDT) interface. The top bar displays the domain 'EBOLA' and a 'Switch Domain' button. The left sidebar contains navigation options: Domain information, Search, WEB, LOAD URLS (selected), SEEDFINDER, LOAD URLS FROM FILE, Filters, and Terms. The main area displays search results for 'ebola', showing about 1575 results. The first result is from 'http://phpa.dhmmh.maryland.gov/oideor/sipor/pages/ebola.aspx' with a 'Deep Crawl' button. The second result is from 'http://www.cnn.com/ebola/' with a 'Deep Crawl' button. The interface includes pagination controls, a 'Tag All' button, and social media sharing options.

You can also upload a file with the list of URLs by clicking on the **LOAD URLS FROM FILE** button. This will bring up a file explorer window where you can select the file to upload. *The list of fully qualified URLs should be entered one per line in the file.* For example:

```
http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1003065
https://bmcpsy psychiatry.biomedcentral.com/articles/10.1186/s12888-017-1280-8
http://www.cdph.ca.gov/programs/cder/Pages/Ebola.aspx
```

Download an example URLs list file for ebola domain [HERE](#). Once the file is selected you can upload them by clicking on **RELEVANT**, **IRRELEVANT**, **NEUTRAL** or **Add Tag** (Add a custom tag). This will annotate the pages correspondingly.



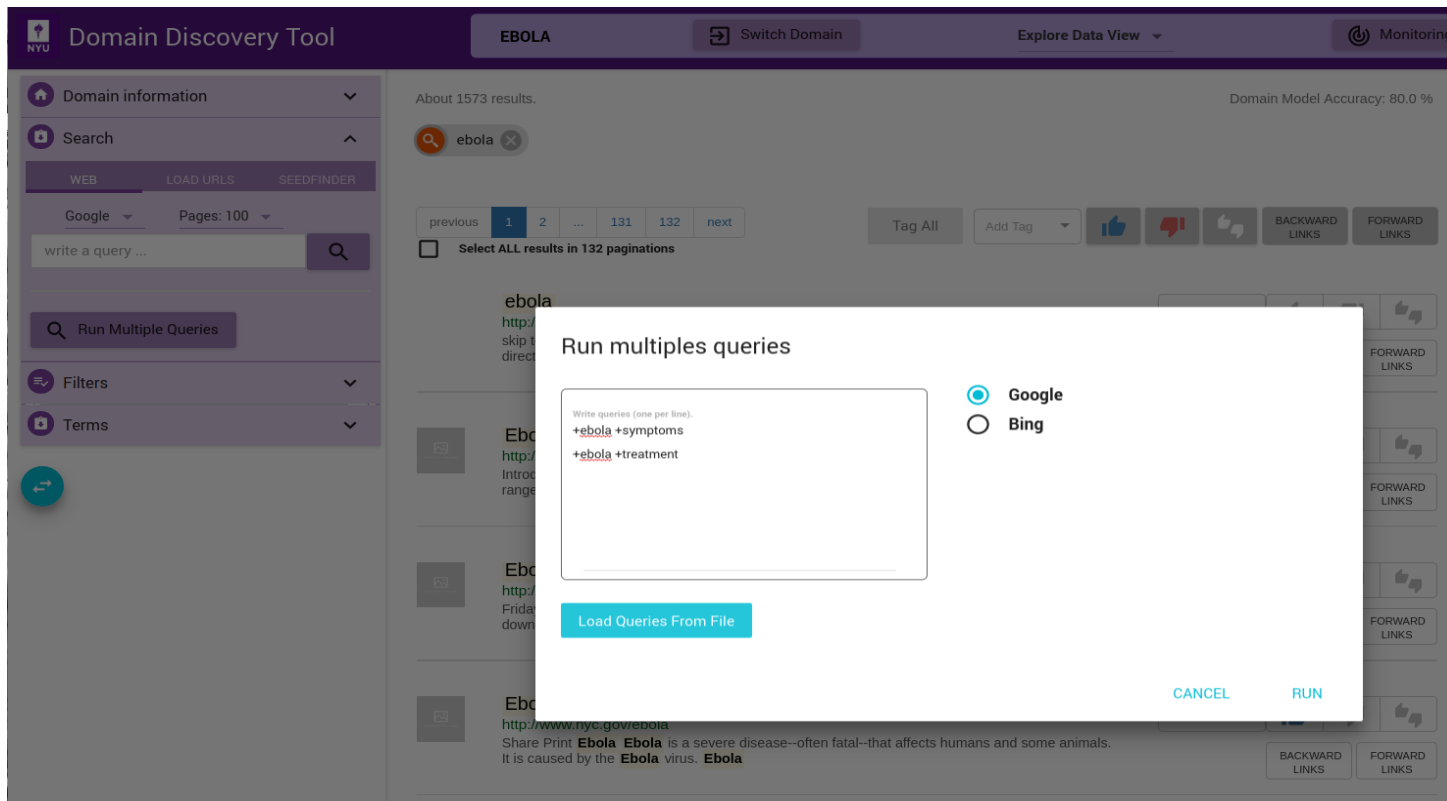
The uploaded URLs are listed in the **Filters** Tab under **Queries** as **Uploaded**.

Web Search

You can do a keywords search on google or bing by clicking on the **WEB** tab. For example, “ebola symptoms”. All queries made are listed in the **Filters** Tab under **Queries**.

The screenshot displays the Domain Discovery Tool (DDT) interface. The top navigation bar is purple and contains the NYU logo, the title 'Domain Discovery Tool', a search bar with 'EBOLA', a 'Switch Domain' button, an 'Explore Data View' dropdown, and a 'Monitor' button. The left sidebar is also purple and contains a 'Domain information' dropdown, a 'Search' dropdown, and a 'Filters' dropdown. The main content area is white and shows search results for 'ebola symptoms'. The search bar at the top of the main area contains 'ebola symptoms' and a magnifying glass icon. Below the search bar, there are tabs for 'WEB', 'LOAD URLS', and 'SEEDFINDER'. The 'WEB' tab is selected. Below the tabs, there is a 'Google' dropdown and a 'Pages: 100' dropdown. A search button with a magnifying glass icon is to the right of the search bar. Below the search bar, there is a 'Run Multiple Queries' button. The search results section shows 'About 86 results.' and a list of results. The first result is 'Signs and Symptoms | Ebola Hemorrhagic Fever | CDC' with a URL 'https://www.cdc.gov/vhf/ebola/symptoms/index.html'. The second result is 'Ebola Virus: Symptoms, Treatment, and Prevention' with a URL 'https://www.webmd.com/a-to-z-guides/ebola-fever-virus-infection'. The third result is 'Ebola virus and Marburg virus - Symptoms and causes - Mayo Clinic' with a URL 'https://www.mayoclinic.org/diseases-conditions/ebola-virus/symptoms-causes/dxc-20338674'. Each result has an 'Add Tag' button and a 'BACKLINK' button. The interface is clean and modern, with a purple and white color scheme.

If you have a multiple search queries then you can load them by clicking on the **Run Multiple Queries** button. This will bring up a window where you can either add the queries one per line in a textbox or upload a file that contains the search queries one per line. You can select the search engine to use (**Google** or **Bing**):



Each of the queries will be issued on Google or Bing (as chosen) and the results made available for exploration and annotation in the **Filters** Tab under **Queries** as **Uploaded**.

Step 3

Annotate Pages

A model is created by annotating pages as **Relevant** or **Irrelevant** for the domain. Currently, the model can only distinguish between relevant and irrelevant pages. You can also annotate pages with custom tags. These can be later grouped as relevant or irrelevant when generating the model. Try to alternate between Steps 3a and 3b to build a model till you reach at least 100 pages for each. This will continuously build a model and you can see the accuracy of the model at the top right corner - **Domain Model Accuracy**.

Step 3a

Tag at least 100 **Relevant** pages for your domain. Refer [How to Annotate](#).

Step 3b

Tag at least 100 **Irrelevant** pages for your domain. Refer [How to Annotate](#).

How to Annotate

In the **Explore Data View** you see the pages for the domain (based on any filters applied) in two ways: through **Snippets** and **Visualizations**, as shown below:

The different mechanisms for annotating pages through **Snippet** are:

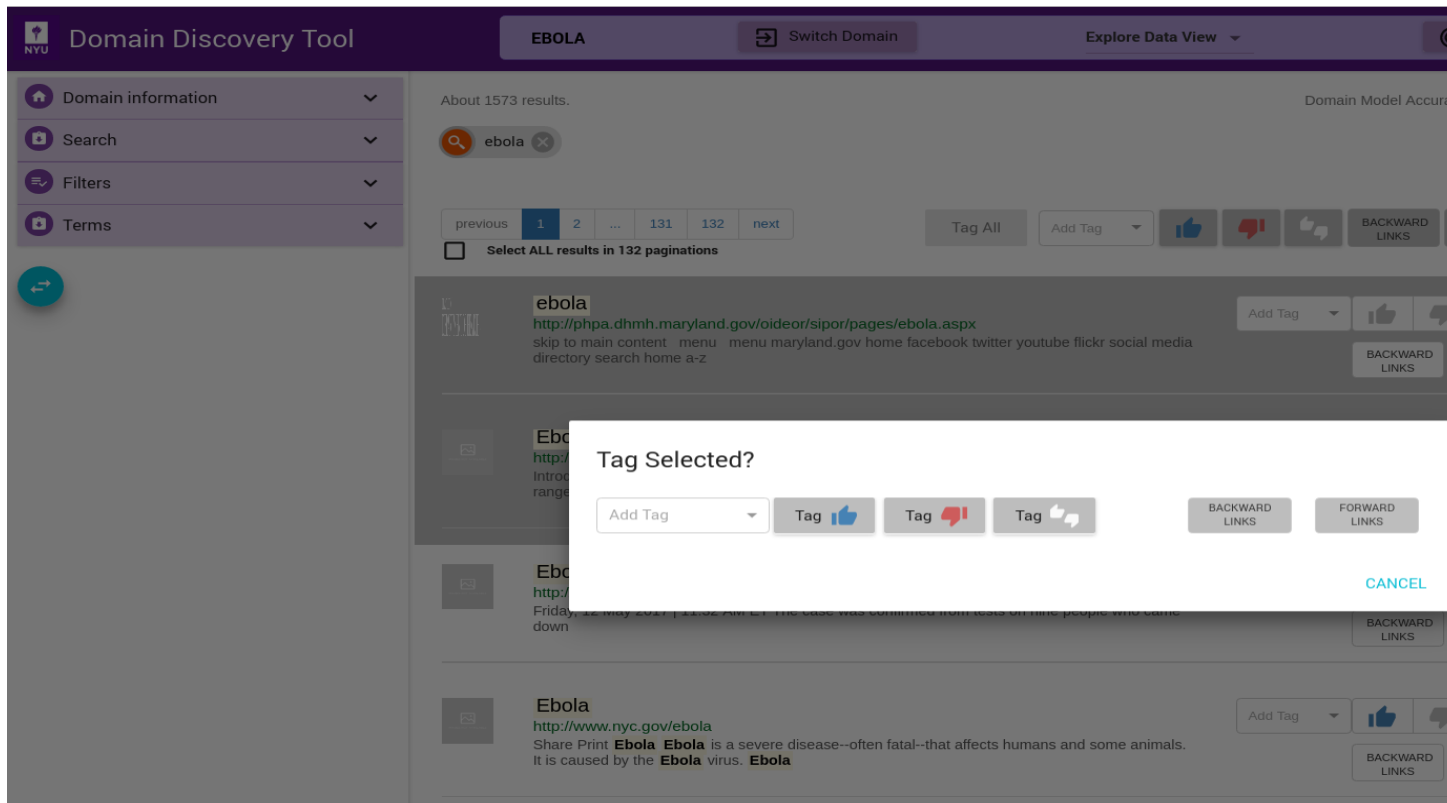
Tag Individual Pages



buttons, along each page, can be used to tag individual pages.

