
ddt Documentation

Release 1.0

Yamuna Krishnamurthy

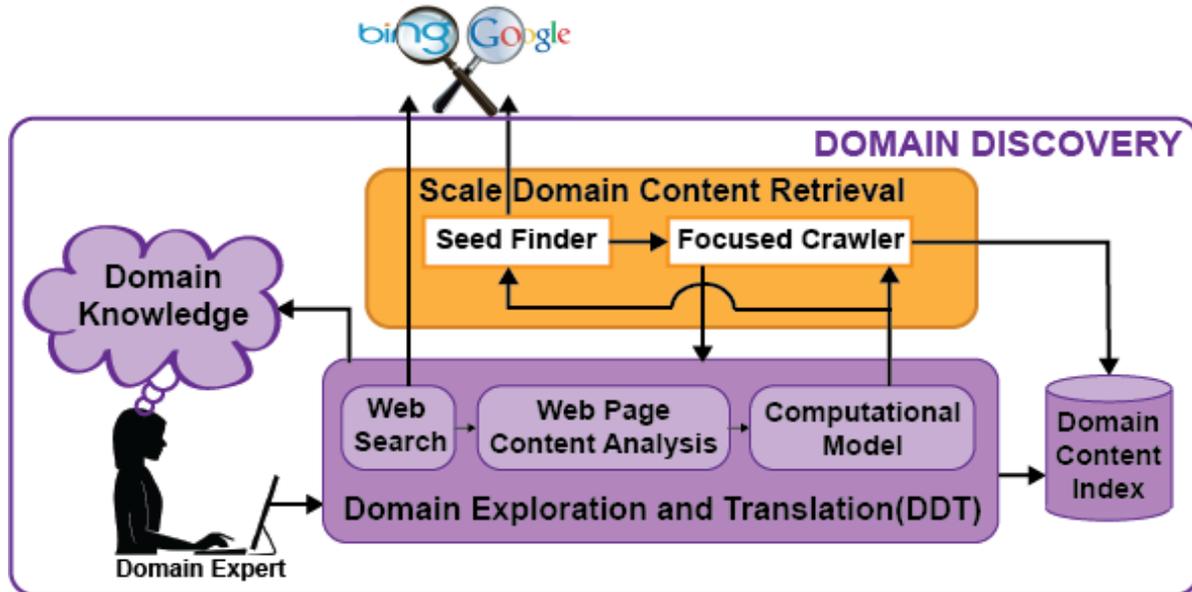
Dec 11, 2021

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
↳ b714e105ccb3a6d5a718c76c2ce1e5a51ea6f10a5f4997a6e5b12b9c7faf50e. 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
↳ 326881fda035692aa0a5c03ec808294aad2f9fd816baa13270d2fe50e7e1e77. 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

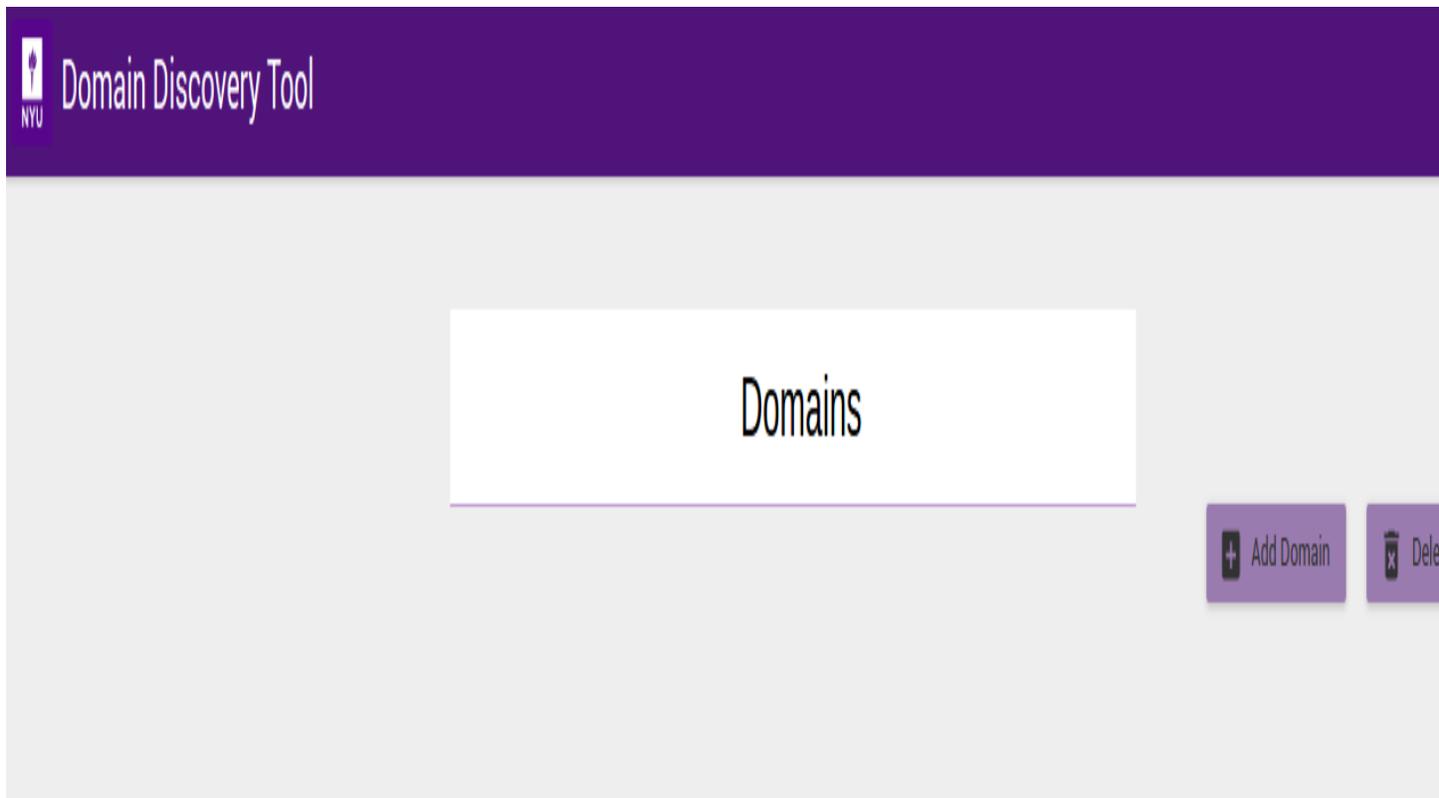
1.2.1 Building Domain Index

Creating a domain specific index involves:

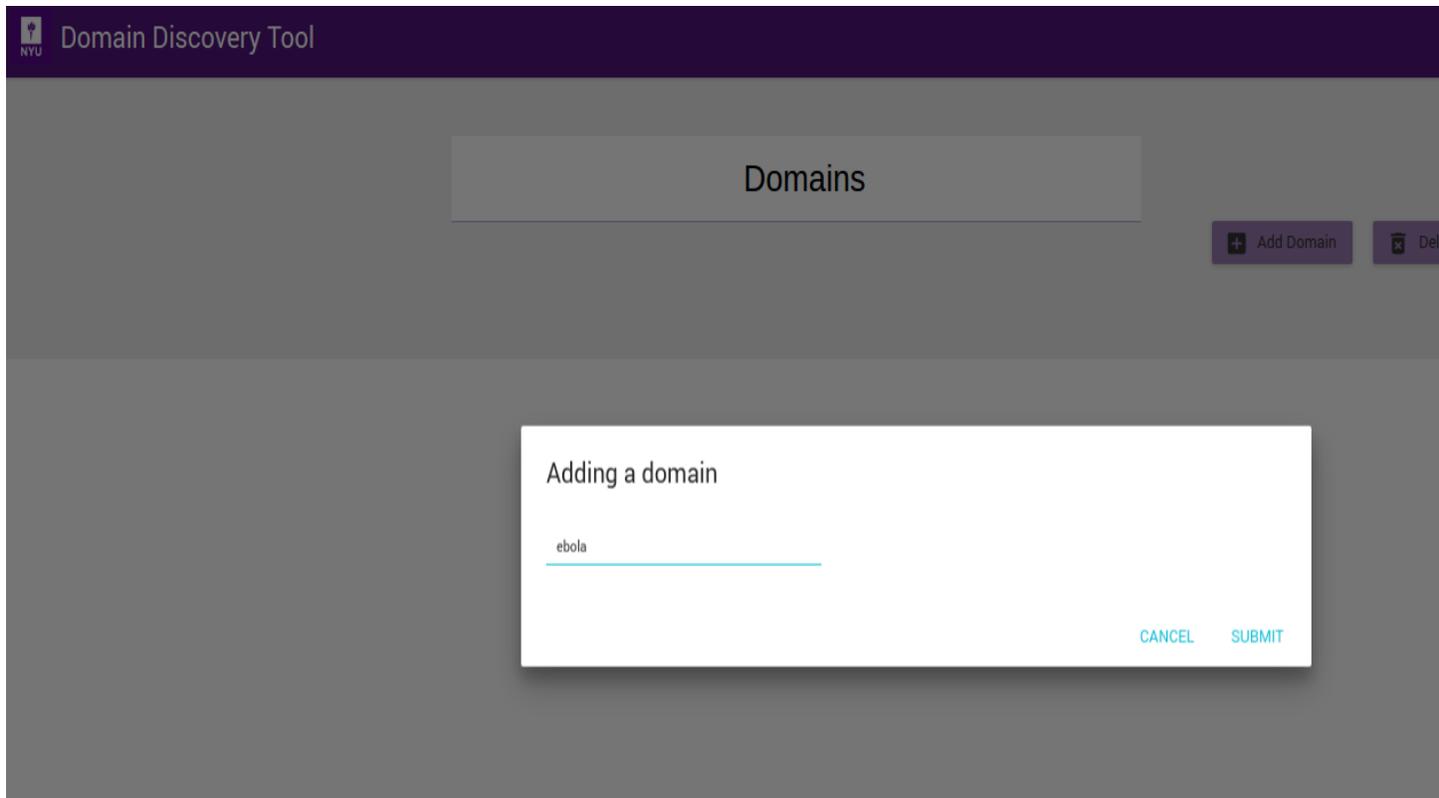
- Uploading known relevant URL domains from which you would like to collect all URLs belonging to that domain. This is called a deep crawl. Follow steps 1,2 and 4, below, for this.
- Create a domain model that can be used for a focused crawl (broad crawl). For this follow all the steps 1-4 below.

