ddt Documentation

Release 1.0

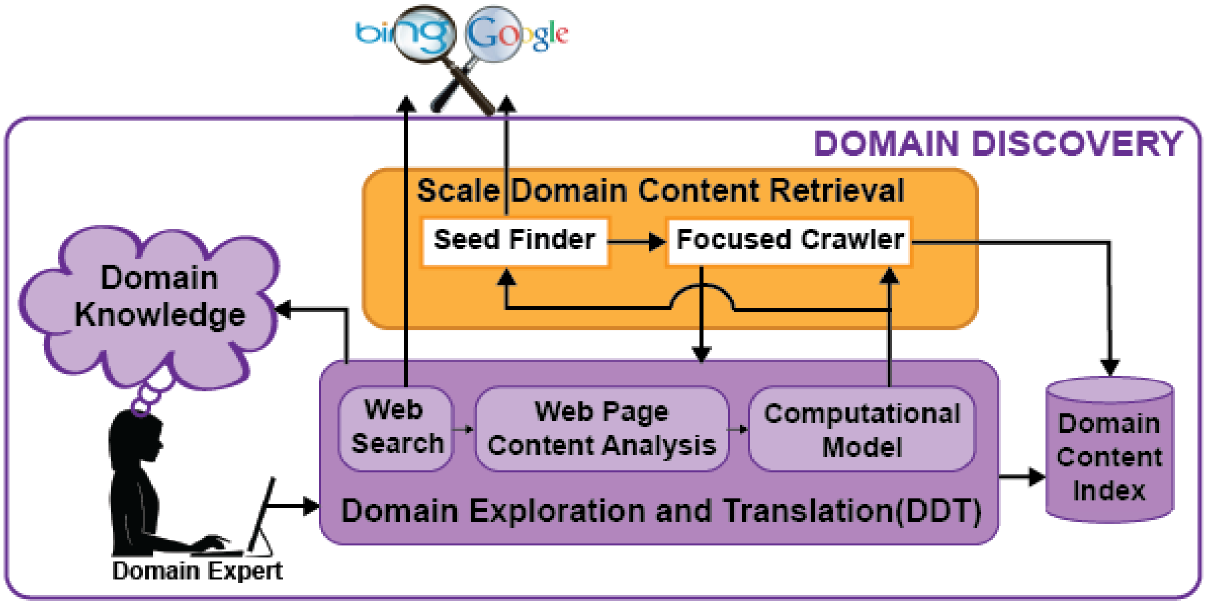
Yamuna Krishnamurthy

Sep 13, 2018

Contents

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.

CHAPTER 1

Contents

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 in step 3. in docker compose install for linux

In order to run the docker version in background download `docker-compose.yml`. Use the following commands to run it:

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

The above commands will start elasticsearch and DDT processes. The elasticsearch and DDT 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}
>>> ./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.