Tag Selected Pages

Select multiple pages by keeping the **ctrl** key pressed and clicking on the pages that you want to select. When done with selecting pages, release the **ctrl** key. This will bring up a window where you can tag the pages as shown below:



Tag All Pages in View



Use the buttons at the top of the list of pages to tag all pages in the current view

Tag All Pages for Current Filter

If you want to tag all pages retrieved for a particular filter (across pagination), then check the **Select ALL results in <total pages> paginations** checkbox below the page list on top left. Then use



buttons to tag all the pages.

Custom Tag

Custom tags can be added using Add Tag text box as shown below. Enter the custom tag in the Add Tag text box and press **enter** key. This adds the tag as a chip below the page info. This can be applied to individual, selected or all pages similar to relevant and irrelevant tags.

The screenshot shows the Domain Discovery Tool (DDT) interface. The top header is purple with the NYU logo and the text 'Domain Discovery Tool'. To the right of the header, there is a search bar containing 'EBOLA', a 'Switch Domain' button, and an 'Explore Data View' dropdown menu. On the left side, there is a sidebar with four menu items: 'Domain information', 'Search', 'Filters', and 'Terms', each with a dropdown arrow. Below the sidebar is a circular icon with two arrows. The main content area shows 'About 5233 results.' and a pagination bar with links for 'previous', '1', '2', '...', '436', '437', and 'next'. Below the pagination bar is a checkbox labeled 'Select ALL results in 437 paginations'. A search result is displayed with a thumbnail image of a green frog, the title 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com', the URL 'http://time.com/4596928/some-people-who-get-ebola-dont-show-symptoms-study/', and the date 'Dec 12, 2016 ... A new report reveals people can get Ebola and not have symptoms.' To the right of the search result is a 'Tag All' button, an 'Add Tag' dropdown menu, and several social media sharing icons (Facebook, Twitter, LinkedIn, etc.). At the bottom of the search result, there is a 'news article' tag with a close button. On the far right, there is a 'Domain Model' dropdown menu and a 'Create option "symptoms"' button.

Tag for Deep Crawl

Some tags such as **Deep Crawl** are pre-configured. User can tag a page (or group of pages) for deep crawl by choosing the tag from the Add Tag drop-down as shown. For example, if user wants to deep crawl all the uploaded pages then they can tag the pages **Deep Crawl**.

The screenshot shows the Domain Discovery Tool (DDT) interface. The top navigation bar is purple and contains the NYU logo, the text 'Domain Discovery Tool', a search bar with 'EBOLA' entered, a 'Switch Domain' button, and a dropdown menu for 'Explore Data View'. On the left, a sidebar menu lists 'Domain information', 'Search', 'Filters', and 'Terms'. The main content area displays 'About 5233 results.' and a pagination bar with links for 'previous', '1', '2', '...', '436', '437', and 'next'. Below the pagination, there is a checkbox labeled 'Select ALL results in 437 paginations'. Two search results are visible. The first result is titled 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com' with a URL and a brief description. The second result is titled 'Ebola symptoms and transmission. :: Washington State Department ...' with a URL and a brief description. Both results have a 'news article' tag. The interface also includes social media sharing icons and a 'Deep Crawl' button.

Step 4

Run Crawler

Once a sufficiently good model is available or pages are tagged for a deep crawl you can change from **Explore Data View** to the **Crawler View** to start the crawl shown below:

The screenshot shows the Domain Discovery Tool (DDT) interface. The top navigation bar is purple and contains the NYU logo, the title 'Domain Discovery Tool', a search bar with 'EBOLA' entered, a 'Switch Domain' button, and a 'Monitor' button. A dropdown menu is open over the search bar, showing 'Explore Data View' and 'Crawling View'. On the left, a sidebar contains links to 'Domain information', 'Search', 'Filters', and 'Terms'. The main content area displays 'About 5233 results.' and a pagination bar with 'previous', '1', '2', '...', '436', '437', and 'next'. Below the pagination, there is a checkbox labeled 'Select ALL results in 437 paginations'. The first search result is a news article titled 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com' with a URL and a brief description. The second result is a document titled 'Ebola symptoms and transmission. :: Washington State Department ...' with a URL and a brief description. Both results have an 'Add Tag' button and a 'BACKLINK' button. A 'Deep Crawl' button is visible at the bottom of the results list.

Step 4a

Deep Crawl

In order to run a *Deep Crawl* annotate pages to be crawled with tag *Deep Crawl* as described in [Tag for Deep Crawl](#).

Domain Discovery Tool | EBOLA | Switch Domain | Crawling View | Monitoring | Search

DEEP CRAWLING | FOCUSED CRAWLING

Domains for crawling

Annotated urls

- http://www.nyc.gov/ebola
- http://phpa.dhmh.maryland.gov/oideor/sipor/pages/ebola.aspx
- http://www.cnn.com/ebola/
- http://rocs.hu-berlin.de/projects/ebola/
- http://www.stanford.edu/group/virus/filo/filo.html

Added urls to deep crawl

Recommendations

DOMAIN	SCORE, COUNT
medicinenet.com	0.054, 10
medicalxpress.com	0.052, 10
nytimes.com	0.049, 11
latimes.com	0.030, 11
cidrap.umn.edu	0.019, 12
time.com	0.016, 11

Min URLs in Don

Start Crawler

Add To Deep Crawl

Loading External Urls

The figure above shows the Deep Crawl View. The list on the left shows all pages annotated as *Deep Crawl* in the Explore Data View. The table on the right shows recommendations of pages that could be added to deep crawl by clicking on the **Add to Deep Crawl**. If keyword terms are added or annotated then recommendations are made based on the score of how many of the keywords they contain. Otherwise the domains are recommended by the number of pages they contain.

The deep crawler can be started by clicking on **Start Crawler** button at the bottom. This starts a deep crawler with all the pages tagged for Deep Crawl.

You can see the results of the crawled data in **Crawled Data** in the Filters Tab. When the crawler is running it can be monitored by clicking on the **Crawler Monitor** button.

Step 4b

Focused Crawl

The figure below shows the Focused Crawler View:

Domain Discovery Tool | EBOLA | Switch Domain | Crawling View | Monitoring

DEEP CRAWLING | FOCUSED CRAWLING

Model Settings

Select positive and negative examples.

Positive

- ☐ Neutral (5065)
- ☒ Relevant (123)
- ☐ Irrelevant (86)
- ☒ Deep Crawl (3)
- ☐ outlier (3)

Negative

- ☐ Neutral (5065)
- ☐ Relevant (123)
- ☒ Irrelevant (86)
- ☐ Deep Crawl (3)
- ☐ outlier (3)

Terms

- ebola virus
- symptoms
- outbreak
- fever
- blood
- ebola
- hemorrhagic
- fluids
- care
- flu
- ebola virus disease

Save **Cancel**

Crawling

Start Crawler

Model

Total Positive: 126
Total Negative: 86
Domain Model (Accuracy): 72.46 %

Ratio/Accuracy: Poor Fair Good Excellent

Export

1. In the 'Model Settings' on the left select the tags that should be considered as relevant(Positive) and irrelevant(Negative). If there sufficient relevant and irrelevant pages (about 100 each), then you can start the crawler by clicking on the **Start Crawler** button.
2. If there are no irrelevant pages then a page classifier model cannot be built. Instead you can either upload keywords by clicking on the 'Add Terms' in the Terms window. You can also annotate the terms extracted from the positive pages by clicking on them. If no annotated terms are available then the top 50 terms are used to build a regular expression model.
3. Once either a page classifier or a regex model is possible start the focused crawler by clicking on the **Start Crawler**.

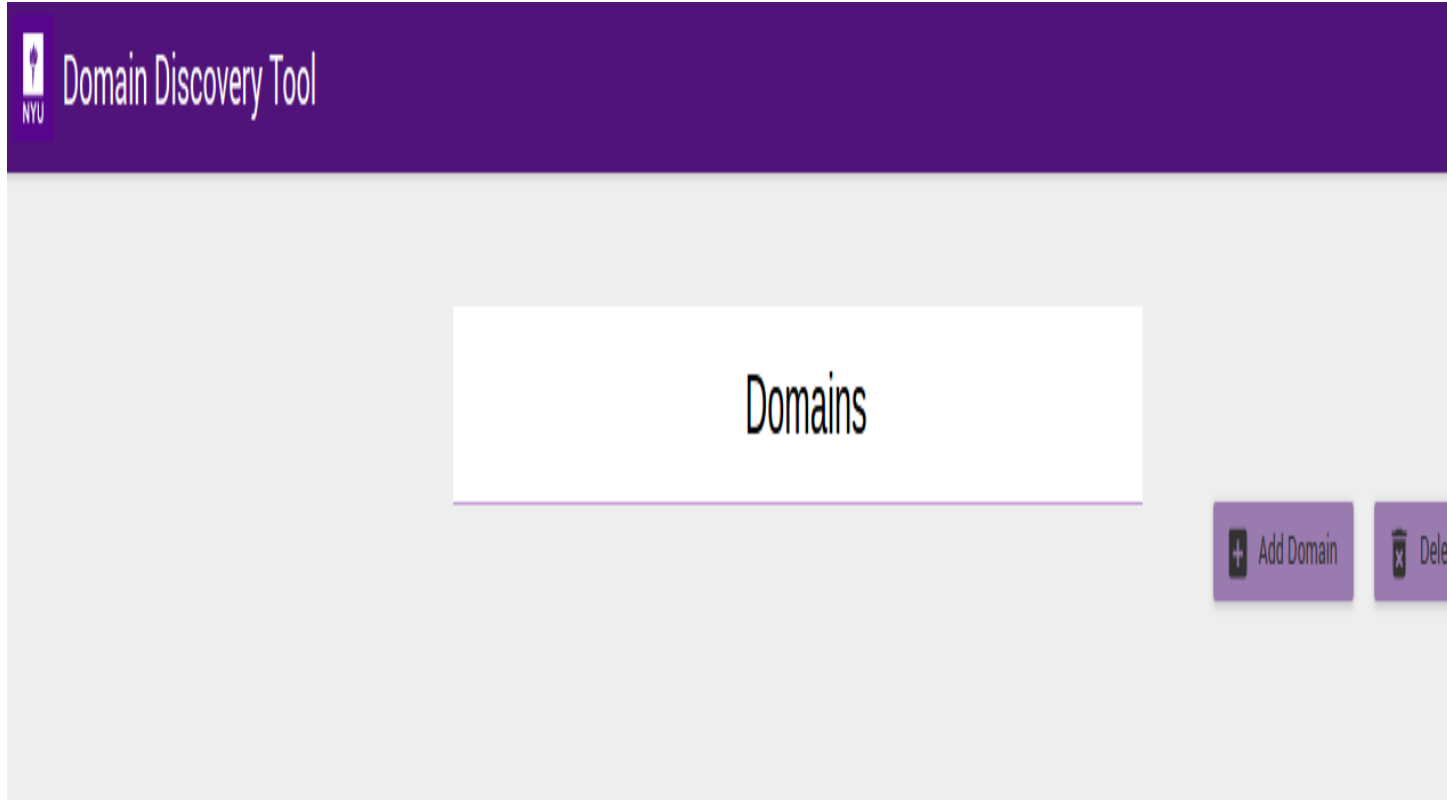
You can see the results of the crawled data in "Crawled Data" in the Filters Tab. When the crawler is running it can be monitored by clicking on the 'Crawler Monitor' button.