Step 1

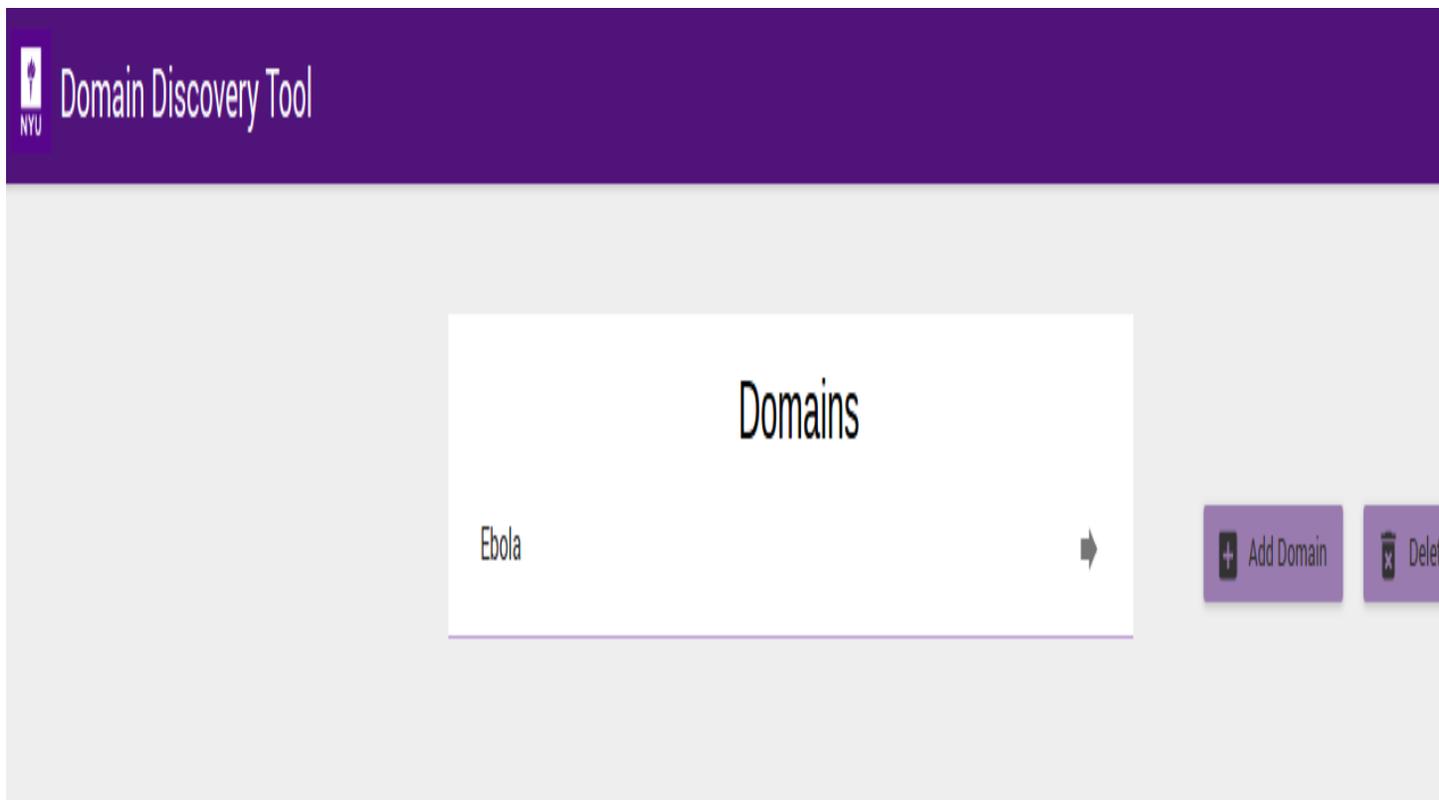
Create Domain



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

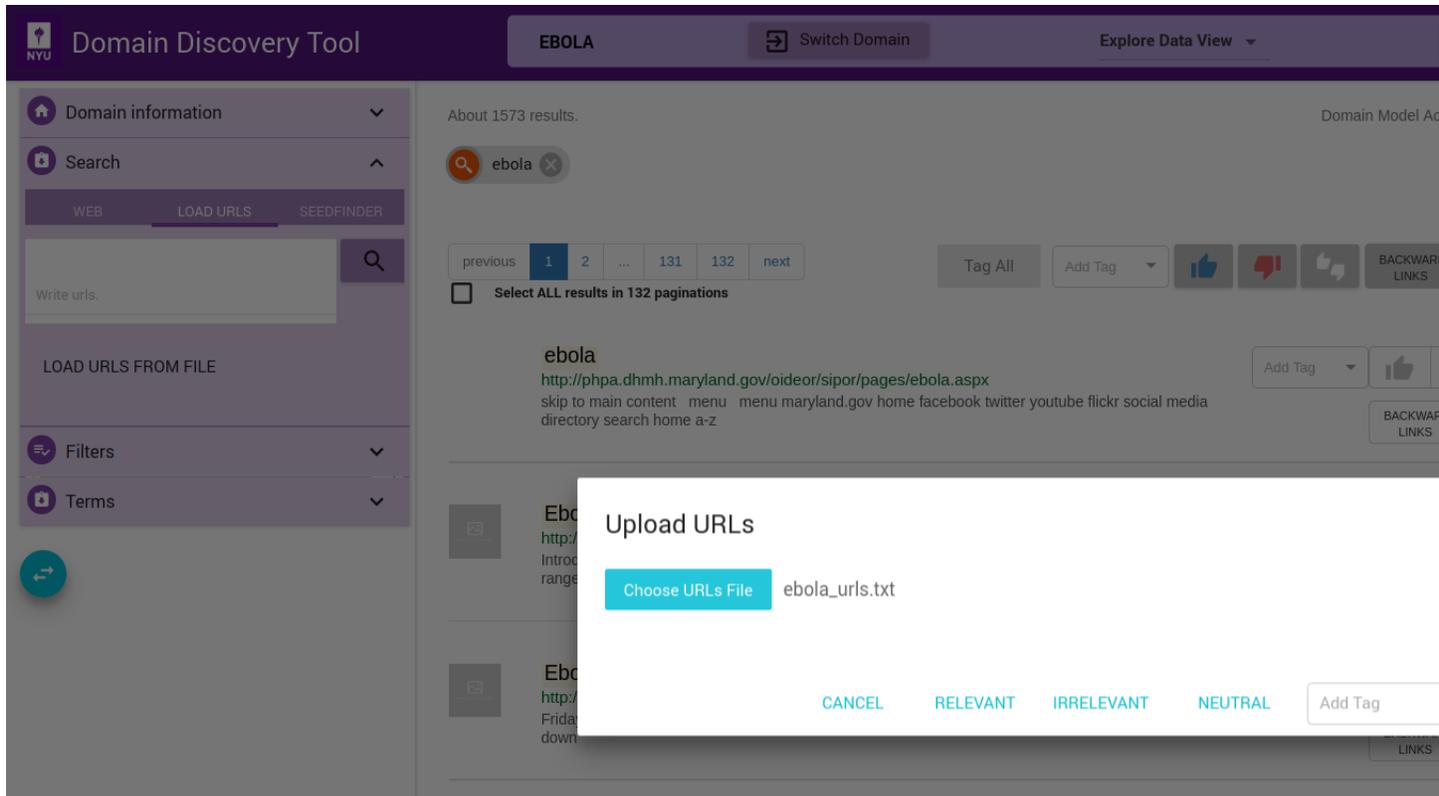
If you have a set of URLs of sites you already know, you can add them from the **LOAD** tab. You can upload the list of URLs in the text box, one fully qualified URL per line, as shown in figure below:

The screenshot shows the Domain Discovery Tool interface. The top navigation bar includes the NYU logo, the tool name 'Domain Discovery Tool', the current domain 'EBOLA', a 'Switch Domain' button, and an 'Explore Data View' dropdown. The sidebar on the left contains navigation options: 'Domain information', 'Search', 'WEB', 'LOAD URLS', 'SEEDFINDER', 'LOAD URLS FROM FILE', 'Filters', and 'Terms'. The main content area shows search results for 'ebola' with 'About 1575 results.' and a search bar containing 'ebola'. Below the search bar are pagination controls (previous, 1, 2, ..., 131, 132, next) and a checkbox for 'Select ALL results in 132 paginations'. Two search results are visible, each with a 'Deep Crawl' button and social media sharing options. The first result is for 'ebola' on 'http://phpa.dhmh.maryland.gov/oideor/sipor/pages/ebola.aspx' and the second is for 'Ebola' on 'http://www.cnn.com/ebola/'.

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://bmcpyschiatry.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**.

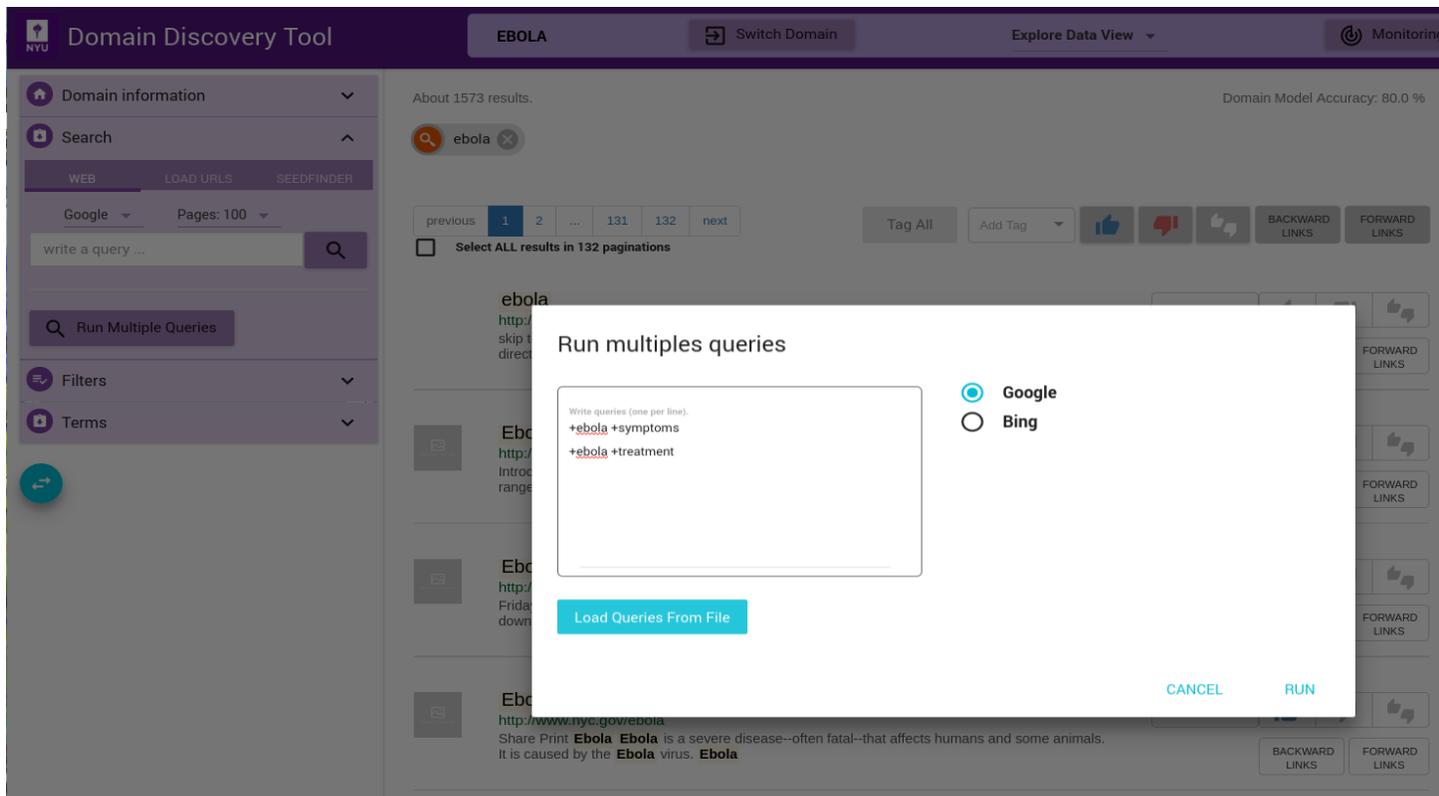
Web Search

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 interface. At the top, the domain 'EBOLA' is selected, with options to 'Switch Domain' and 'Explore Data View'. The left sidebar contains navigation options: 'Domain information', 'Search', 'Filters', and 'Terms'. Under 'Search', there are tabs for 'WEB', 'LOAD URLS', and 'SEEDFINDER'. A search box contains 'ebola symptoms' with a search icon. Below the search box is a 'Run Multiple Queries' button. The main content area shows 'About 86 results.' and a search query 'ebola symptoms'. There are pagination controls (previous, 1, 2, ..., 7, 8, next) and a 'Select ALL results in 8 paginations' checkbox. Three search results are visible, each with a title, URL, date, and a brief description. The results are:

- 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

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 screenshot shows the 'Domain Discovery Tool' interface. The top navigation bar includes 'EBOLA', 'Switch Domain', 'Explore Data View', and 'Monitoring'. The left sidebar contains 'Domain information', 'Search', 'Filters', and 'Terms'. The main content area is titled 'SNIPPETS' and shows search results for 'ebola symptoms'. The results list four items:

- 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 result includes a snippet of text and an 'Add Tag' button. The interface also features pagination controls (previous, 1, 2, ..., 6, 7, next) and a 'Select ALL results in 7 paginations' checkbox.