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:

```
>>> 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:

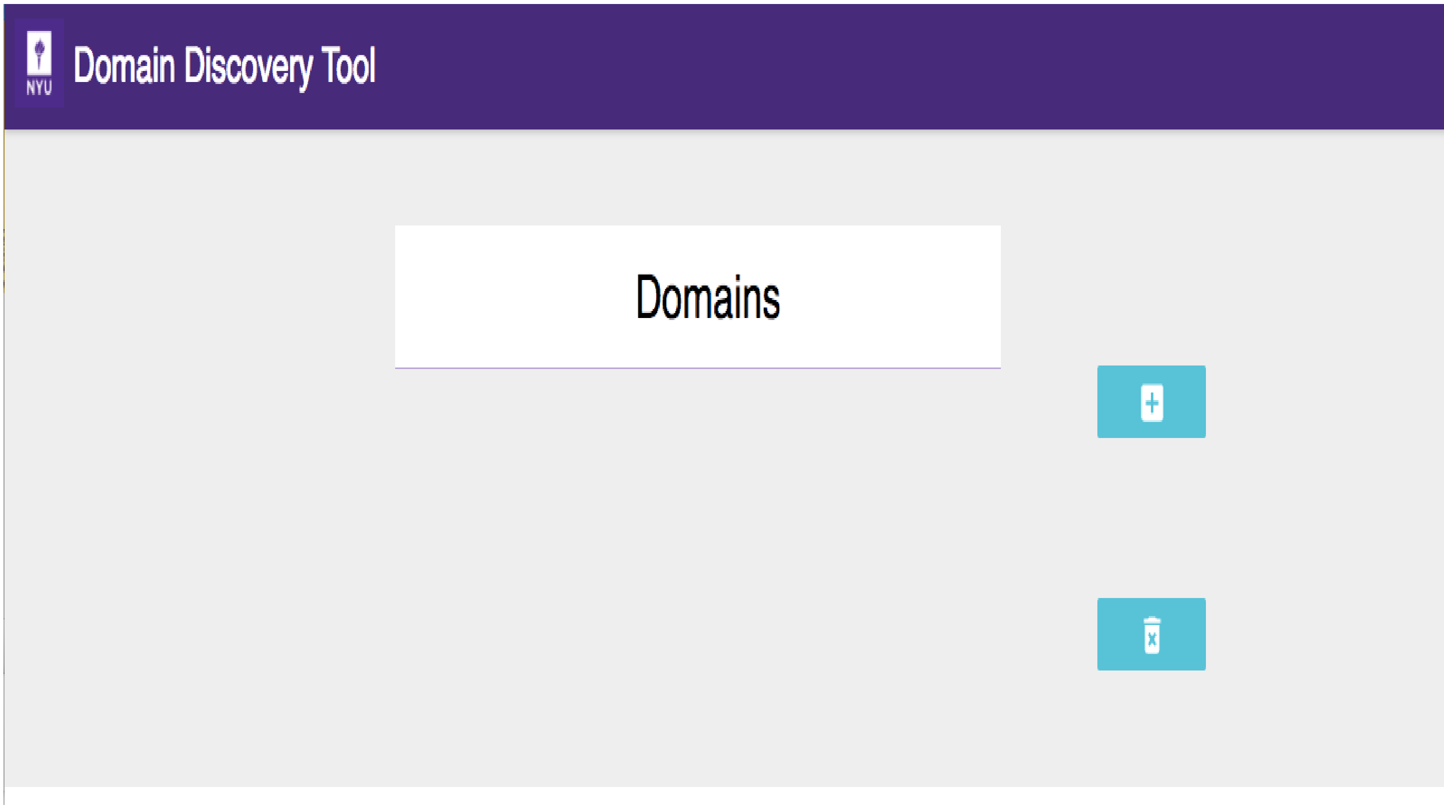
```
>>> ./bin/ddt-dev
```

Use Domain Discovery Tool

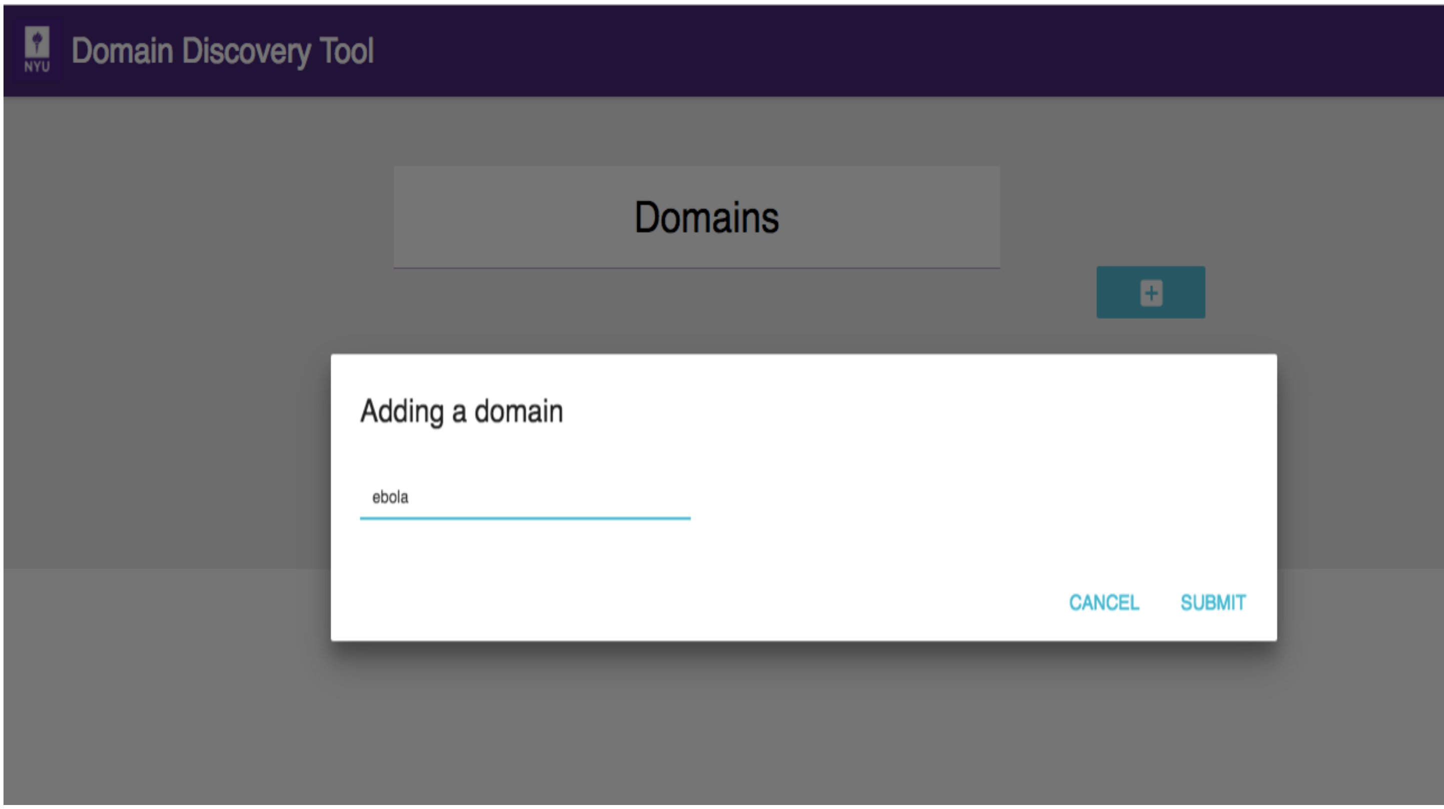
1.2 Using the Domain Discovery Tool

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