The Model info on the bottom right shows how good a domain model is if there are both relevant and irrelevant pages annotated. The color bar shows the strength of the model based on the balance of relevant and irrelevant pages and the classifier accuracy of the model.

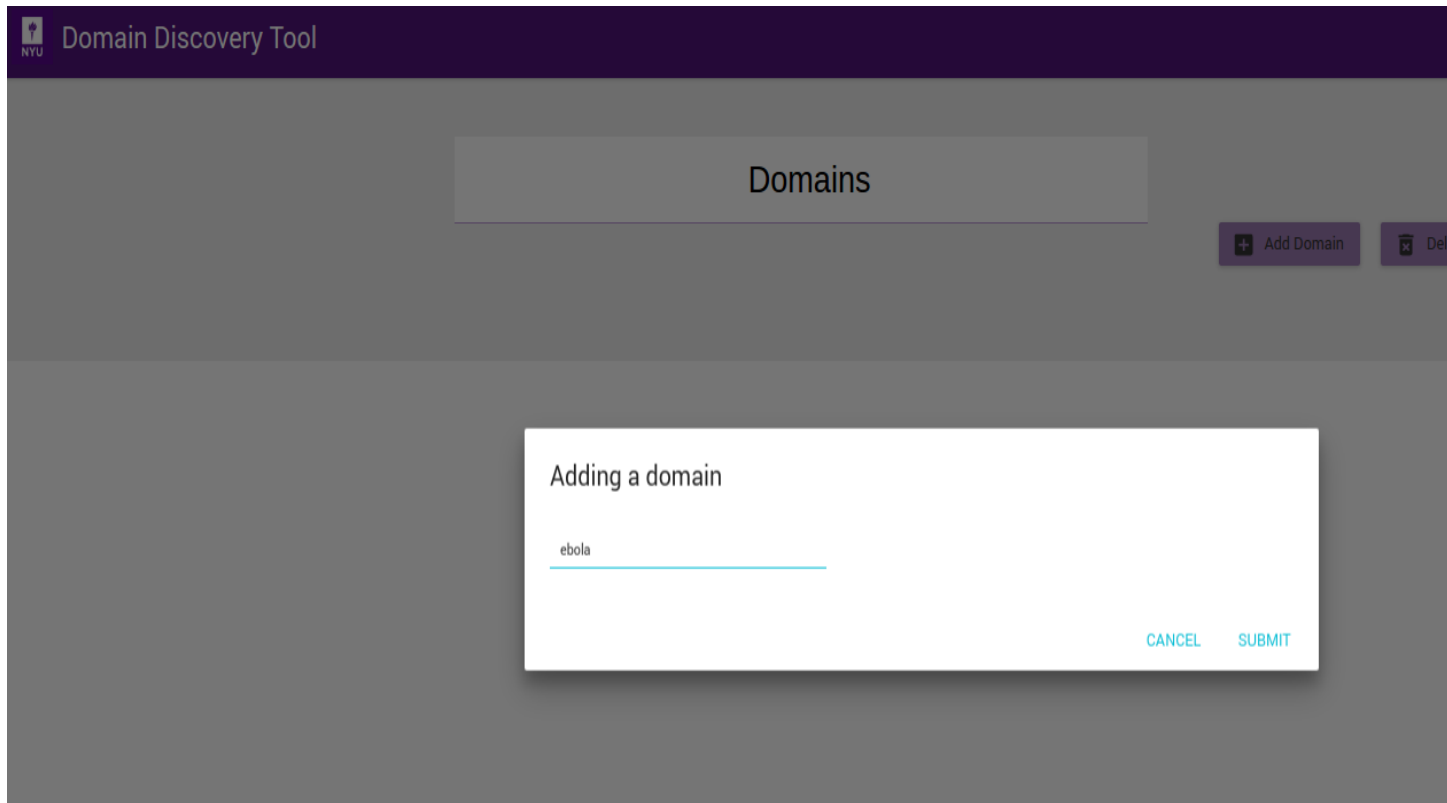
1.3 How To

Now you should be able to head to <http://<hostname>:8084/> to interact with the tool.

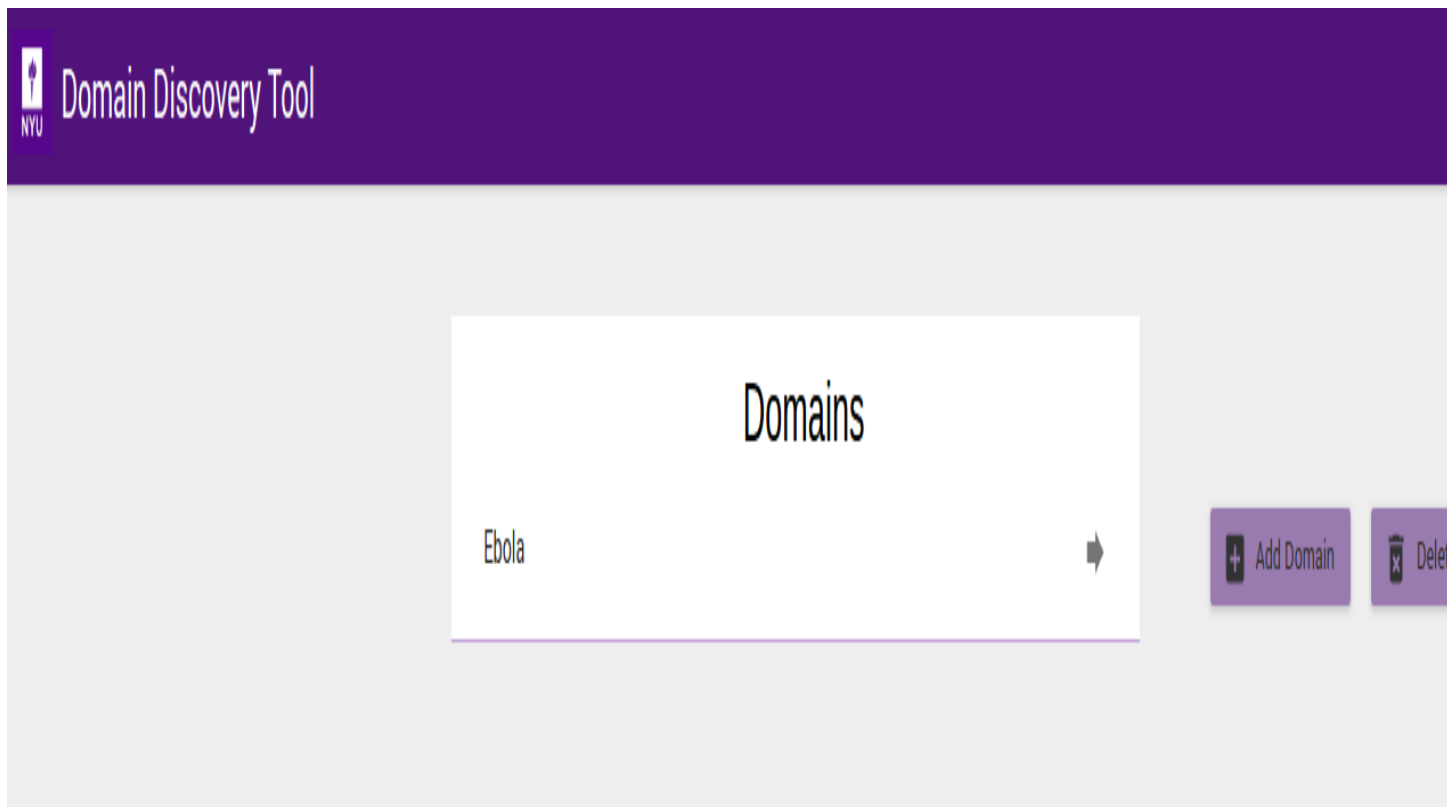
1.3.1 Create Domain



Begin by adding a domain on the Domains page (initial page), shown in the figure above, by clicking on the **Add Domain** button. Domain maintains context of domain discovery.



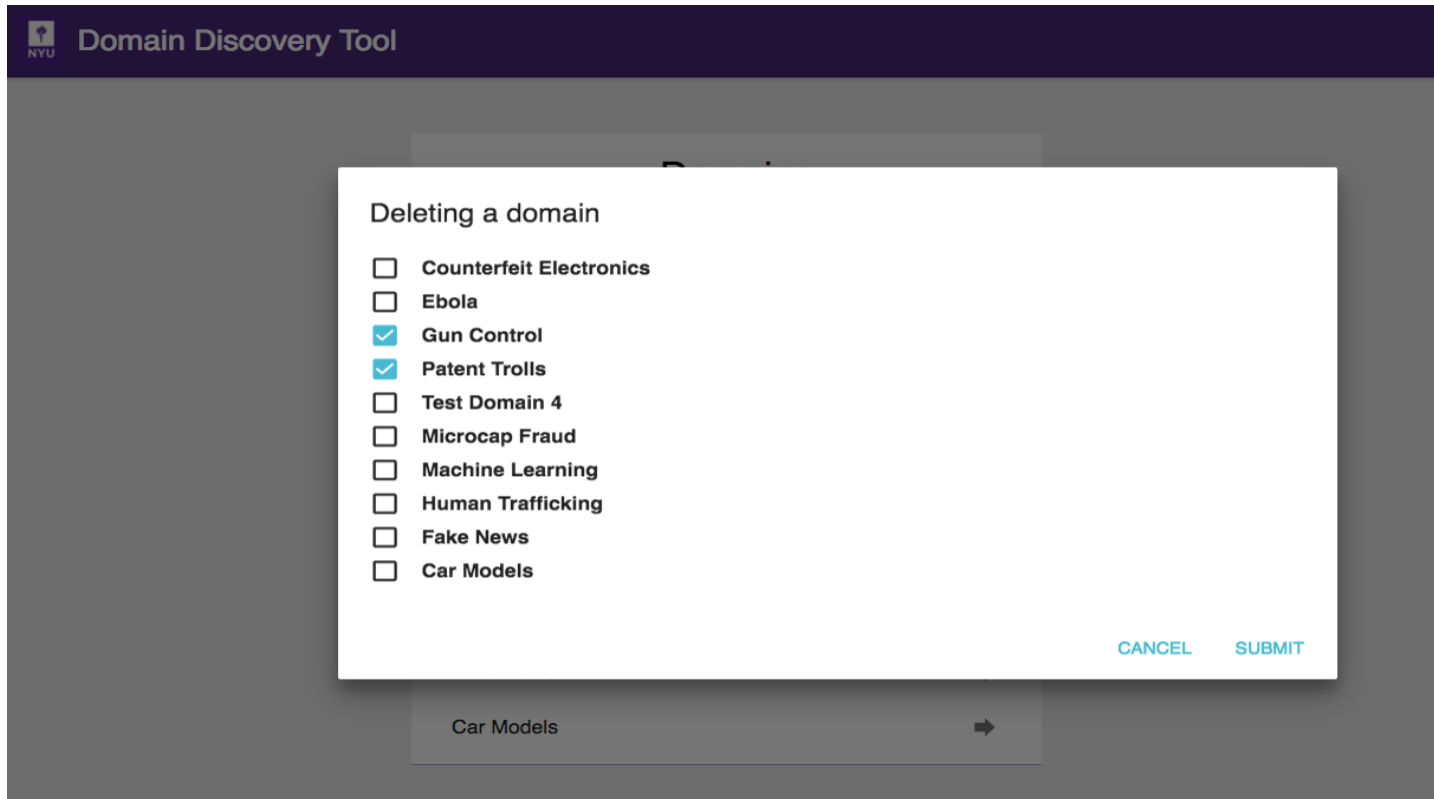
On the **Adding a domain** dialog shown in figure above, enter the name of the domain you would like to create, for example **Ebola**, and click on **Submit** button. You should now see the new domain you added in the list of domains as shown below.