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

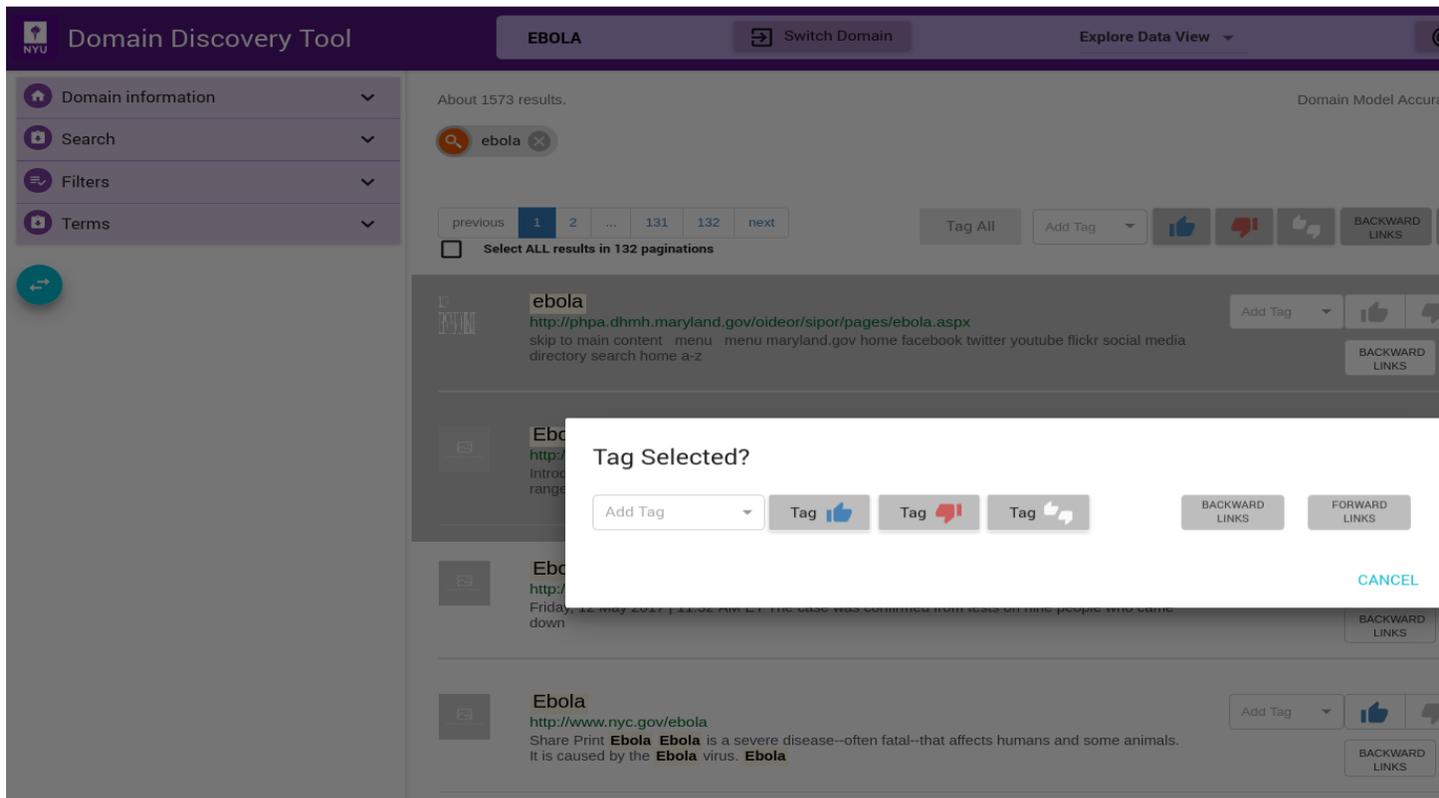
Tag Individual Pages



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 interface. The top navigation bar is purple and contains the NYU logo, the text 'Domain Discovery Tool', the search term 'EBOLA', a 'Switch Domain' button, and an 'Explore Data View' dropdown menu. On the left side, there is a vertical sidebar with four menu items: 'Domain information', 'Search', 'Filters', and 'Terms', each with a downward arrow. Below the sidebar is a circular blue button with a white double-headed arrow. The main content area shows 'About 5233 results.' and 'Domain Model'. Below this is a pagination control with buttons for 'previous', '1', '2', '...', '436', '437', and 'next'. To the right of the pagination are buttons for 'Tag All', 'Add Tag' (with a dropdown arrow), and social media icons (Facebook, Twitter, LinkedIn). Below the pagination is a checkbox labeled 'Select ALL results in 437 paginations'. The main search result is a news article titled 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com' with a URL 'http://time.com/4596928/some-people-who-get-ebola-dont-show-symptoms-study/' and a snippet 'Dec 12, 2016 ... A new report reveals people can get Ebola and not have symptoms.' Below the article is a tag 'news article' with a close button. On the right side, there is a search bar containing 'symptoms|' and a dropdown arrow, and a button labeled 'Create option "symptoms"'. There are also 'BACKW LINK' buttons on the right side.

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' interface. The top navigation bar is purple and contains the NYU logo, the text 'Domain Discovery Tool', a search box with 'EBOLA' entered, a 'Switch Domain' button, and an 'Explore Data View' dropdown menu. On the left, a sidebar menu includes 'Domain information', 'Search', 'Filters', and 'Terms'. The main content area displays 'About 5233 results.' and a pagination control with buttons for 'previous', '1', '2', '...', '436', '437', and 'next'. Below the pagination is a checkbox labeled 'Select ALL results in 437 paginations'. Two search results are visible: 1) A news article from Time.com titled 'Ebola: People Can Get the Virus and Not Have Symptoms' with a green snake icon, dated Dec 12, 2016. 2) A document from the Washington State Department of Health titled 'Ebola symptoms and transmission' with a blue logo. Each result has an 'Add Tag' dropdown menu and a 'Deep Crawl' button. The 'Deep Crawl' button for the second result is highlighted.

Step 4

Run Crawler

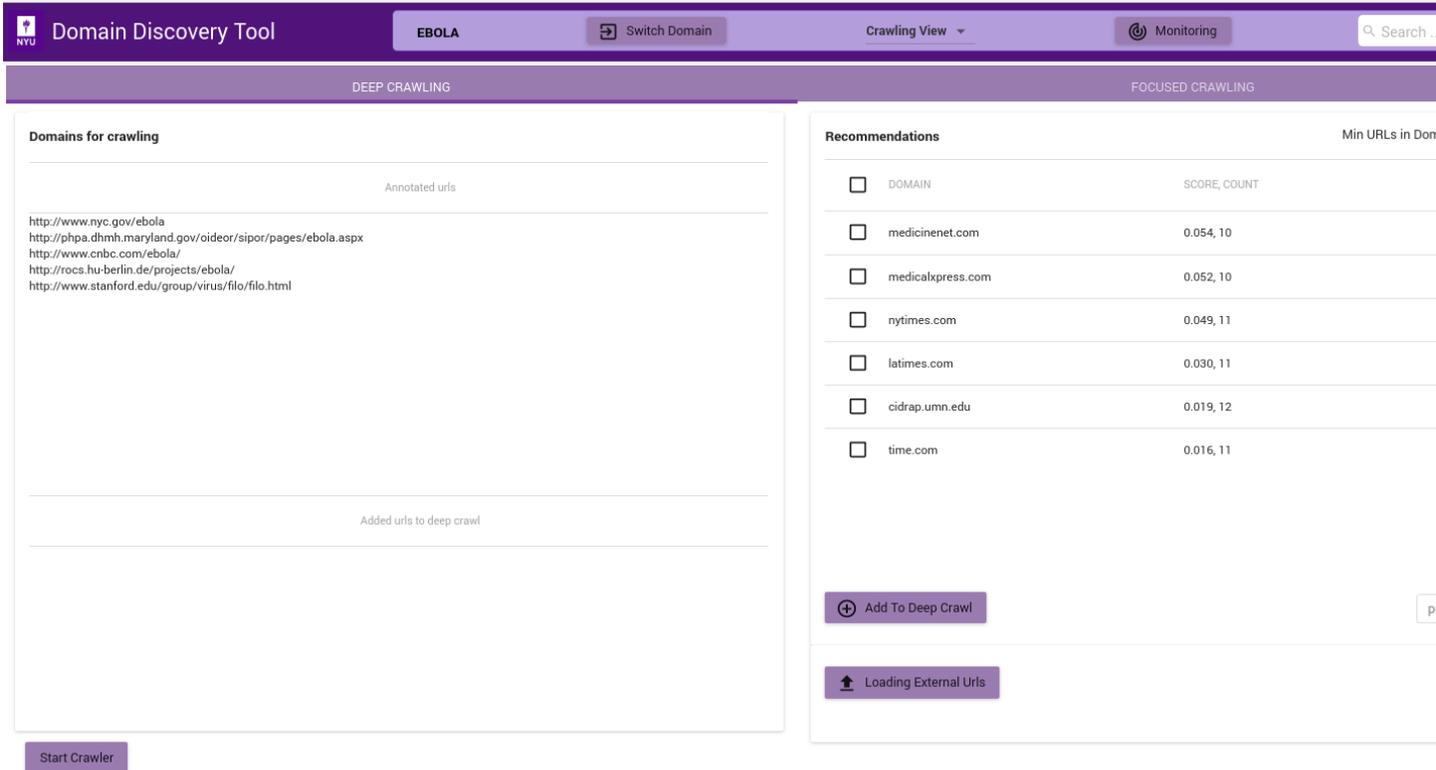
Once a sufficiently good model is available or pages are tagged for a deep crawl you can change from **Explore Data View** to the **Crawler View** to start the crawl shown below:

The screenshot shows the 'Domain Discovery Tool' interface. At the top, there is a purple header with the NYU logo and the text 'Domain Discovery Tool'. To the right of the header, the search term 'EBOLA' is displayed in a light purple box, along with a 'Switch Domain' button. A dropdown menu is open, showing 'Explore Data View' and 'Crawling View'. On the left side, there is a sidebar with navigation options: 'Domain information', 'Search', 'Filters', and 'Terms', each with a downward arrow. Below these is a circular refresh button. The main content area shows 'About 5233 results.' and a pagination control with buttons for 'previous', '1', '2', '...', '436', '437', and 'next'. A checkbox is checked, labeled 'Select ALL results in 437 paginations'. Below this, there are two search results. The first result is from Time.com, titled 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com', with a URL and a snippet. It has a 'news article' tag. The second result is from the Washington State Department of Health, titled 'Ebola symptoms and transmission. :: Washington State Department ...', with a URL and a snippet. It has a 'Deep Crawl' tag. To the right of each result are buttons for 'Add Tag', a thumbs-up icon, and a 'BACKLINK' button.