1.2.1 Create Domain



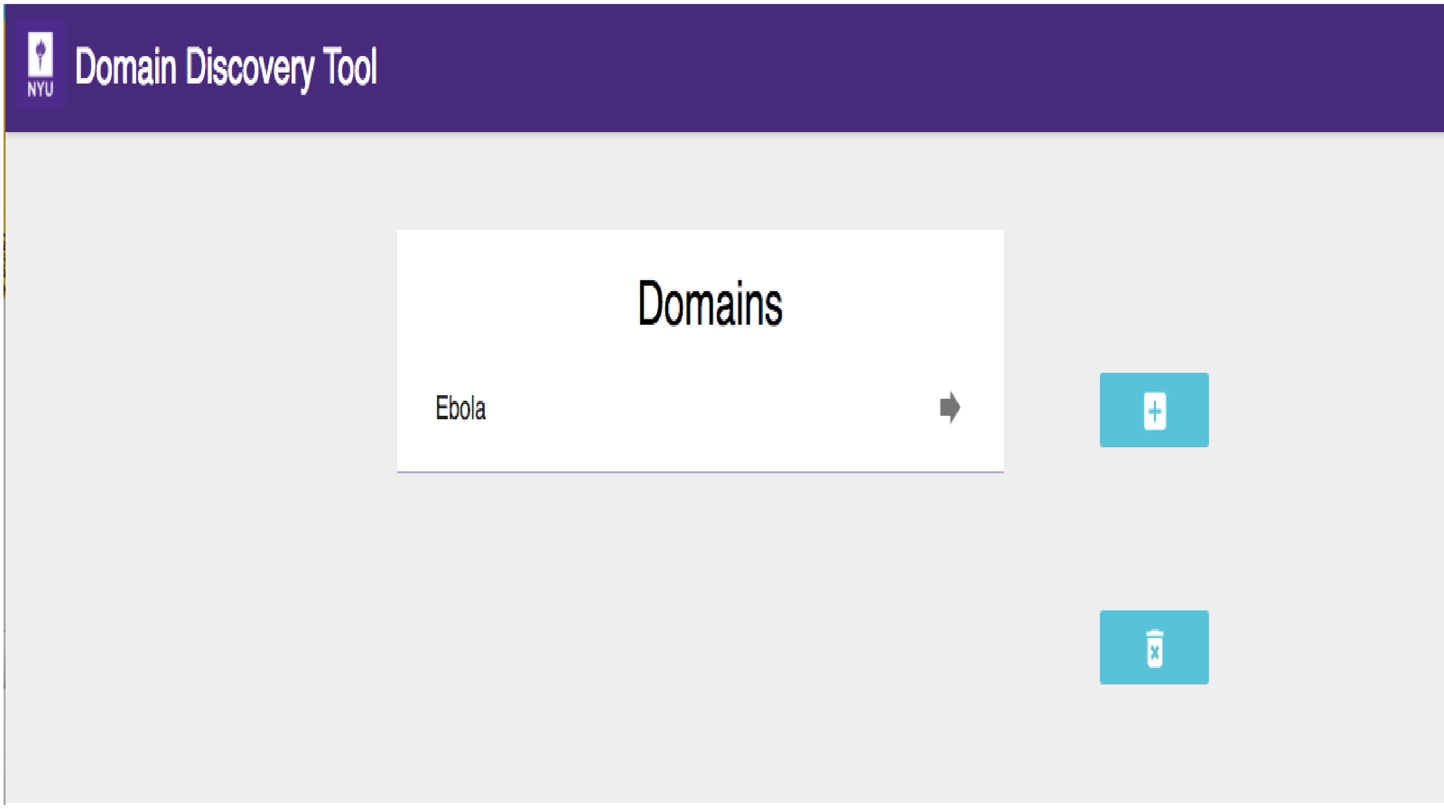
Begin by adding a domain on the Domains page (initial page), shown in the figure above, by clicking on the ⊞ button. Domain maintains context of domain discovery.



Adding a domain

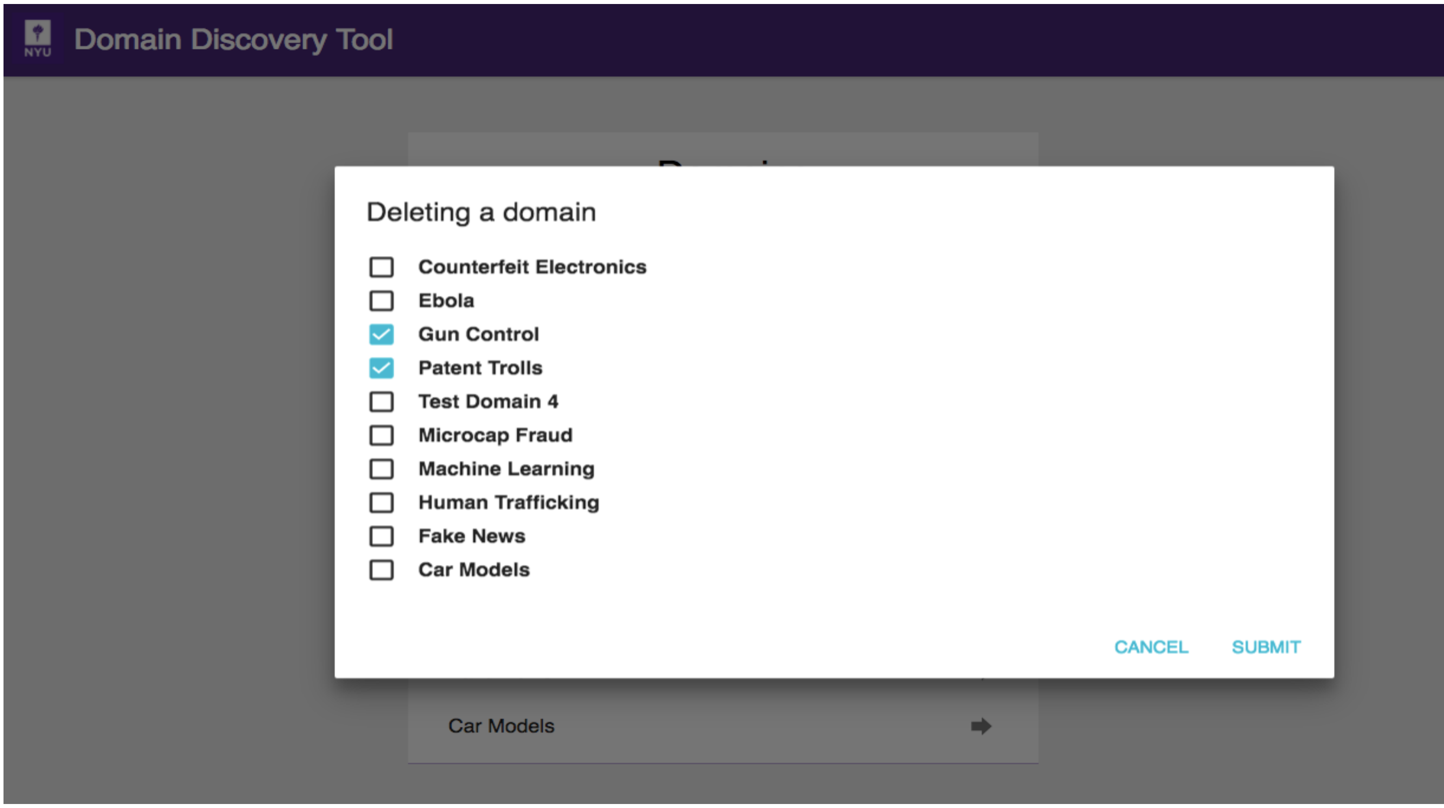
ebola

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.