Once domain is added click on domain name in the list of domains to collect, analyse and annotate web pages.

1.3.2 Delete Domain

Domains can be deleted by clicking on the **Delete Domain** button.



On the **Deleting a domain** dialog select the domains to be deleted in the list of current domains and click on **Submit** button. They will no longer appear on the domains list.

NOTE: This will delete all the data collected for that domain.

1.3.3 Acquire Data

Continuing with our example of the **Ebola** domain, we show here the methods of uploading data. Expand the Search tab on the left panel. You can add data to the domain in the following ways:

Upload URLs

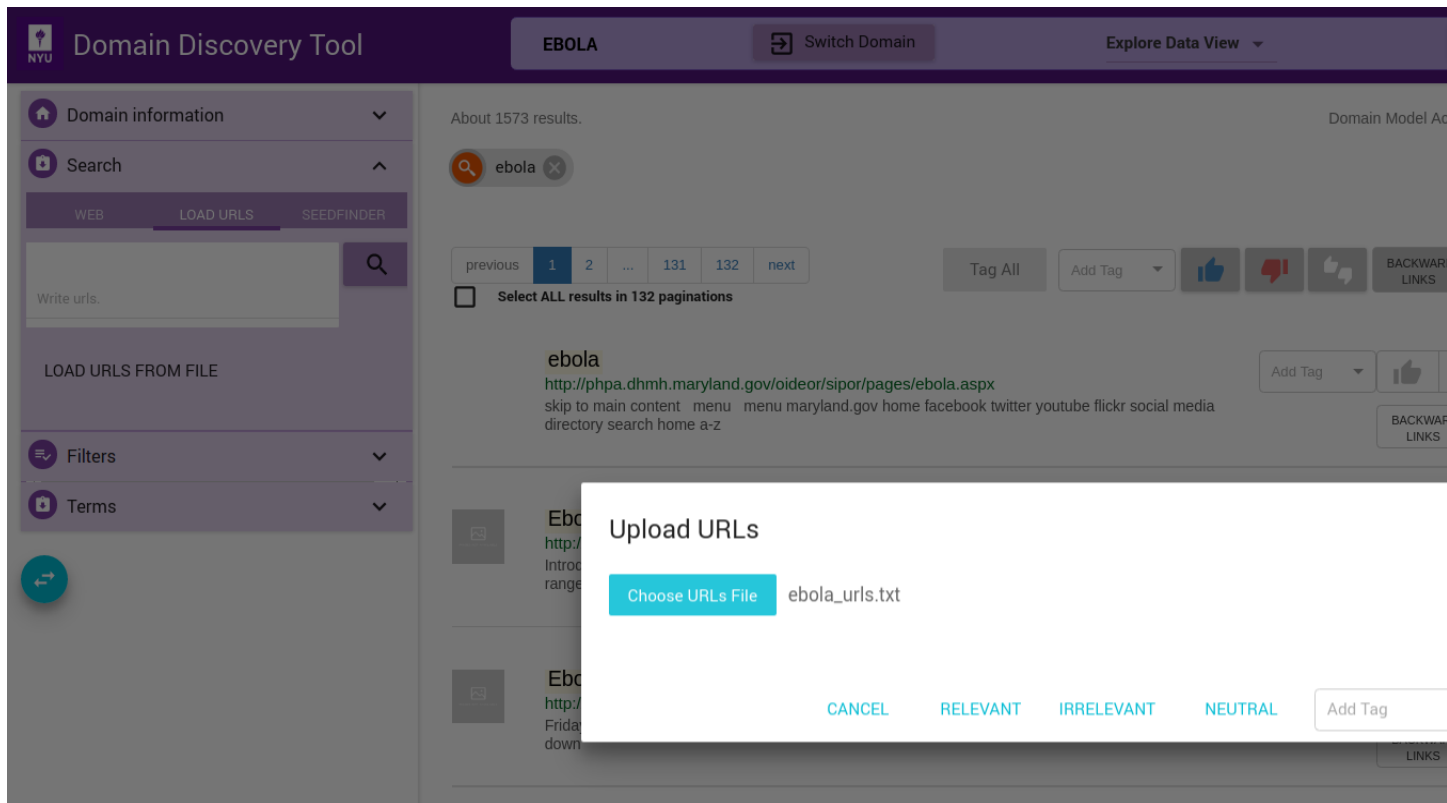
If you have a set of URLs of sites you already know, you can add them from the **LOAD** tab. You can upload the list of URLs in the text box, one fully qualified URL per line, as shown in figure below:

The screenshot shows the 'Domain Discovery Tool' interface. The left sidebar contains navigation options: 'Domain information', 'Search', 'WEB', 'LOAD URLS', 'SEEDFINDER', 'Filters', and 'Terms'. The 'LOAD URLS' section is active, showing a text input field with a URL and a 'LOAD URLS FROM FILE' button. The main content area displays search results for 'ebola'. It includes a search bar with 'ebola' entered, a pagination control showing 'previous', '1', '2', '...', '131', '132', and 'next', and a checkbox to 'Select ALL results in 132 paginations'. Two search results are visible: one from 'http://phpa.dhmmh.maryland.gov/oideor/sipor/pages/ebola.aspx' and another from 'http://www.cnn.com/ebola/'. Each result has a 'Deep Crawl' button and social media sharing options.

You can also upload a file with the list of URLs by clicking on the **LOAD URLS FROM FILE** button. This will bring up a file explorer window where you can select the file to upload. *The list of fully qualified URLs should be entered one per line in the file.* For example:

```
http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1003065
https://bmcpsy psychiatry.biomedcentral.com/articles/10.1186/s12888-017-1280-8
http://www.cdph.ca.gov/programs/cder/Pages/Ebola.aspx
```

Download an example URLs list file for ebola domain [HERE](#). Once the file is selected you can upload them by clicking on **RELEVANT**, **IRRELEVANT**, **NEUTRAL** or **Add Tag** (Add a custom tag). This will annotate the pages correspondingly.



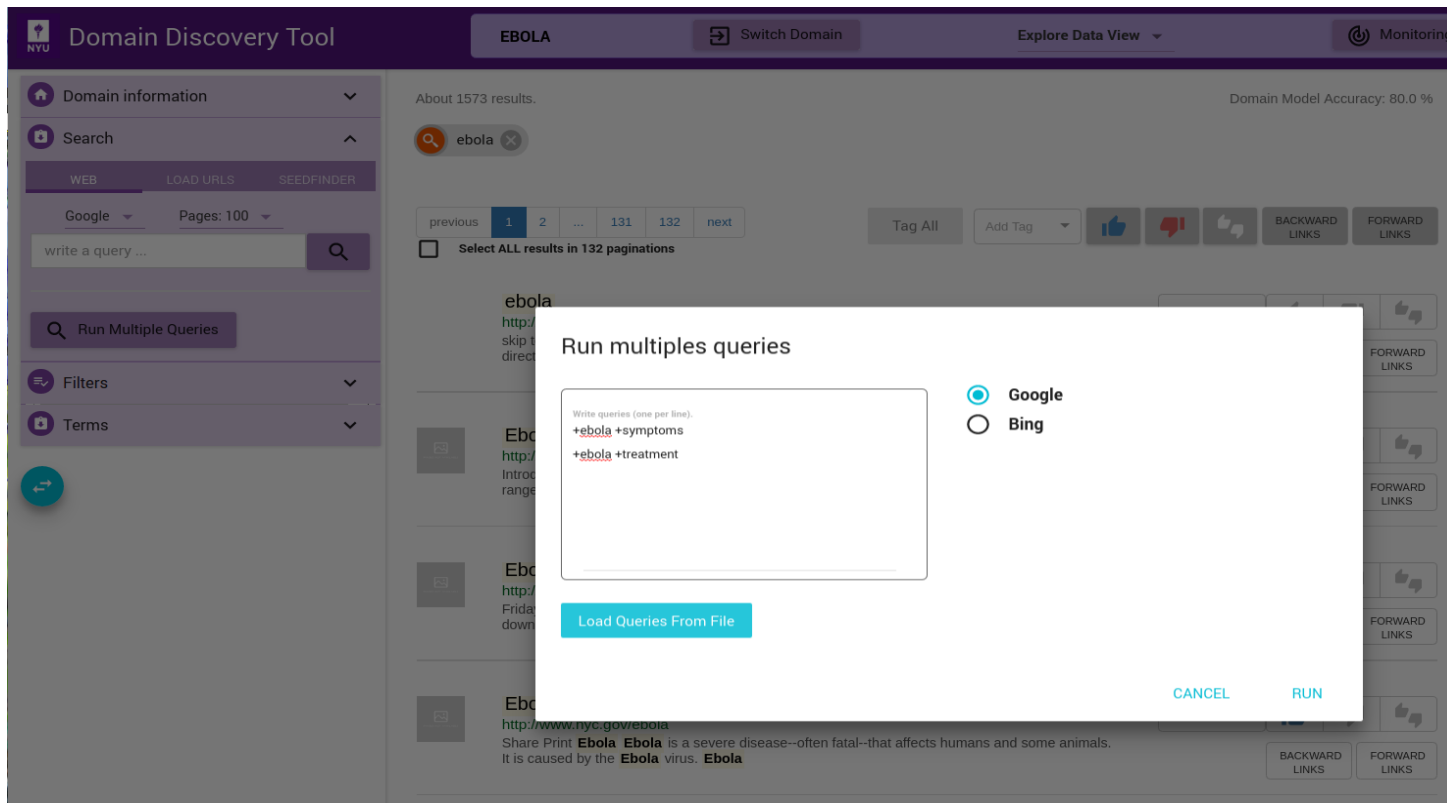
The uploaded URLs are listed in the **Filters** Tab under **Queries** as **Uploaded**.

Web Search

You can do a keywords search on google or bing by clicking on the **WEB** tab. For example, “ebola symptoms”. All queries made are listed in the **Filters** Tab under **Queries**.

The screenshot displays the Domain Discovery Tool (DDT) interface. The top navigation bar is purple and contains the NYU logo, the title 'Domain Discovery Tool', a search bar with 'EBOLA', a 'Switch Domain' button, an 'Explore Data View' dropdown, and a 'Monitor' button. The left sidebar is also purple and contains a 'Domain information' dropdown, a 'Search' dropdown, and a 'WEB' tab. Below the 'WEB' tab, there are options for 'Google' and 'Pages: 100', a search input field with 'ebola symptoms', a 'Run Multiple Queries' button, and 'Filters' and 'Terms' dropdowns. The main content area is white and shows search results for 'ebola symptoms'. It includes a pagination bar with 'previous', '1', '2', '...', '7', '8', and 'next'. Below the pagination bar, there are three search results, each with a title, a URL, a date, and a brief description. The first result is 'Signs and Symptoms | Ebola Hemorrhagic Fever | CDC' with a URL 'https://www.cdc.gov/vhf/ebola/symptoms/index.html' and a date 'Nov 2, 2014'. The second result is 'Ebola Virus: Symptoms, Treatment, and Prevention' with a URL 'https://www.webmd.com/a-to-z-guides/ebola-fever-virus-infection' and a date 'Oct 1, 2016'. The third result is 'Ebola virus and Marburg virus - Symptoms and causes - Mayo Clinic' with a URL 'https://www.mayoclinic.org/diseases-conditions/ebola-virus/symptoms-causes/dxc-20338674' and a date 'Jul 15, 2017'. Each result has an 'Add Tag' button and a 'BACKLINK' button.