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*.



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:

The screenshot shows the Domain Discovery Tool interface. At the top, there's a navigation bar with 'Domain Discovery Tool', 'EBOLA', 'Switch Domain', 'Crawling View', and 'Monitoring'. Below this, there are two tabs: 'DEEP CRAWLING' and 'FOCUSED CRAWLING'. The main content area is divided into three sections:

- Model Settings:** A section for selecting positive and negative examples. It has two columns: 'Positive' and 'Negative'. Under 'Positive', there are checkboxes for 'Neutral (5065)', 'Relevant (123)', 'Irrelevant (86)', 'Deep Crawl (3)', and 'outlier (3)'. Under 'Negative', there are checkboxes for 'Neutral (5065)', 'Relevant (123)', 'Irrelevant (86)', 'Deep Crawl (3)', and 'outlier (3)'. There are 'Save' and 'Cancel' buttons at the bottom.
- Crawling:** A section with a 'Start Crawler' button.
- Model:** A section showing 'Total Positive: 126', 'Total Negative: 86', and 'Domain Model (Accuracy): 72.46 %'. Below this is a 'Ratio/Accuracy' color bar ranging from 'Poor' (red) to 'Excellent' (green). There is an 'Export' button at the bottom.

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**.

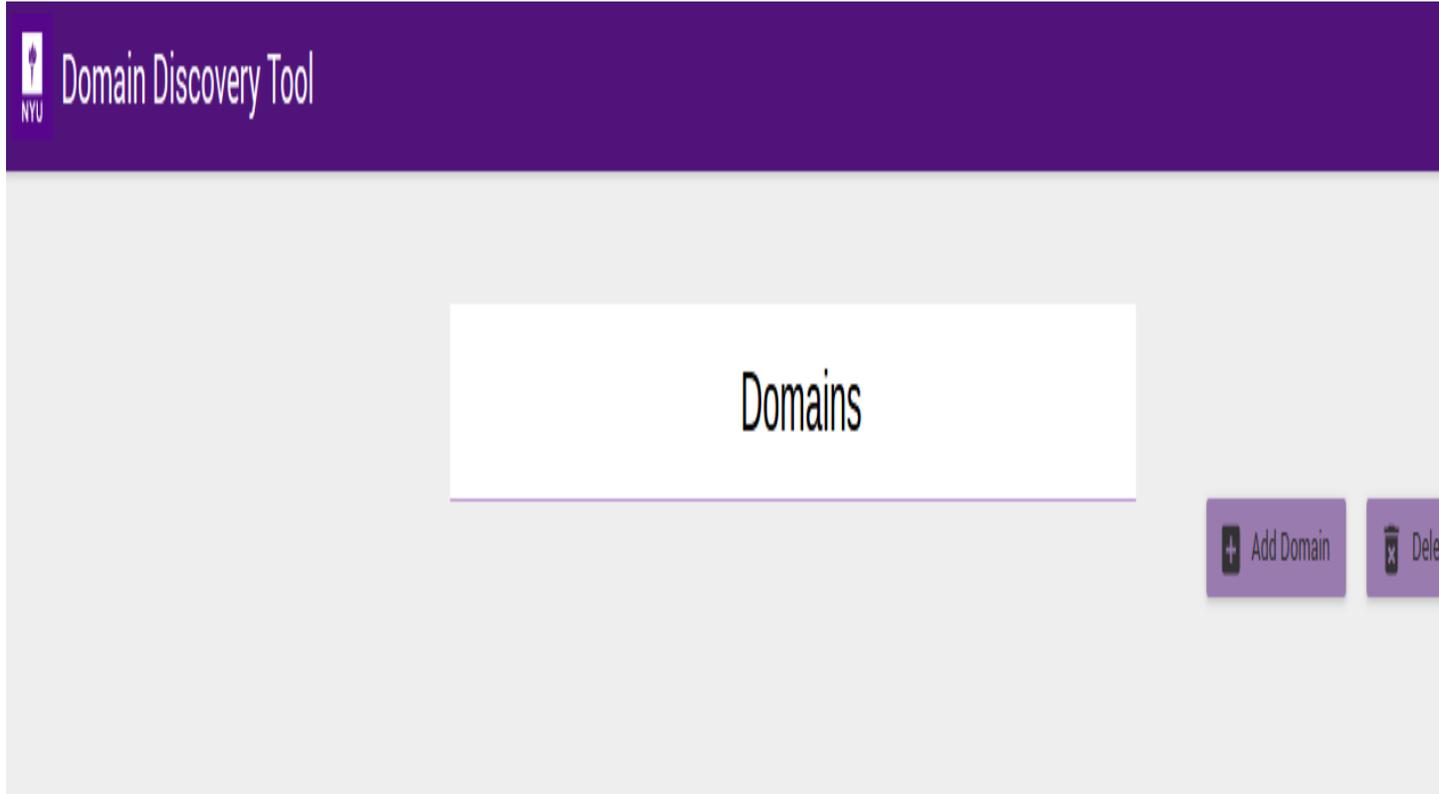
You can see the results of the crawled data in "Crawled Data" in the Filters Tab. When the crawler is running it can be monitored by clicking on the 'Crawler Monitor' button.

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.

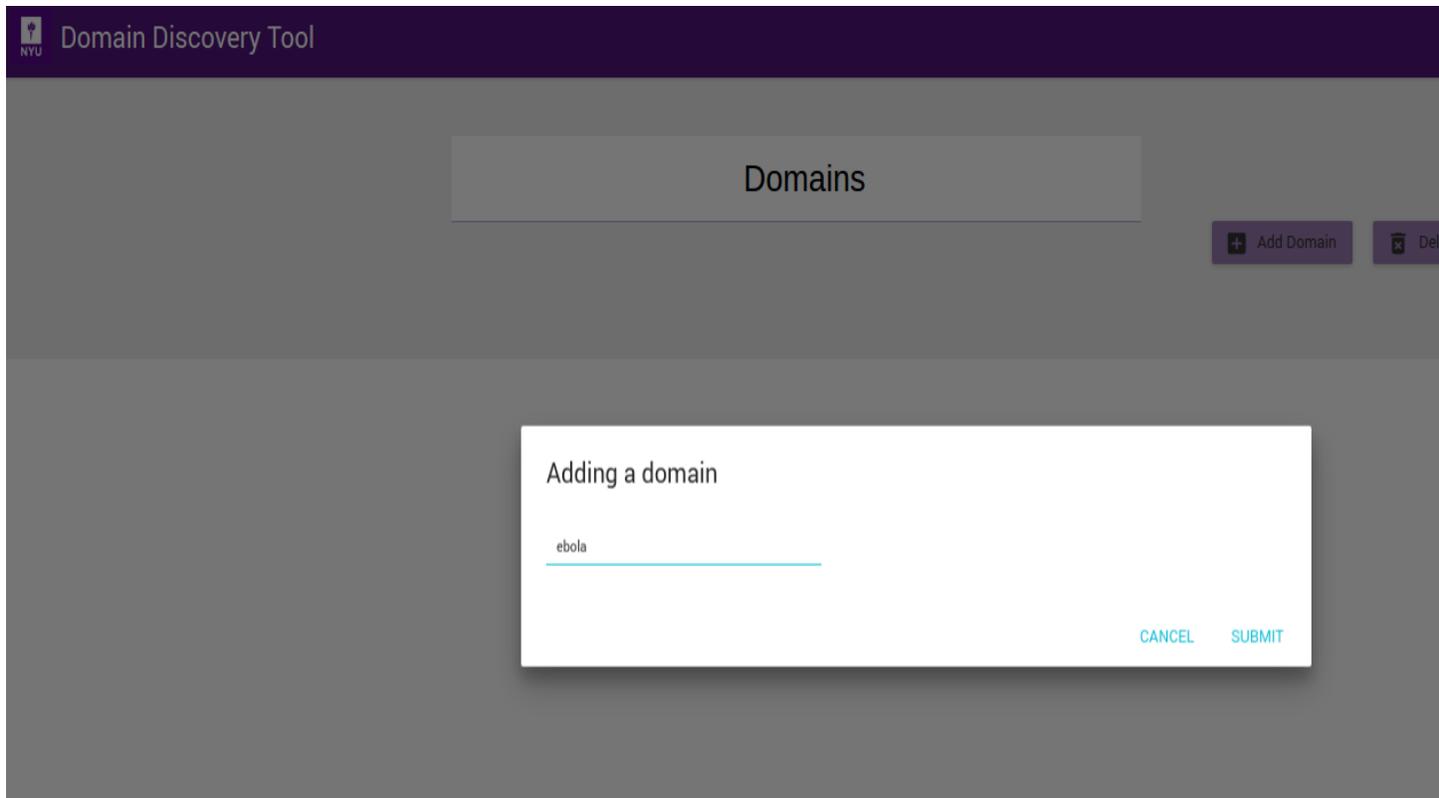
1.3 How To

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

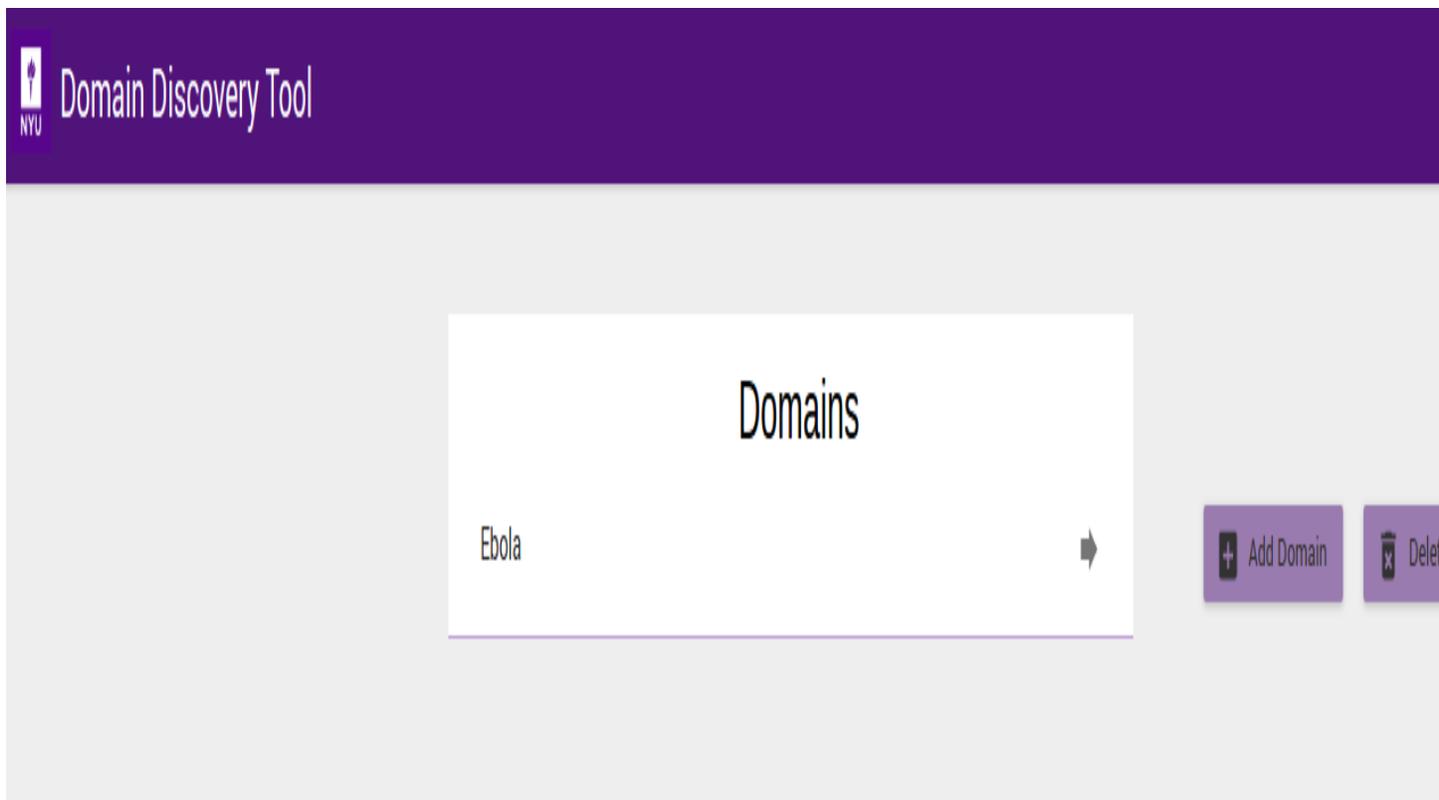
1.3.1 Create Domain



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.



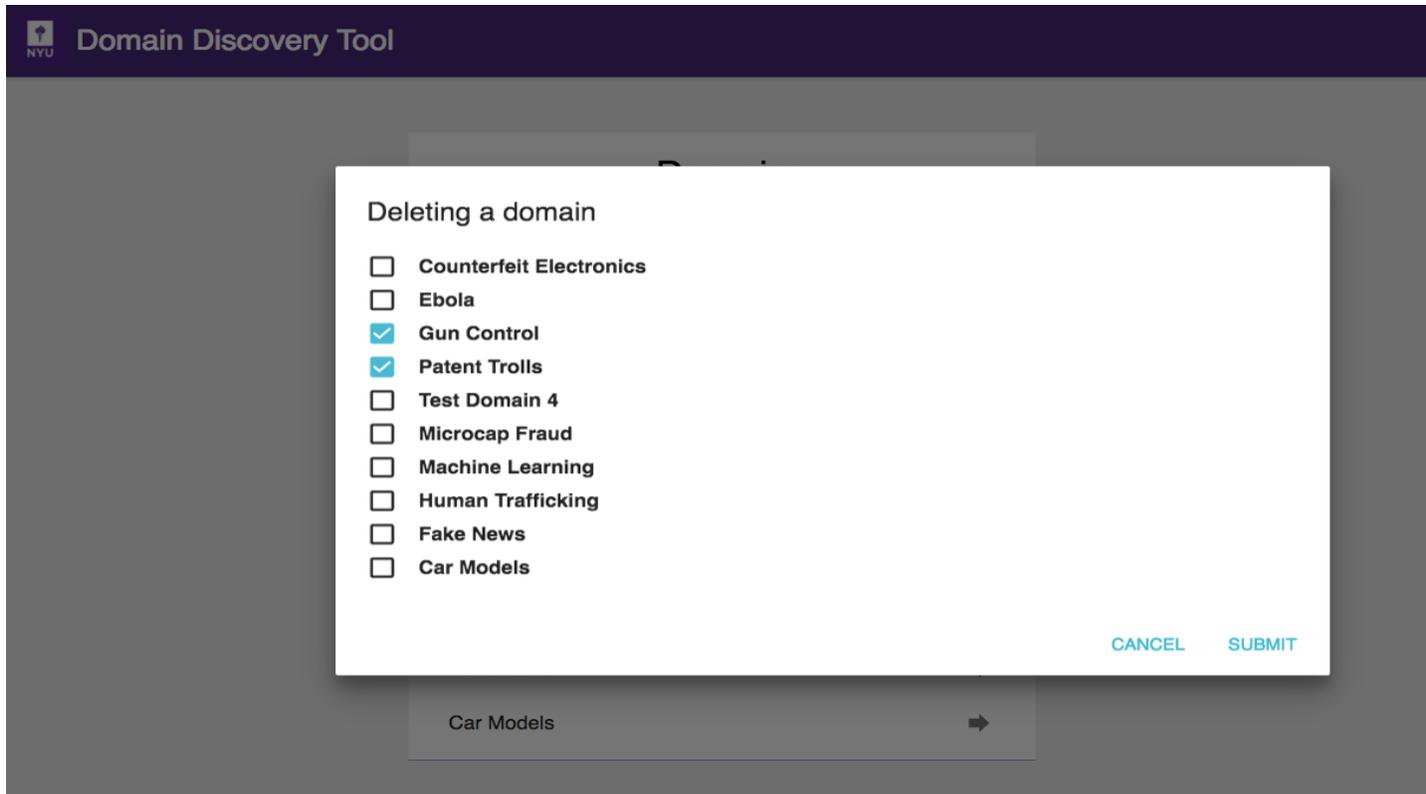
On the **Adding a domain** dialog shown in figure above, enter the name of the domain you would like to create, for example **Ebola**, and click on **Submit** button. You should now see the new domain you added in the list of domains as shown below.



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

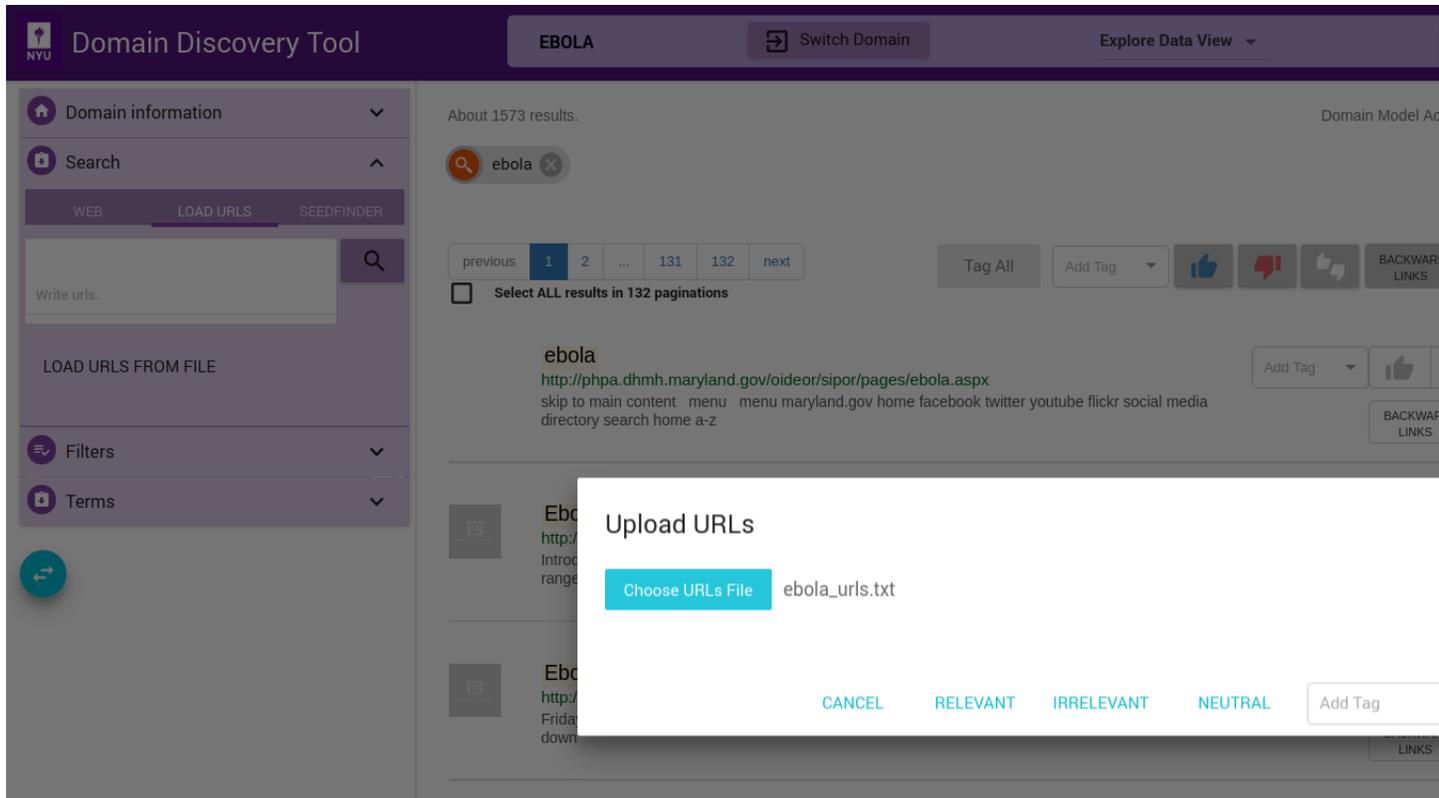
If you have a set of URLs of sites you already know, you can add them from the **LOAD** tab. You can upload the list of URLs in the text box, one fully qualified URL per line, as shown in figure below:

The screenshot shows the Domain Discovery Tool interface. The top navigation bar includes the NYU logo, the tool name 'Domain Discovery Tool', the current domain 'EBOLA', a 'Switch Domain' button, and an 'Explore Data View' dropdown. The left sidebar contains navigation options: 'Domain information', 'Search', 'WEB', 'LOAD URLS', 'SEEDFINDER', 'LOAD URLS FROM FILE', 'Filters', and 'Terms'. The main content area displays search results for 'ebola', showing 'About 1575 results.' and a search bar with 'ebola' entered. Below the search bar are pagination controls (previous, 1, 2, ..., 131, 132, next) and a checkbox for 'Select ALL results in 132 paginations'. Two search results are visible, each with a 'Deep Crawl' button and social media sharing options. The first result is for 'ebola' on 'http://phpa.dhmh.maryland.gov/oideor/sipor/pages/ebola.aspx'. The second result is for 'Ebola' on 'http://www.cnn.com/ebola/'.