If you have a multiple search queries then you can load them by clicking on the **Run Multiple Queries** button. This will bring up a window where you can either add the queries one per line in a textbox or upload a file that contains the search queries one per line. You can select the search engine to use (**Google** or **Bing**):

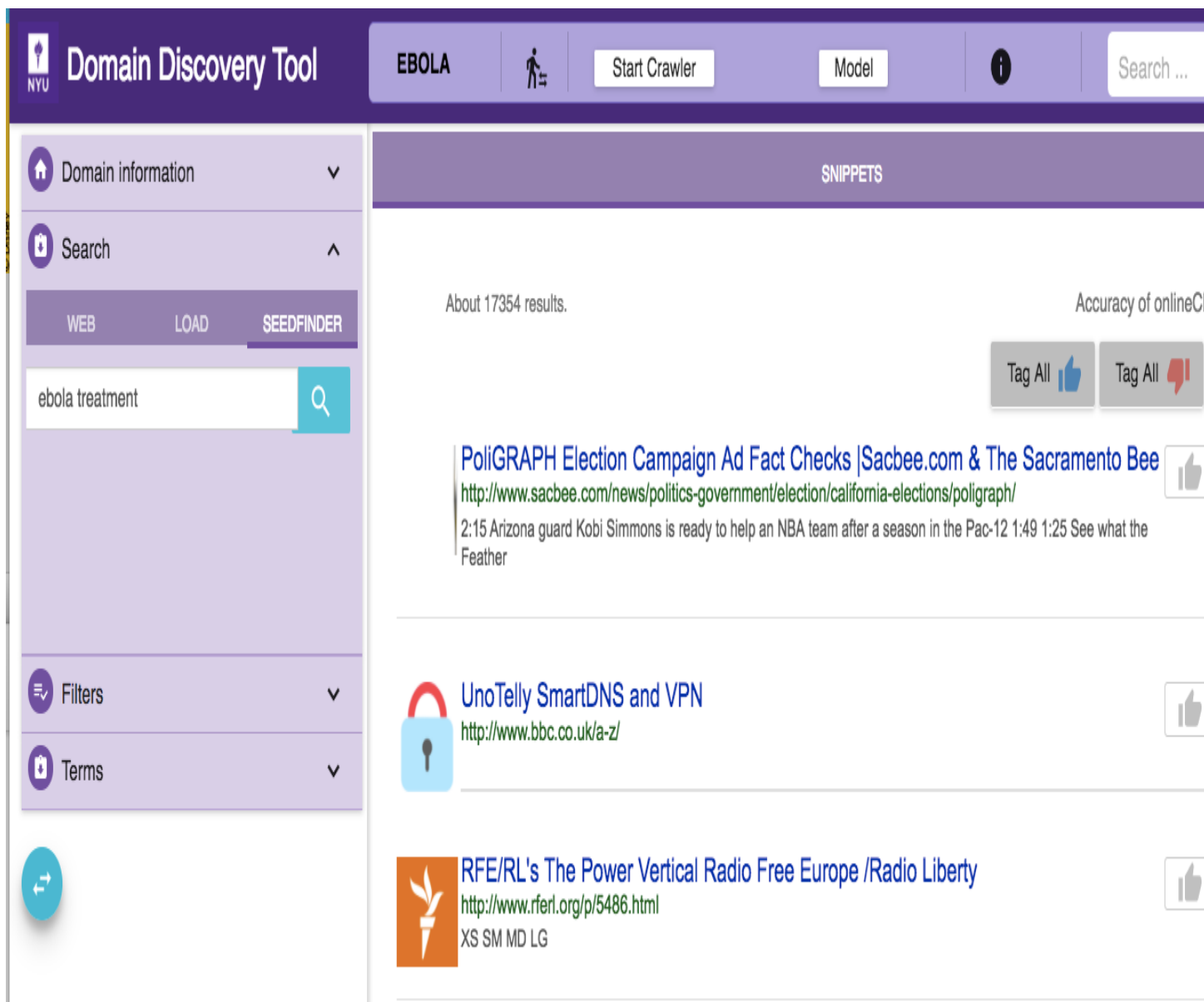


Each of the queries will be issued on Google or Bing (as chosen) and the results made available for exploration and annotation in the **Filters** Tab under **Queries** as **Uploaded**.

SeedFinder

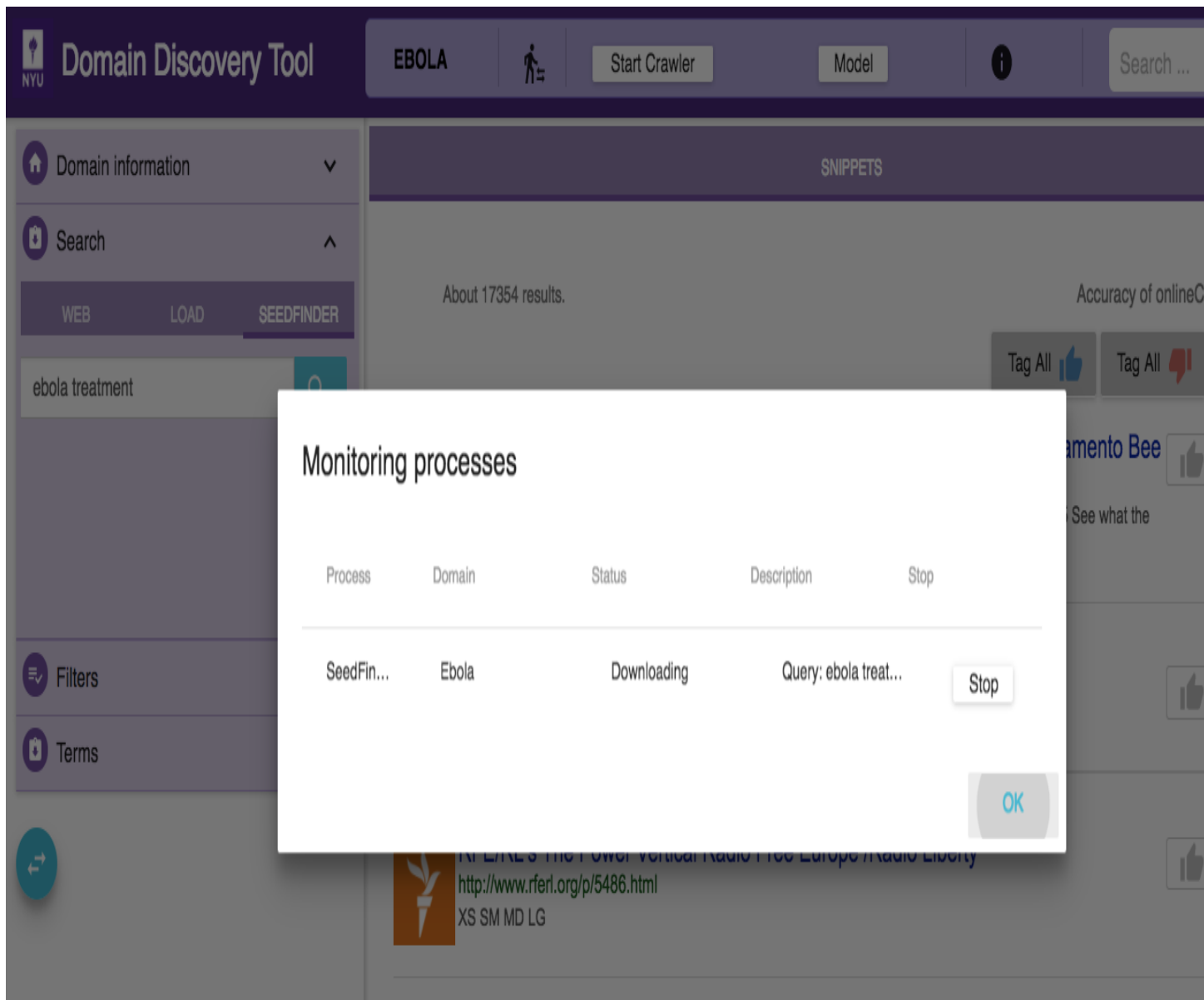
Instead of making multiple queries to Google/Bing yourself you can trigger automated keyword search on Google/Bing and collect more web pages for the domain using the SeedFinder. This requires a domain model. So once you have annotated sufficient pages, indicated by a non-zero accuracy on the top right corner, you can use the SeedFinder functionality.

To start a SeedFinder search click on the SEEDFINDER tab.



Enter the initial search query keywords, for example **ebola treatment**, as shown in the figure above. The SeedFinder issues this query to Google/Bing. It applies the domain model to the pages returned by Google/Bing. From the pages labeled relevant by the domain model the SeedFinder extracts keywords to form new queries which it again issues to Google/Bing. This iterative process terminates when no more relevant pages are retrieved or the max number of queries configured is exceeded.

You can monitor the status of the SeedFinder in the **Process Monitor** that can be accessed by clicking on the  on the top as shown below:



You can also stop the seedfinder process from the **Process Monitor** by clicking on the stop button shown along the corresponding process.

All queries made are listed in the **Filters** Tab under **SeedFinder Queries**. These pages can now be analysed and annotated just like the other web pages.

Crawl Forward and Backward

This allows the user to crawl one level forward or backward for all the selected URLs.

Forward Links - Forward links are all the links contained in a given page. When you crawl one level forward it downloads all the pages corresponding to the links contained in the page.

Backward Links - Backward links are all the links that contain a link to the given page. When you crawl one level backward it first finds all the links that contain a link to the selected page and then downloads all the pages corresponding to the links contained in the all the backward link pages.

The motivation for backward and forward crawling is the assumption that links containing the selected pages (back links) and links contained in the selected page (forward links) would be about similar topic as the selected page.