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://bmcpyschiatry.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

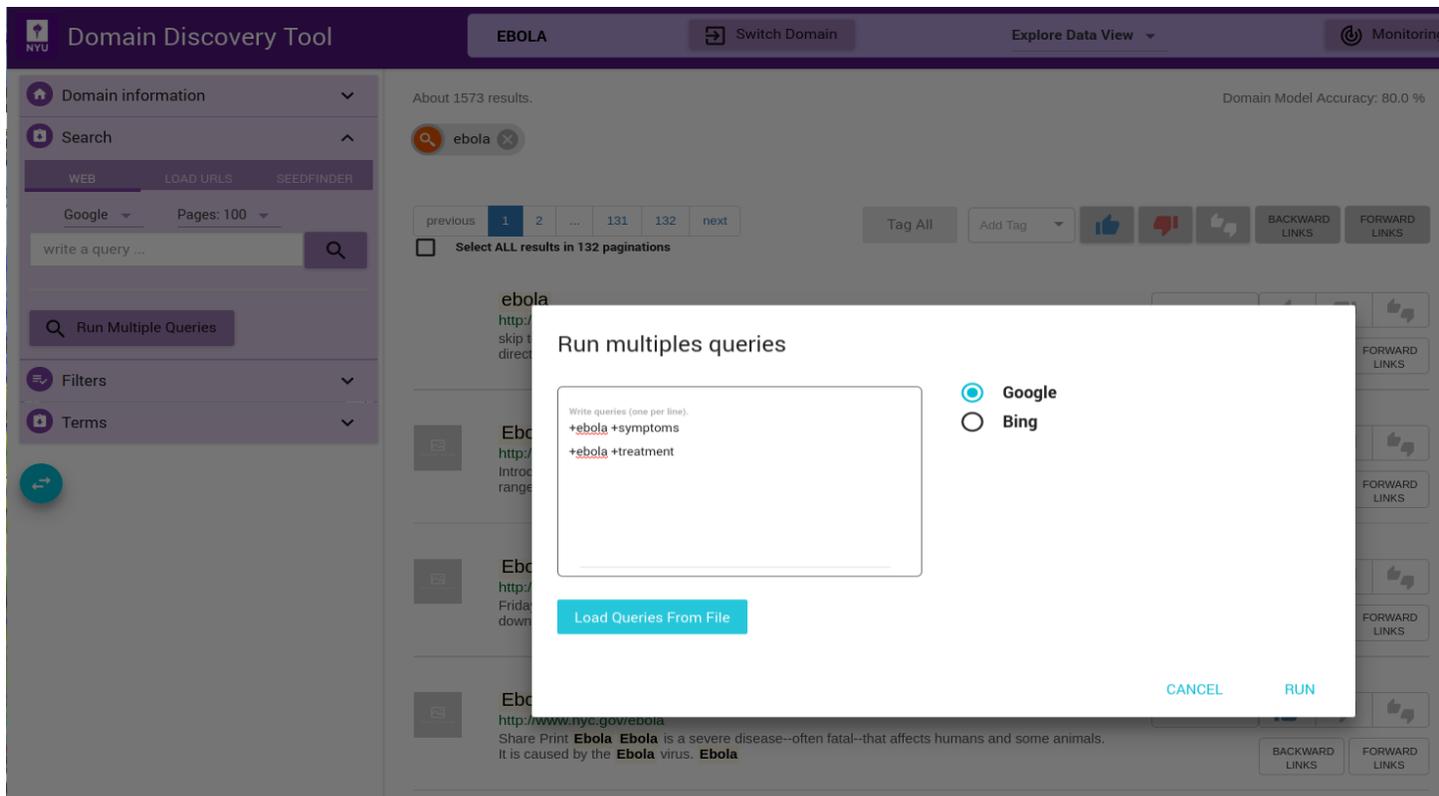
You can do a keywords search on google or bing by clicking on the **WEB** tab. For example, “ebola symptoms”. All queries made are listed in the **Filters** Tab under **Queries**.

The screenshot shows the Domain Discovery Tool interface. The top navigation bar includes the NYU logo, the tool name 'Domain Discovery Tool', the current domain 'EBOLA', a 'Switch Domain' button, and an 'Explore Data View' dropdown. A sidebar on the left contains navigation options: 'Domain information', 'Search', 'Filters', and 'Terms'. The 'Search' section is active, showing a search engine dropdown set to 'Google', a page limit of '100', and a search input field containing 'ebola symptoms'. A 'Run Multiple Queries' button is visible below the search field. The main content area displays 'About 86 results.' and a search bar with 'ebola symptoms'. Below this is a pagination control showing 'previous', '1', '2', '...', '7', '8', and 'next'. A checkbox option 'Select ALL results in 8 paginations' is present. The search results list includes three entries:

- Signs and Symptoms | Ebola Hemorrhagic Fever | CDC** (https://www.cdc.gov/vhf/ebola/symptoms/index.html) with a date of Nov 2, 2014.
- Ebola Virus: Symptoms, Treatment, and Prevention** (https://www.webmd.com/a-to-z-guides/ebola-fever-virus-infection) with a date of Oct 1, 2016.
- Ebola virus and Marburg virus - Symptoms and causes - Mayo Clinic** (https://www.mayoclinic.org/diseases-conditions/ebola-virus/symptoms-causes/dxc-20338674) with a date of Jul 15, 2017.

 Each result entry includes a title, URL, date, and a brief description of symptoms. To the right of each result are buttons for 'Add Tag', a thumbs-up icon, 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**):

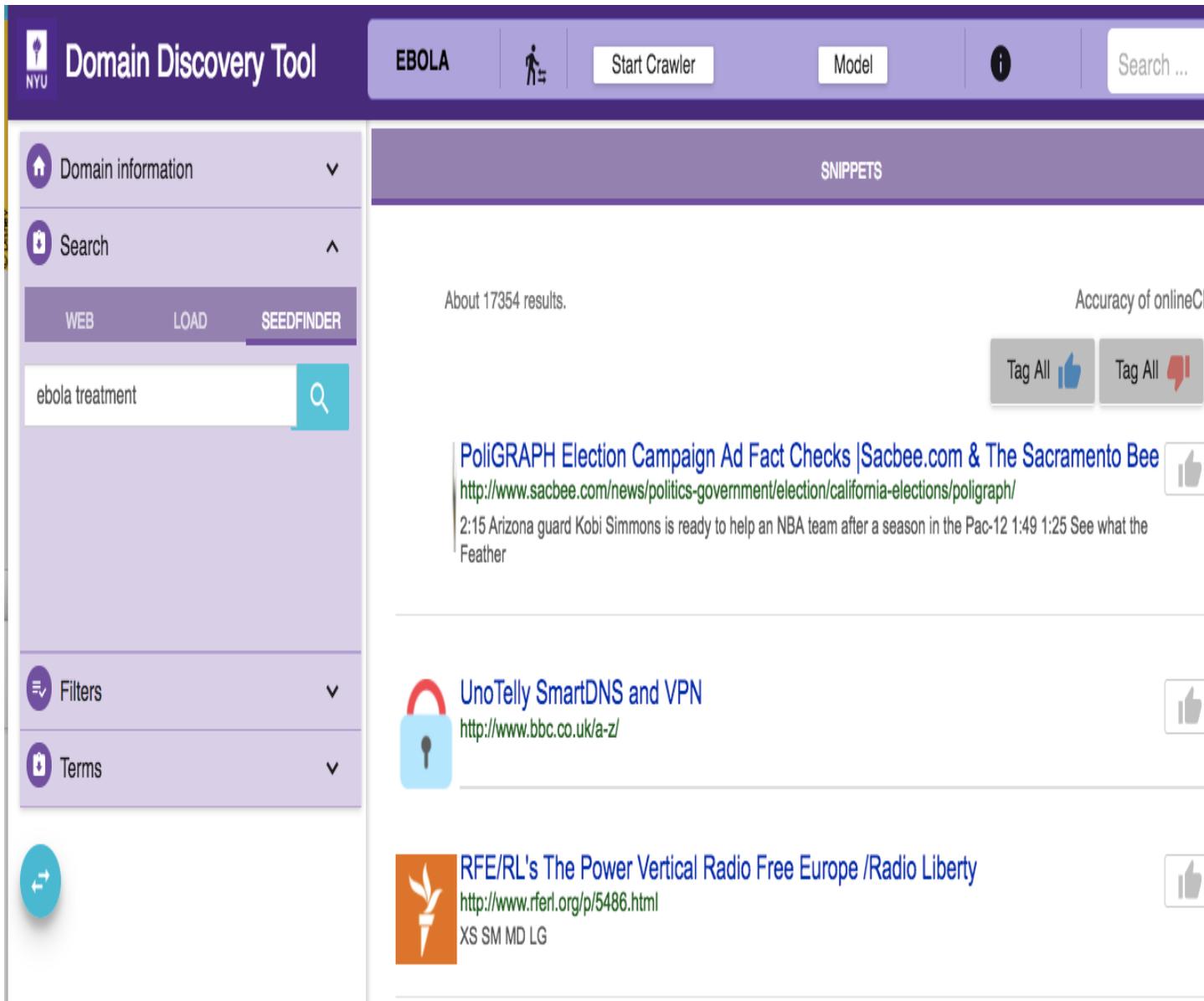


Each of the queries will be issued on Google or Bing (as chosen) and the results made available for exploration and annotation in the **Filters** Tab under **Queries** as **Uploaded**.

SeedFinder

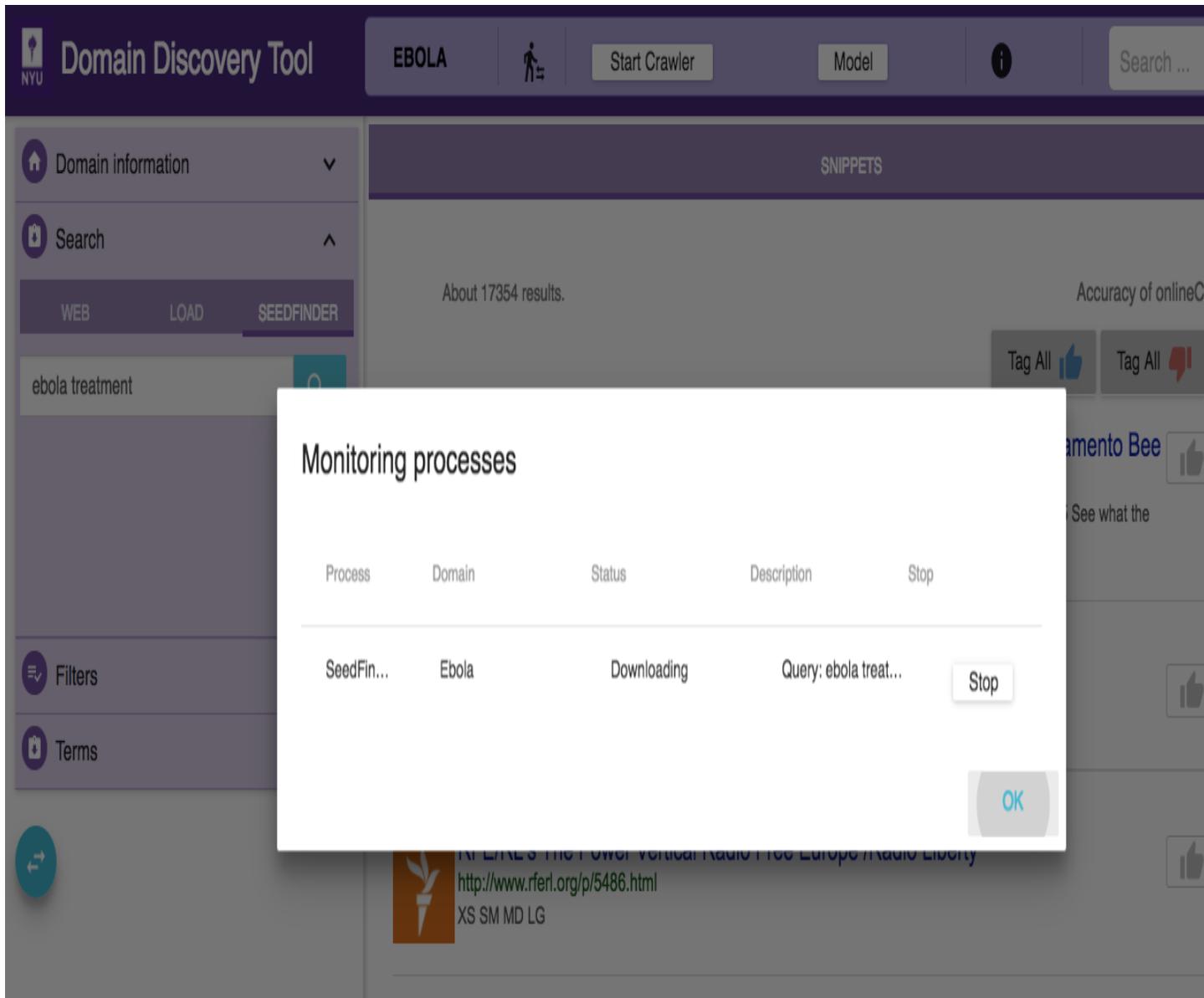
Instead of making multiple queries to Google/Bing yourself you can trigger automated keyword search on Google/Bing and collect more web pages for the domain using the SeedFinder. This requires a domain model. So once you have annotated sufficient pages, indicated by a non-zero accuracy on the top right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