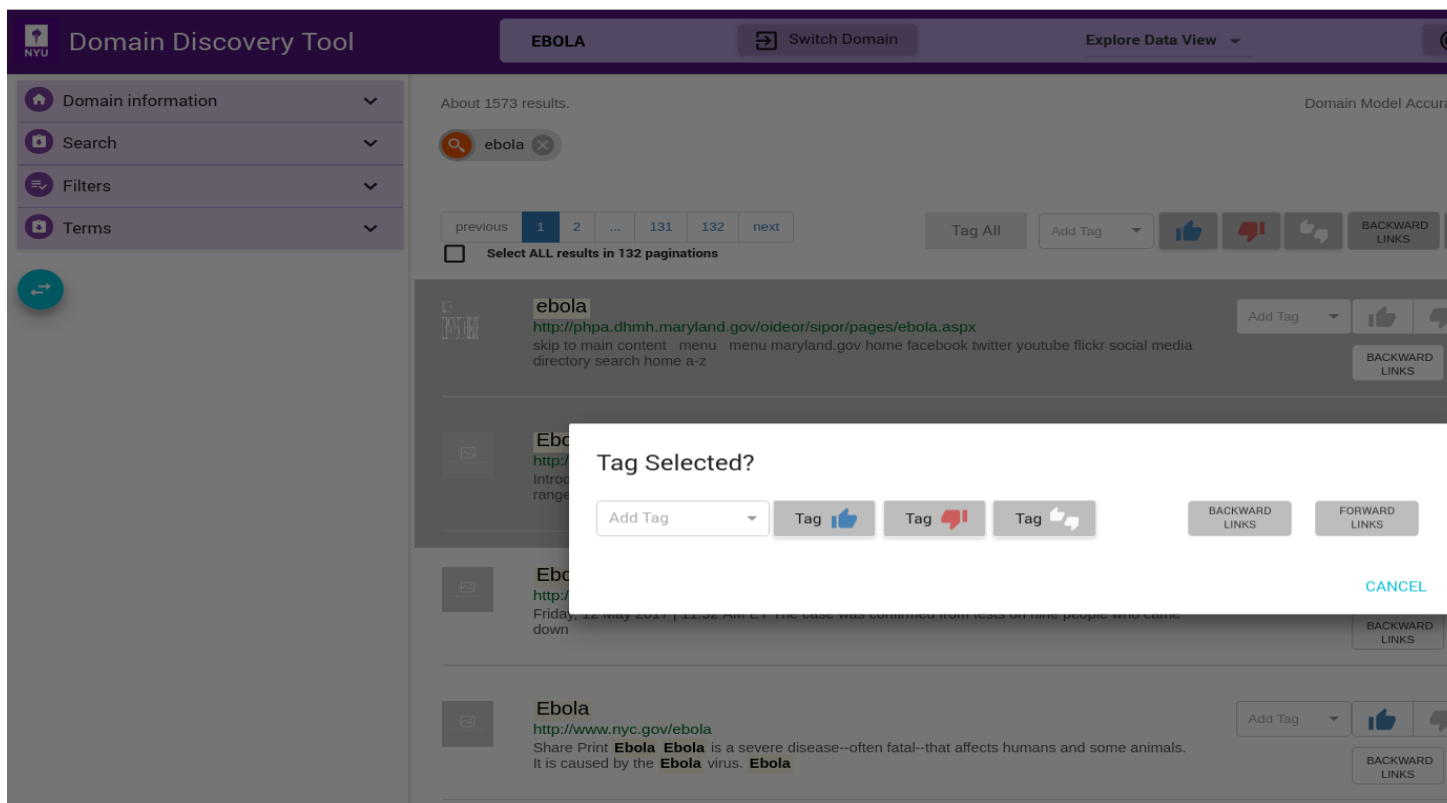
Crawl Individual Pages



buttons, along each page, can be used to crawl backward or forward links in individual pages.

Crawl Selected Pages

Select multiple pages by keeping the **ctrl** key pressed and clicking on the pages that you want to select. When done with selecting pages, release the **ctrl** key. This will bring up a window where you can choose to crawl forward or backward the pages as shown below:



Crawl All Pages



Use the buttons at the top of the list of pages to crawl backward or forward links on all pages in the current view.

Crawl All Pages for Current Filter

If you want to crawl forward or backward all pages retrieved for a particular filter (across pagination), then check the **Select ALL results in <total pages> paginations** checkbox below the page list on top left. Then use



buttons to crawl all the pages.

1.3.4 Explore Data (Filters)

The screenshot displays the 'Domain Discovery Tool' interface for the domain 'EBOLA'. The left sidebar contains a 'Filters' section with the following options:

- Domain information
- Search
- Filters
 - Queries
 - Crawled Data (CD)
 - Tags
 - news (12)
 - Neutral (403)
 - Relevant (47)
 - Irrelevant (22)
 - Deep Crawl (14)
 - Domains
 - Model Tags
- Terms

The main panel shows search results for 'ebola symptoms'. The top bar includes 'EBOLA', 'Switch Domain', 'Explore Data View', and 'Monitoring'. The search results are displayed in a list format, showing titles, URLs, and snippets. The first result is 'Ebola :Virus:Symptoms,Treatment,and PreventionSSSlideshow: Ebo...', followed by 'Ebola :Causes,Symptoms & TreatmentLive ScienceLive Science', 'Ebola :Symptoms,treatment,and causes', and 'Ebola Virus Causes,Symptoms,History & Vaccine'. Each result has an 'Add Tag' button and a set of action buttons (thumbs up, thumbs down, double thumbs up, backward links, forward links). The bottom of the panel shows a pagination bar with 'previous', '1', and 'next' buttons, and a checkbox for 'Select ALL results in 1 paginations'.

Once some pages are loaded into the domain, they can be analyzed and spliced with various filters available in the Filters tab on the left panel. The available filters are:

Queries

This lists all the web search queries and uploaded URLs made to date in the domain. You can select one or more of these queries to get pages for those specific queries.

Tags

This lists the annotations made to data. Currently the annotations can be either **Relevant**, **Irrelevant** or **Neutral**.

Domains

This lists all the top level domains of all the pages in the domain. For example, the top level domain for URL <https://ebolaresponse.un.org/data> is **ebolaresponse.un.org**.

Model Tags

You can expand the **Model Tags** and click the **Udate Model Tags** button that appears below, to apply the domain model to a random selection of 500 unlabeled pages. The predicted labels for these 500 pages could be:

- **Maybe Relevant:** These are pages that have been labeled relevant by the model with a high confidence
- **Maybe Irrelevant:** These are pages that have been labeled irrelevant by the model with a high confidence
- **Unsure:** These are pages that were marked relevant or irrelevant by the domain model but with low confidence. Experiments have shown that labeling these pages helps improve the domain model's ability to predict labels for similar pages with higher confidence.

NOTE: This will take a few seconds to apply the model and show the results.

Annotated Terms

This lists all the terms that are either added, uploaded in the Terms Tab. It also lists the terms from the extracted terms in the Terms Tab that are annotated.

SeedFinder Queries

This lists all the seedfinder queries made to date in the domain. You can select one or more of these queries to get pages for those specific queries.

Crawled Data

This lists the relevant and irrelevant crawled data. The relevant crawled data, **CD Relevant**, are those crawled pages that are labeled relevant by the domain model. The irrelevant crawled data, **CD Irrelevant**, are those crawled pages that are labeled irrelevant by the domain model.

Search for Keywords

The screenshot displays the Domain Discovery Tool (DDT) interface. On the left is a sidebar with navigation options: Domain information, Search, Filters, and Terms. The Search section is active, showing a search bar with 'ebola symptoms' and a 'Run Multiple Queries' button. The main area shows search results for the domain 'EBOLA'. At the top, it says 'About 86 results.' and 'Domain info'. Below this is a search bar with 'ebola symptoms' and a 'Switch Domain' button. A pagination bar shows 'previous', '1', '2', '...', '7', '8', and 'next'. There are buttons for 'Tag All', 'Add Tag', and social media icons. The results list includes three items:

- Signs and Symptoms | Ebola Hemorrhagic Fever | CDC**
<https://www.cdc.gov/vhf/ebola/symptoms/index.html>
 Nov 2, 2014 ... **Symptoms** of **Ebola** include. Fever; Severe headache; Muscle pain; Weakness; Fatigue; Diarrhea; Vomiting; Abdominal (stomach) pain...
- Ebola Virus: Symptoms, Treatment, and Prevention**
<https://www.webmd.com/a-to-z-guides/ebola-fever-virus-infection>
 Oct 1, 2016 ... WebMD explains the latest info on the rare but deadly disease **Ebola**, including how it's spread.
- Ebola virus and Marburg virus - Symptoms and causes - Mayo Clinic**
<https://www.mayoclinic.org/diseases-conditions/ebola-virus/symptoms-causes/dxc-20338674>
 Jul 15, 2017 ... Signs and **symptoms** typically begin abruptly within five to 10 days of infection with **Ebola** or

Search by keywords within the page content text. This search is available on the top right corner as shown in the figure above. It can be used along with the other filters. The keywords are searched not only in the content of the page but also the title and URL of the page.

1.3.5 Annotate Pages

A model is created by annotating pages as **Relevant** or **Irrelevant** for the domain. Currently, the model can only distinguish between relevant and irrelevant pages. You can also annotate pages with custom tags. These can be later grouped as relevant or irrelevant when generating the model. Try to alternate between Steps 3a and 3b to build a model till you reach at least 100 pages for each. This will continuously build a model and you can see the accuracy of the model at the top right corner - **Domain Model Accuracy**.

In the **Explore Data View** you see the pages for the domain (based on any filters applied) in two ways: through **Snippets** and **Visualizations**, as shown below:

The screenshot shows the Domain Discovery Tool (DDT) interface. The top navigation bar includes the NYU logo, 'Domain Discovery Tool', and tabs for 'EBOLA', 'Switch Domain', 'Explore Data View', and 'Monitoring'. A search bar is on the right. The left sidebar contains navigation links: 'Domain information', 'Search', 'Filters', and 'Terms'. The main content area is titled 'SNIPPETS' and shows 'About 76 results.' for the search term 'ebola symptoms'. Below this is a pagination bar with links for 'previous', '1', '2', '...', '6', '7', and 'next'. A checkbox labeled 'Select ALL results in 7 paginations' is present. The results list includes four entries, each with a snippet, a URL, and an 'Add Tag' button. The entries are:

- Signs and Symptoms | Ebola Hemorrhagic Fever | CDC** (https://www.cdc.gov/vhf/ebola/symptoms/index.html)
- Ebola Virus: Symptoms, Treatment, and Prevention** (https://www.webmd.com/a-to-z-guides/ebola-fever-virus-infection)
- Ebola : Causes, Symptoms & Treatment** (https://www.livescience.com/48311-ebola-causes-symptoms-treatment.html)
- Ebola virus and Marburg virus - Symptoms and causes - Mayo Clinic** (https://www.mayoclinic.org/diseases-conditions/ebola-virus/symptoms-causes/dxc-20338674)

 Each entry also has a 'BACKW LINKS' button.

The different mechanisms for annotating pages through **Snippet** are:

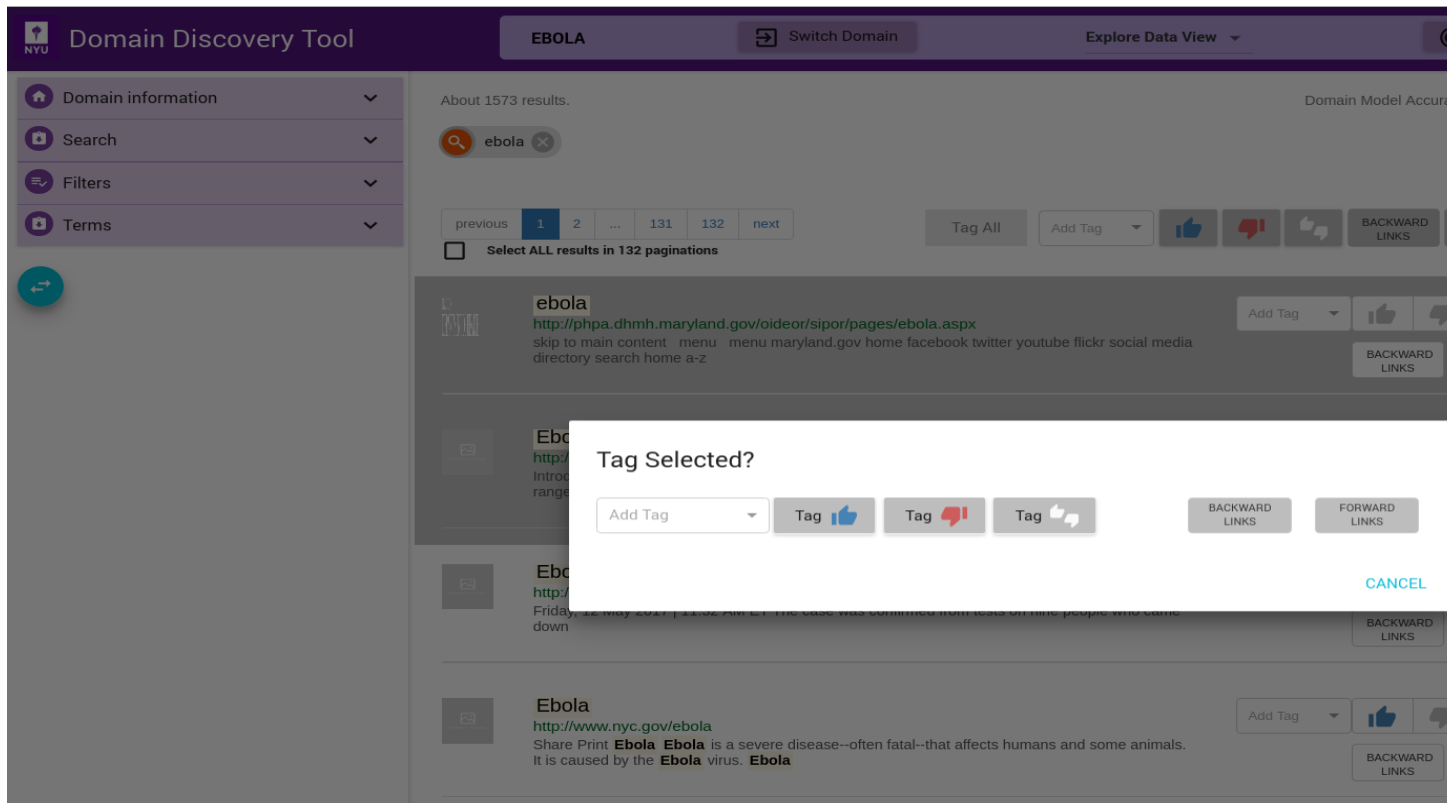
Tag Individual Pages



buttons, along each page, can be used to tag individual pages.

Tag Selected Pages

Select multiple pages by keeping the **ctrl** key pressed and clicking on the pages that you want to select. When done with selecting pages, release the **ctrl** key. This will bring up a window where you can tag the pages as shown below:



Tag All Pages in View



Use the **Tag All** buttons at the top of the list of pages to tag all pages in the current view

Tag All Pages for Current Filter

If you want to tag all pages retrieved for a particular filter (across pagination), then check the **Select ALL results in <total pages> paginations** checkbox below the page list on top left. Then use



buttons to tag all the pages.


Custom Tag

Custom tags can be added using Add Tag text box as shown below. Enter the custom tag in the Add Tag text box and press **enter** key. This adds the tag as a chip below the page info. This can be applied to individual, selected or all pages similar to relevant and irrelevant tags.

The screenshot shows the Domain Discovery Tool (DDT) interface. The top header is purple with the NYU logo and the text 'Domain Discovery Tool'. To the right of the header, there is a search bar containing 'EBOLA', a 'Switch Domain' button, and an 'Explore Data View' dropdown menu. On the left side, there is a sidebar with four menu items: 'Domain information', 'Search', 'Filters', and 'Terms', each with a downward arrow. Below the sidebar is a circular icon with a double arrow. The main content area displays 'About 5233 results.' and 'Domain Model'. Below this, there is a pagination bar with 'previous', '1', '2', '...', '436', '437', and 'next'. To the right of the pagination bar are buttons for 'Tag All', 'Add Tag', and social media icons (Facebook, Twitter, LinkedIn). Below the pagination bar is a checkbox labeled 'Select ALL results in 437 paginations'. The main content area also displays a search result for 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com' with a thumbnail image of a green frog. Below the search result is a 'news article' tag with a close button. On the right side, there is a 'symptoms' dropdown menu with a 'Create option "symptoms"' button and a 'BACKW LINK' button.

Tag for Deep Crawl

Some tags such as **Deep Crawl** are pre-configured. User can tag a page (or group of pages) for deep crawl by choosing the tag from the Add Tag drop-down as shown. For example, if user wants to deep crawl all the uploaded pages then they can tag the pages **Deep Crawl**.




Domain Discovery Tool

Domain information

Search

Filters

Terms



EBOLA

Switch Domain

Explore Data View

Monitor

About 5233 results.

Domain Model

previous

1

2

...


436


437


next

Tag All

Add Tag








BACK LINK

☐ Select ALL results in 437 paginations




Ebola: People Can Get the Virus and Not Have Symptoms | Time.com

<http://time.com/4596928/some-people-who-get-ebola-dont-show-symptoms-study/>


Dec 12, 2016 ... A new report reveals people can get Ebola and not have symptoms.

news article

Add Tag



BACK LINK



Ebola symptoms and transmission. :: Washington State Department ...

<https://www.doh.wa.gov/YouandYourFamily/IllnessandDisease/Ebola>

What are the symptoms of Ebola? Symptoms include fever, headache, body aches, diarrhea, vomiting, stomach pain and sometimes abnormal bleeding.


Deep Crawl

Add Tag

news article

Deep Crawl

outlier



BACK LINK

1.3. How To

37

1.3.6 Extracted Terms Summary

The screenshot displays the Domain Discovery Tool (DDT) interface. The top navigation bar includes the NYU logo, the title 'Domain Discovery Tool', and buttons for 'EBOLA SYMPTOMS', 'Start Crawler', 'Model', and a search bar. The left sidebar contains a 'Terms' tab, which is active, showing a list of terms with corresponding frequency bars. The terms listed are: marburg, ebola symptom, influenza, flu, virus disease, disease, virus, fever, outbreaks, information, and infection. The right panel shows the search results for 'ebola AND symptoms', displaying two snippets from eMedTV and Everyday Health.

The most relevant terms and phrases (unigrams, bigrams and trigrams) are extracted from the pages in the current view of DDT and listed in the Terms Tab on the left panel, as shown in the figure above. This provides a summary of the pages currently in view. Initially, when there are no annotated terms, the top 40 terms with the highest TFIDF (term frequency-inverse document frequency) are selected. The terms are displayed with their frequency of occurrence in relevant (blue) and irrelevant (red) pages (bars to the right of the Terms panel). This helps the expert to select terms that are more discerning of relevant pages.

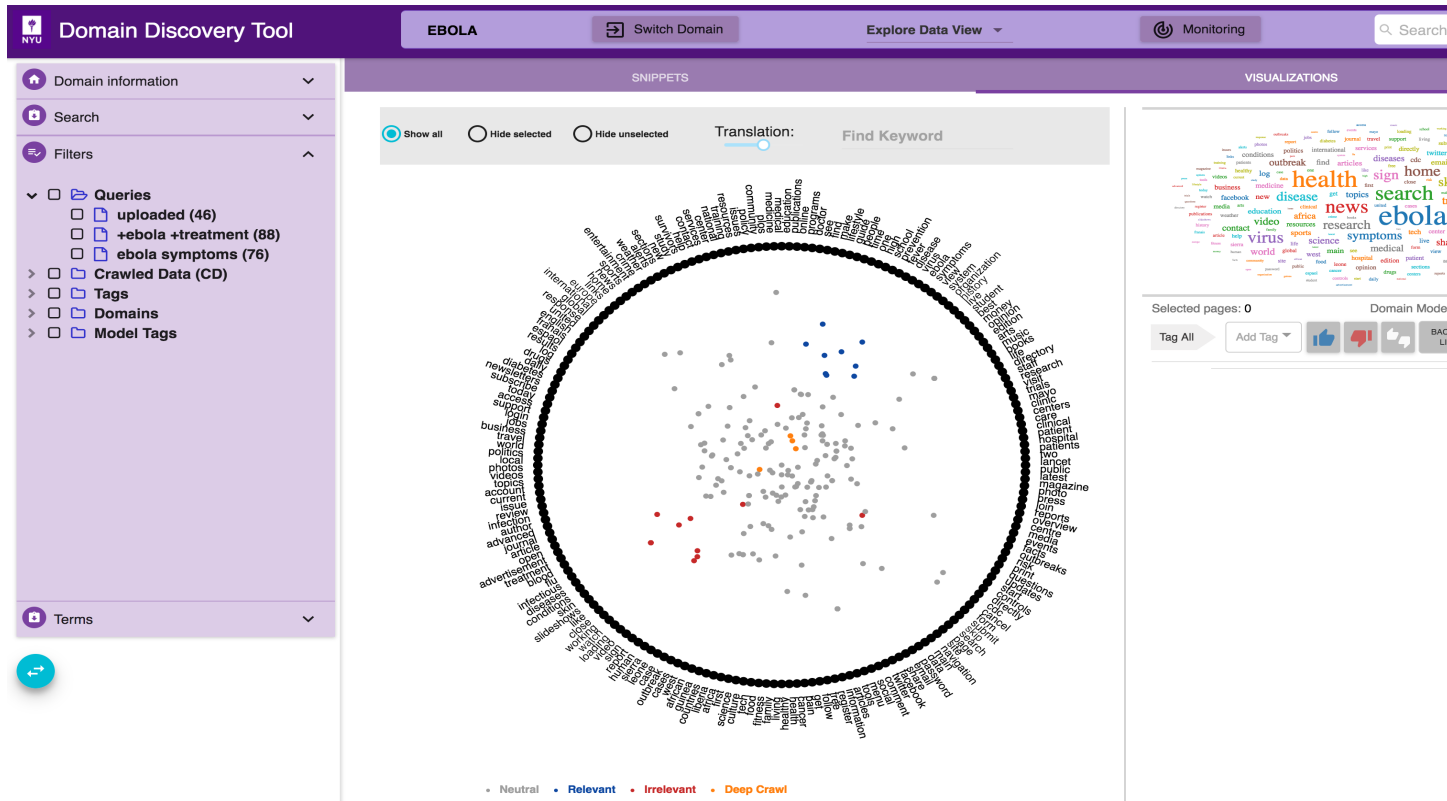
Terms can be tagged as 'Positive' and 'Negative' by 1-click and 2-click respectively. The tags are stored in the active data source. When the update terms button is clicked, the positively and negatively annotated terms are used to re-rank the other terms. Terms help the expert understand and discover new information about the domains of interest. The terms can be used to refine the Web search or start new sub topic searches.

Custom relevant and irrelevant terms can be added by clicking the + button to boost the extraction of more relevant terms. These custom terms are distinguished by the delete icon before them which can be clicked to delete the custom term.

Hovering the mouse over the terms in the Terms window displays the context in which they appear on the pages. This again helps the expert understand and disambiguate the relevant terms. Inspect the terms extracted in the "Terms" window. Clicking on the stop button pins the context to the corresponding term.