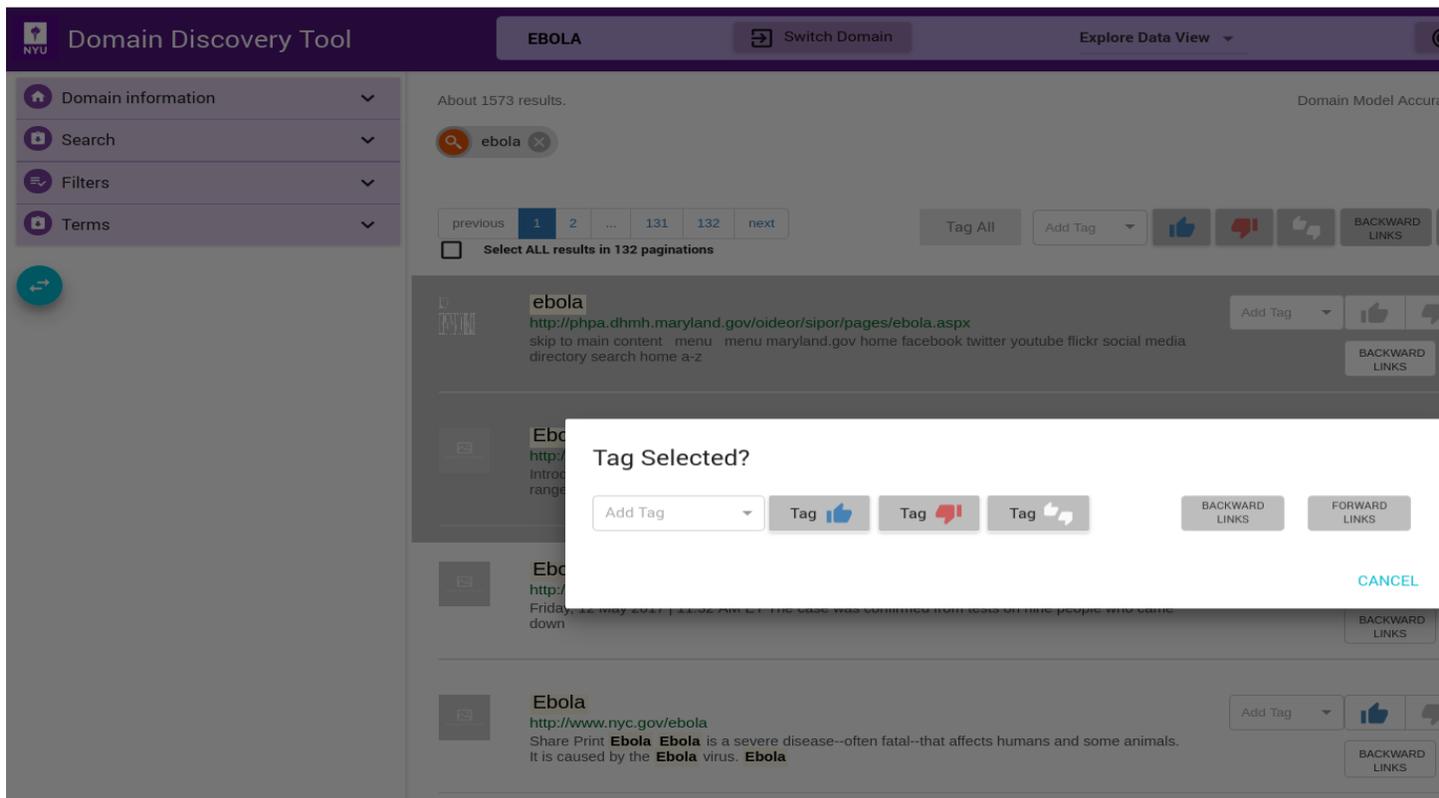
Crawl Individual Pages



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)

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 interface. The top navigation bar includes the NYU logo, the tool name 'Domain Discovery Tool', the current domain 'EBOLA', a 'Switch Domain' button, and the view mode 'Explore Data View'. A sidebar on the left contains navigation options: 'Domain information', 'Search', 'Filters', and 'Terms'. The 'Search' section is active, showing a search bar with 'ebola symptoms' and a search icon. Below the search bar are options for 'Google' and 'Pages: 100'. A 'Run Multiple Queries' button is also present. The main content area shows search results for 'ebola symptoms'. It indicates 'About 86 results.' and shows a search filter 'ebola symptoms'. Below this is a pagination control with 'previous', '1', '2', '...', '7', '8', and 'next'. A 'Tag All' button and an 'Add Tag' dropdown are visible. The results list includes three entries:

- Signs and Symptoms | Ebola Hemorrhagic Fever | CDC** with URL <https://www.cdc.gov/vhf/ebola/symptoms/index.html>. The snippet mentions symptoms like fever, severe headache, muscle pain, weakness, fatigue, diarrhea, vomiting, and abdominal pain.
- Ebola Virus: Symptoms, Treatment, and Prevention** with URL <https://www.webmd.com/a-to-z-guides/ebola-fever-virus-infection>. The snippet states that WebMD explains the latest info on the rare but deadly disease Ebola, including its spread.
- Ebola virus and Marburg virus - Symptoms and causes - Mayo Clinic** with URL <https://www.mayoclinic.org/diseases-conditions/ebola-virus/symptoms-causes/dxc-20338674>. The snippet notes that signs and symptoms typically begin abruptly within five to ten days of infection with Ebola or Marburg virus.

 Each result has an 'Add Tag' button and a 'BACKLINK' button.

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' interface. The top navigation bar includes 'EBOLA', 'Switch Domain', 'Explore Data View', and 'Monitoring'. The left sidebar contains '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 the search bar, there are pagination controls (previous, 1, 2, ..., 6, 7, next) and a checkbox for 'Select ALL results in 7 paginations'. The results list includes four entries, each with a title, URL, snippet, 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)

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

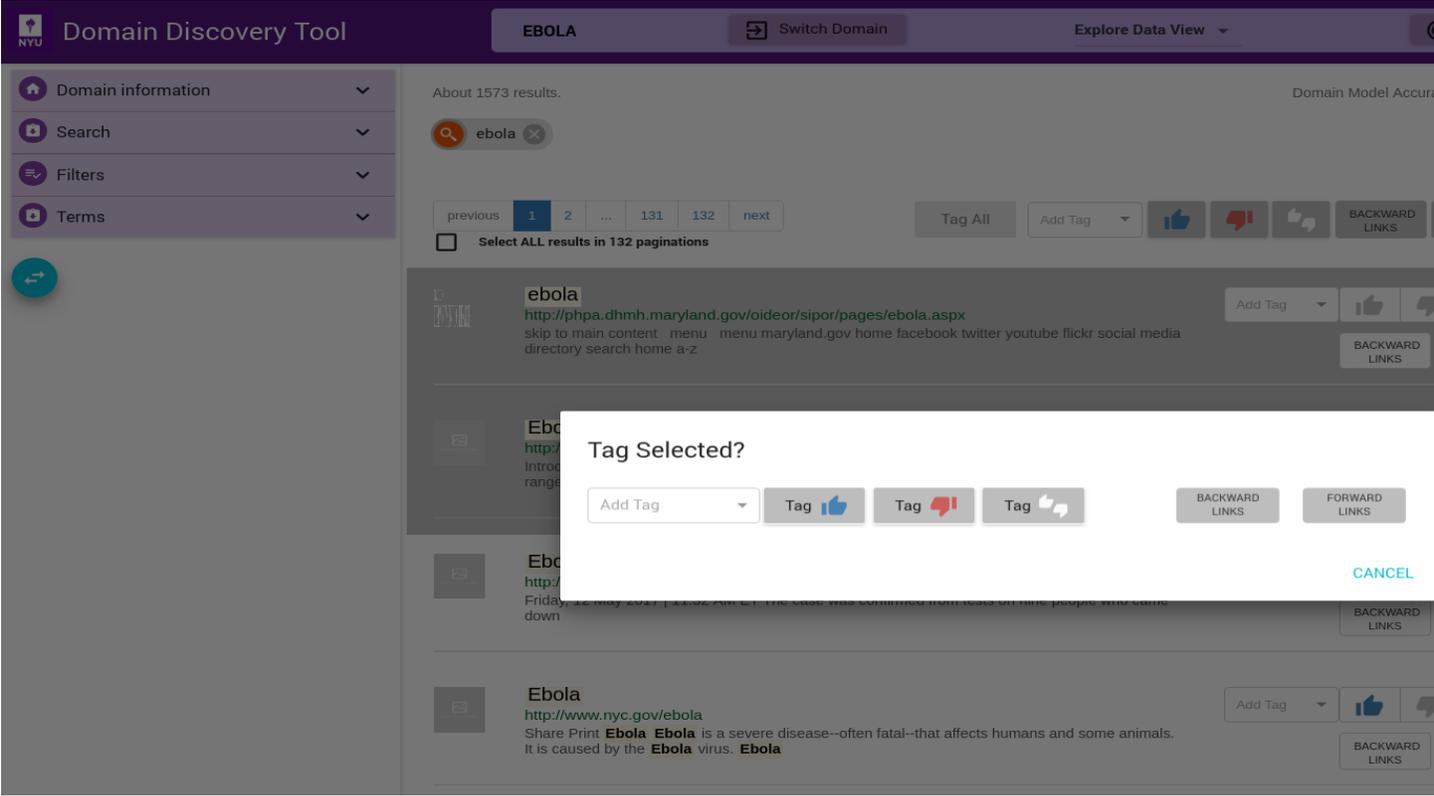
Tag Individual Pages



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 interface. The top navigation bar is purple and contains the NYU logo, the text 'Domain Discovery Tool', the search term 'EBOLA', a 'Switch Domain' button, and a dropdown menu for 'Explore Data View'. On the left, a sidebar menu includes 'Domain information', 'Search', 'Filters', and 'Terms'. The main content area displays 'About 5233 results.' and a pagination control with buttons for 'previous', '1', '2', '...', '436', '437', and 'next'. Below the pagination is a checkbox labeled 'Select ALL results in 437 paginations'. A search result is shown with a thumbnail 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 result are social media icons (Facebook, Twitter, etc.) and a 'Tag All' button. A 'Create option "symptoms"' button is visible on the far right. At the bottom of the result area, there is a 'news article' tag with a close 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 information
- Search
- Filters
- Terms



About 5233 results.