1.3.7 Visualization through RadViz

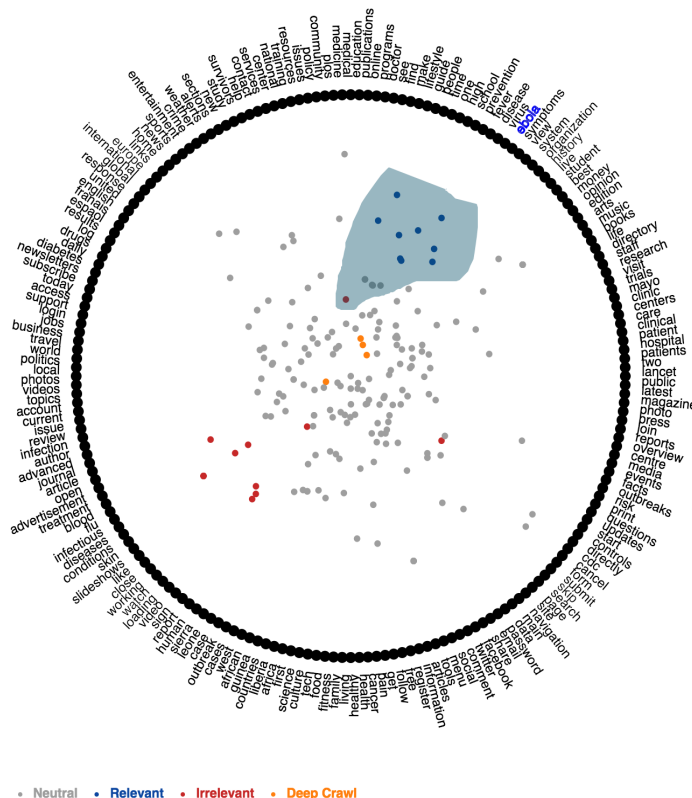
Select the **Visualization Tab** in the **Explore Data View** to see the multidimensional scaling visualization using RadViz.



RadViz is a data visualization that enables users to explore and analyze samples in a data set (such as a corpus of web pages in the case of DDT), represented as points in the visualization, in terms of similarity relations among semantic descriptors (keywords on the pages). Keywords are located along a circle, and pages are represented as points in the circle. The more similar the pages the closer the distance between them. Also, the greater the proximity of a page to a keyword, the greater the frequency of occurrence of that keyword in that page. This kind of analysis allows users to identify regions of interest in the data set according to the most relevant features of the sample.

Explore Pages

In order to explore the pages in the visualization you would need to select the pages that you want to see the various details for.



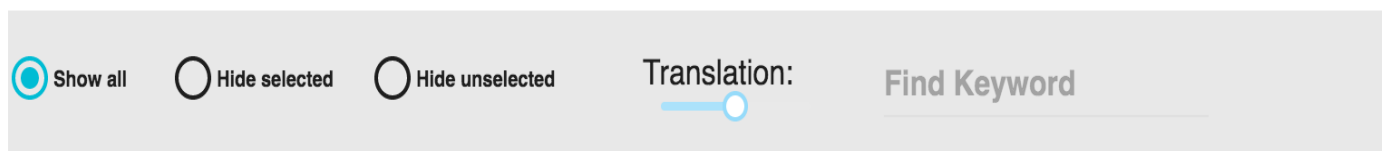
Selection of a group of pages is done using lasso selection. For this, the user simply drags a freehand selection around the pages located in the circle in RadViz, similar to how you would outline something on a piece of paper with a pen or pencil. To start the lasso selection users simply have to click at the spot where you want to begin the selection, then continue holding your mouse button down and drag to draw a freeform selection outline. To complete the selection, simply release your mouse button. You don't necessarily have to return the same spot you started from, but if you don't, RadViz will automatically close the selection for you by drawing a straight line from the point where you released your mouse button to the point where you began, so in most cases, you will want to finish where you started.

When the pages are selected, you will observe the following:

- Keywords contained in the selected pages will be highlighted along the circle.
- A WordCloud of all the top keywords contained in the selected pages is generated in the right top corner. The font size of the keyword in the word cloud is proportional to the frequency of occurrence of the word
- Snippets of selected pages are shown at the right bottom corner

Pages can be tagged through RadViz as 'Positive' and 'Negative', and even Custom Tag, by drawing lasso around any region of interest, which made the selection of a sub-group of pages very easy, and then users can tag the selected pages as 'Positive', 'Negative' and Custom Tag.

ToolBar RadViz



This visualization has five controls to interact with, whose functionality are described below.

Showing data in RadViz



radio buttons, can be used to show or hide data on RadViz.

Show all: Show all is selected by default in this visualization. It shows all the pages present in the data collection.

Hide selected: This option hides the selected pages of the current view.

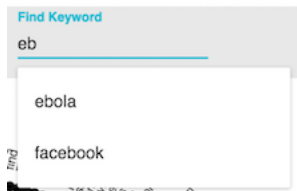
Hide unselected: This option hides the pages that are not selected.

Translation in RadViz



slider allows to calibrate the degree of denseness or sparseness of the representations of the pages in the visualization.

Find Keyword in RadViz



auto-complete text-field allows to search a keyword over all keywords in the visualization. Blue font color is used to highlight the keyword (shown below). This functionality is supported by an autocomplete process using all keyword used in th current view of RadViz.



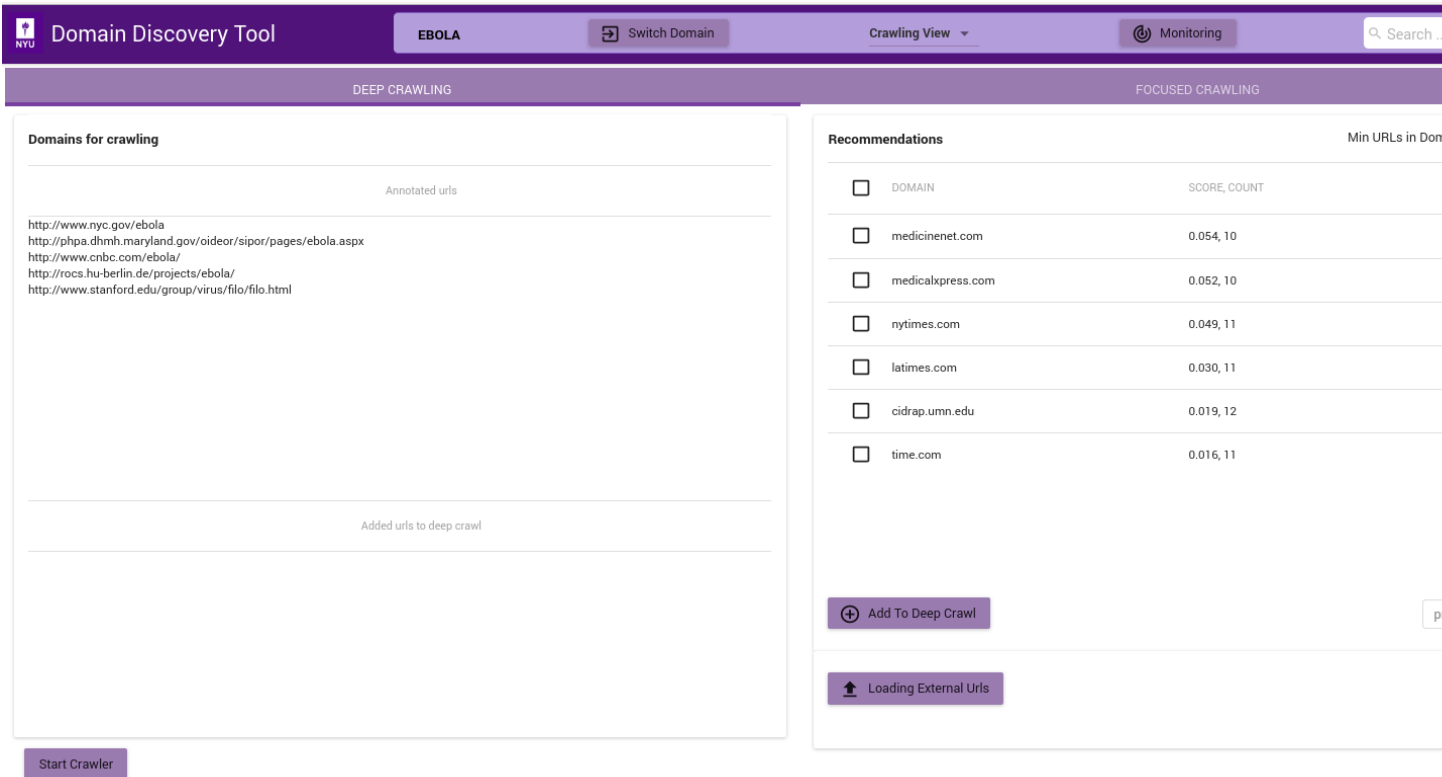
1.3.8 Run Crawler

Once a sufficiently good model is available or pages are tagged for a deep crawl you can change from **Explore Data View** to the **Crawler View** to start the crawl shown below:

The screenshot shows the Domain Discovery Tool (DDT) interface. The top navigation bar is purple and contains the NYU logo, the title "Domain Discovery Tool", a search bar with "EBOLA" entered, a "Switch Domain" button, and a "Monitor" button. A dropdown menu is open over the search bar, showing "Explore Data View" and "Crawling View". On the left, a sidebar menu lists "Domain information", "Search", "Filters", and "Terms". The main content area displays "About 5233 results." and a pagination bar with links for "previous", "1", "2", "...", "436", "437", and "next". Below the pagination bar is a checkbox labeled "Select ALL results in 437 paginations". The first search result is a news article titled "Ebola: People Can Get the Virus and Not Have Symptoms | Time.com" with a URL and a brief description. The second result is a page titled "Ebola symptoms and transmission. :: Washington State Department ..." with a URL and a brief description. Both results have "Add Tag" buttons and "BACKLINK" buttons. A "news article" tag is visible below the first result, and a "Deep Crawl" tag is visible below the second result.

Deep Crawl

In order to run a *Deep Crawl* annotate pages to be crawled with tag *Deep Crawl* as described in [Tag for Deep Crawl](#).



The figure above shows the Deep Crawl View. The list on the left shows all pages annotated as *Deep Crawl* in the Explore Data View. The table on the right shows recommendations of pages that could be added to deep crawl by clicking on the **Add to Deep Crawl**. If keyword terms are added or annotated then recommendations are made based on the score of how many of the keywords they contain. Otherwise the domains are recommended by the number of pages they contain.

The deep crawler can be started by clicking on **Start Crawler** button at the bottom. This starts a deep crawler with all the pages tagged for Deep Crawl.

You can see the results of the crawled data in **Crawled Data** in the Filters Tab. When the crawler is running it can be monitored by clicking on the **Crawler Monitor** button.

Focused Crawl

The figure below shows the Focused Crawler View:

1. In the ‘Model Settings’ on the left select the tags that should be considered as relevant(Positive) and irrelevant(Negative). If there sufficient relevant and irrelevant pages (about 100 each), then you can start the crawler by clicking on the **Start Crawler** button.
2. If there are no irrelevant pages then a page classifier model cannot be built. Instead you can either upload keywords by clicking on the ‘Add Terms’ in the Terms window. You can also annotate the terms extracted from the positive pages by clicking on them. If no annotated terms are available then the top 50 terms are used to build a regular expression model.
3. Once either a page classifier or a regex model is possible start the focused crawler by clicking on the **Start Crawler**.

You can see the results of the crawled data in “Crawled Data” in the Filters Tab. When the crawler is running it can be monitored by clicking on the ‘Crawler Monitor’ button.

The Model info on the bottom right shows how good a domain model is if there are both relevant and irrelevant pages annotated. The color bar shows the strength of the model based on the balance of relevant and irrelevant pages and the classifier accuracy of the model.

1.4 Publication

Yamuna Krishnamurthy, Kien Pham, Aecio Santos, and Juliana Freire. 2016. [Interactive Web Content Exploration for Domain Discovery](#) (Interactive Data Exploration and Analytics (IDEA) Workshop at Knowledge Discovery and Data Mining (KDD), San Francisco, CA).

1.5 Contact

DDT Development Team [ddt-dev@vgc.poly.edu]

CHAPTER 2

Links

- [GitHub repository](#)

CHAPTER 3

Indices and tables

- `genindex`
- `modindex`
- `search`