Domain Model

previous 1 2 ... 436 437 next

Tag All Add Tag [thumbs up] [thumbs down] [share] 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.

Add Tag [thumbs up] [thumbs down] BACK LINK

news article



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.

Add Tag [thumbs up] [thumbs down] BACK LINK
 news article
 Deep Crawl
 outlier

Deep Crawl

1.3.6 Extracted Terms Summary

The screenshot displays the Domain Discovery Tool (DDT) interface. The top navigation bar includes the NYU logo, the tool name 'Domain Discovery Tool', and a search bar containing 'EBOLA SYMPTOMS'. Below the search bar are buttons for 'Start Crawler', 'Model', and a search icon. The left sidebar contains navigation tabs: 'Domain information', 'Search', 'Filters', and 'Terms'. The 'Terms' tab is active, showing a list of terms with horizontal bars indicating their frequency in relevant (blue) and irrelevant (red) pages. The main content area shows search results for 'ebola AND symptoms', with 'About 395 results.' displayed. Two search results are visible: 'Ebola Symptoms' from eMedTV and 'Ebola Symptoms |Everyday Health' from Everyday Health. Each result includes a thumbnail, title, URL, and a brief description.

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.

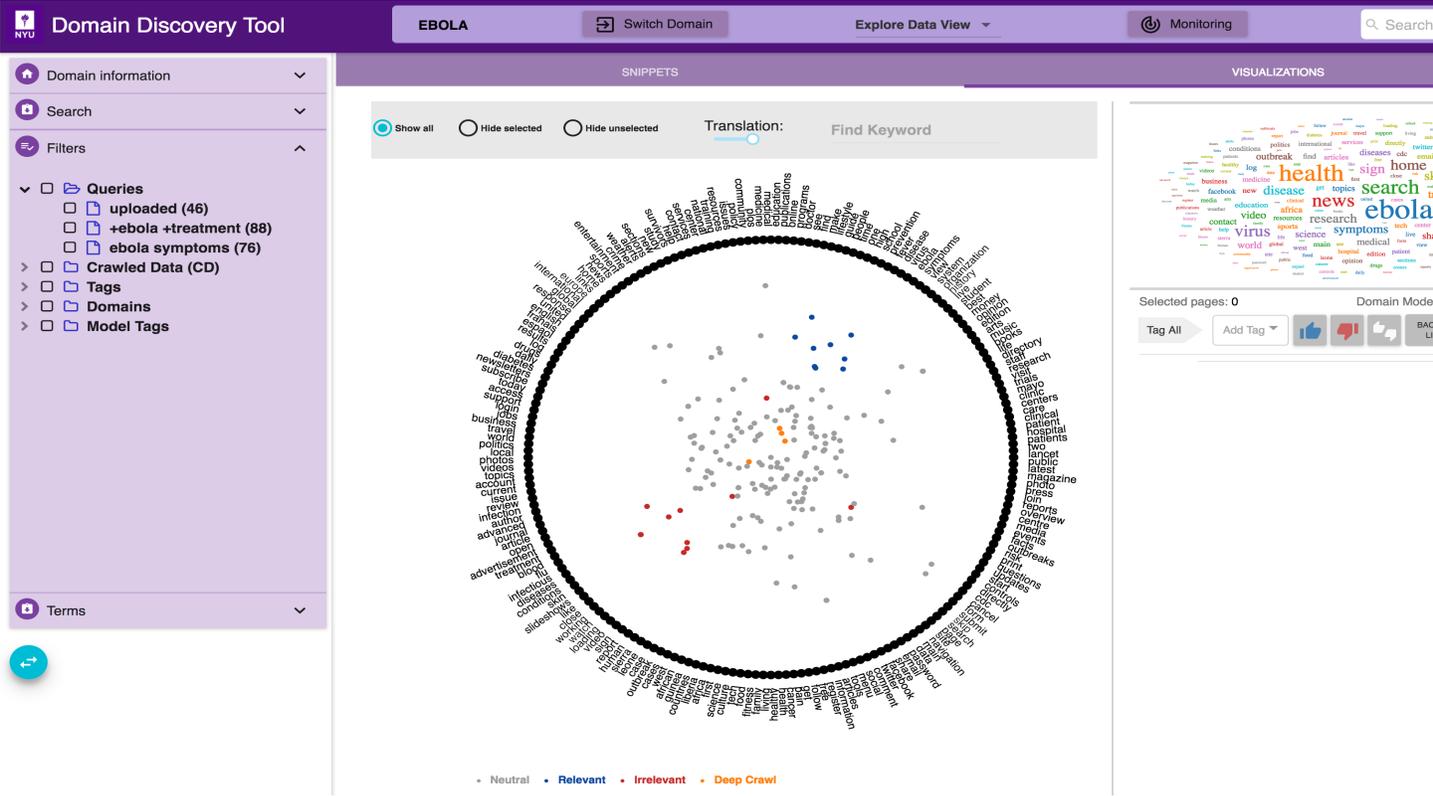
Terms can be tagged as 'Positive' and 'Negative' by 1-click and 2-click respectively. The tags are stored in the active data source. When the update terms button is clicked, the positively and negatively annotated terms are used to re-rank the other terms. Terms help the expert understand and discover 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

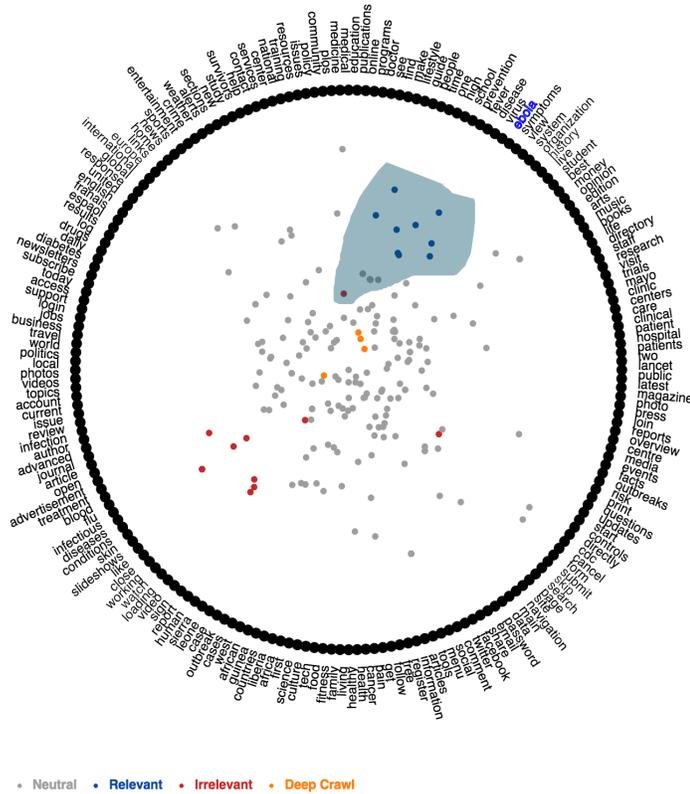
Select the **Visualization Tab** in the **Explore Data View** to see the multidimensional scaling visualization using RadViz.



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 a 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.

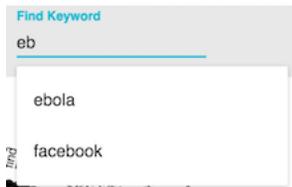
Hide unselected: This option hides the pages that are not selected.

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.

The screenshot shows the Domain Discovery Tool interface. At the top, there is a purple header with the NYU logo and the text 'Domain Discovery Tool'. To the right of the header, the search term 'EBOLA' is displayed, along with a 'Switch Domain' button. A dropdown menu is open, showing 'Explore Data View' and 'Crawling View'. On the left side, there is a sidebar with navigation options: 'Domain information', 'Search', 'Filters', and 'Terms', each with a dropdown arrow. Below the sidebar is a circular refresh button. The main content area shows 'About 5233 results.' and a pagination bar with 'previous', '1', '2', '...', '436', '437', and 'next'. There is a checkbox labeled 'Select ALL results in 437 paginations'. Below this, two search results are visible. The first result is from Time.com, titled 'Ebola: People Can Get the Virus and Not Have Symptoms | Time.com', with a URL and a date of Dec 12, 2016. The second result is from the Washington State Department of Health, titled 'Ebola symptoms and transmission. :: Washington State Department ...', with a URL and a description of symptoms. Each result has an 'Add Tag' button and a 'BACKLINK' button. There are also social media icons (thumbs up, thumbs down, share) for each result. At the bottom of each result, there is a tag: 'news article' for the first result and 'Deep Crawl' for 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 screenshot displays the 'Domain Discovery Tool' interface. At the top, there is a navigation bar with 'Domain Discovery Tool', 'EBOLA', 'Switch Domain', 'Crawling View', and 'Monitoring'. Below this, the interface is divided into two main sections: 'DEEP CRAWLING' and 'FOCUSED CRAWLING'.

The 'DEEP CRAWLING' section on the left contains a 'Domains for crawling' area. It has two sub-sections: 'Annotated urls' and 'Added urls to deep crawl'. The 'Annotated urls' section lists several URLs related to Ebola, such as '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/', and 'http://www.stanford.edu/group/virus/filo/filo.html'. Below this, there is a 'Start Crawler' button.

The 'FOCUSED CRAWLING' section on the right features a 'Recommendations' table. The table has two columns: 'DOMAIN' and 'SCORE, COUNT'. It lists several domains with their respective scores and counts:

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

Below the table, there are two buttons: 'Add To Deep Crawl' and '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.

Focused Crawl

The figure below shows the Focused Crawler View:

The screenshot displays the Domain Discovery Tool interface. At the top, there is a navigation bar with 'Domain Discovery Tool' on the left, 'EBOLA' in the center, and 'Switch Domain', 'Crawling View', and 'Monitoring' on the right. Below this, a secondary bar indicates 'DEEP CRAWLING' and 'FOCUSED CRAWLING' options.

The main content area is divided into three sections:

- Model Settings:** A section for selecting positive and negative examples. It has two columns: 'Positive' and 'Negative'. Under 'Positive', 'Relevant (123)' and 'Deep Crawl (3)' are checked. Under 'Negative', 'Irrelevant (86)' is checked. There are 'Save' and 'Cancel' buttons at the bottom.
- Crawling:** A section with a 'Start Crawler' button.
- Model:** A section showing 'Total Positive: 126', 'Total Negative: 86', and 'Domain Model (Accuracy): 72.46 %'. Below this is a 'Ratio/Accuracy' color bar ranging from 'Poor' (red) to 'Excellent' (green). An 'Export' button is at the bottom.

The 'Terms' window on the right lists terms like 'ebola virus', 'symptoms', 'outbreak', 'fever', 'blood', 'ebola', 'hemorrhagic', 'fluids', 'care', 'flu', and 'ebola virus disease', each with a corresponding bar chart.

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`