Domains can be deleted by clicking on the 🗑 button.



Deleting a domain

- □ Counterfeit Electronics
- □ Ebola
- ☑ Gun Control
- ☑ Patent Trolls
- □ Test Domain 4
- □ Microcap Fraud
- □ Machine Learning
- □ Human Trafficking
- □ Fake News
- □ Car Models

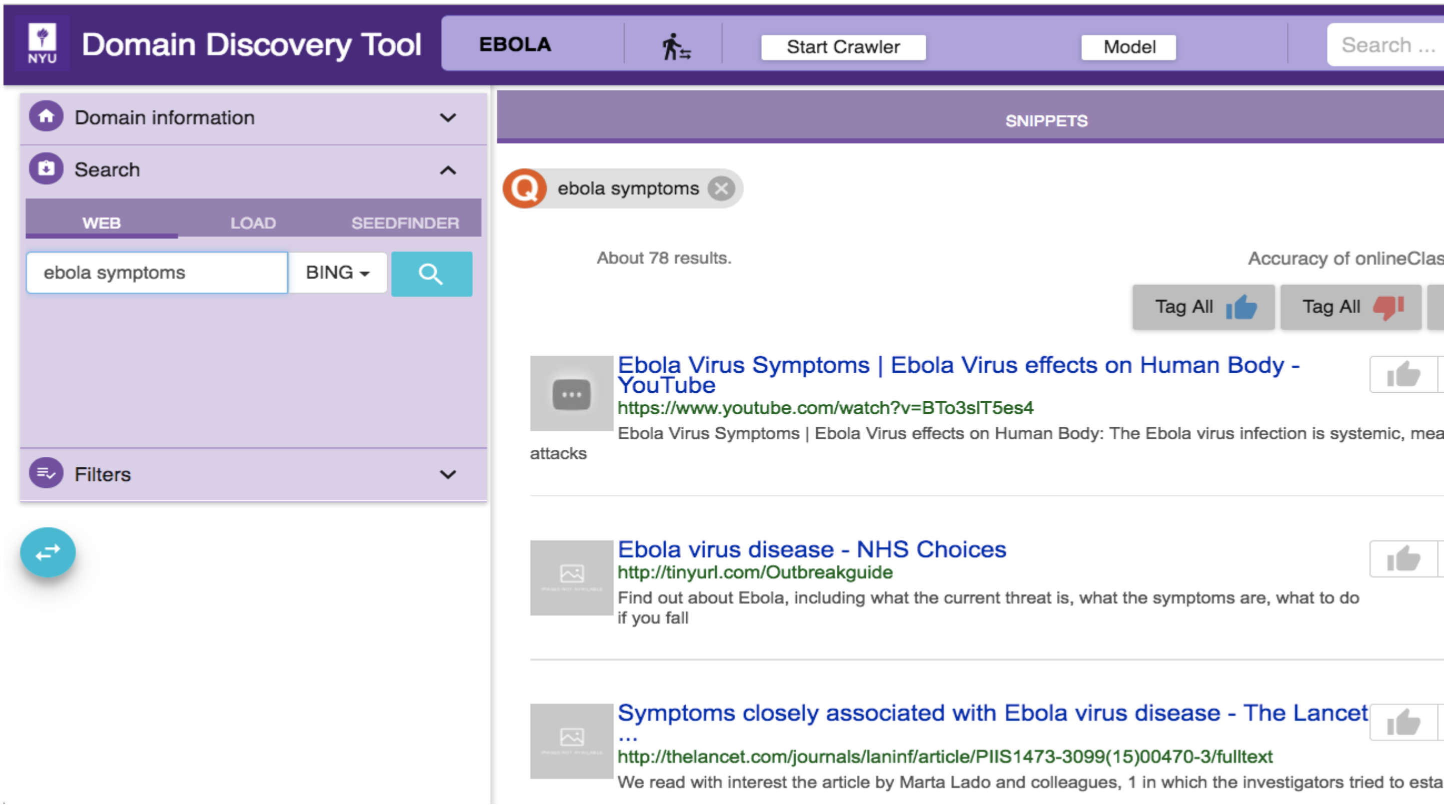
On the **Deleting a domain** dialog select the domains to be deleted in the list of current domains and click on **Submit** button. They will no longer appear on the domains list.

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

1.2.2 Acquire Data

Continuing with our example of the **Ebola** domain, we show here the 3 methods of uploading data. Expand the Search tab on the left panel. You can add data to the domain in the following ways:

Web Search



ebola symptoms

About 78 results.

Accuracy of onlineClas

Ebola Virus Symptoms | Ebola Virus effects on Human Body - YouTube
https://www.youtube.com/watch?v=BTo3slT5es4
Ebola Virus Symptoms | Ebola Virus effects on Human Body: The Ebola virus infection is systemic, mea
attacks

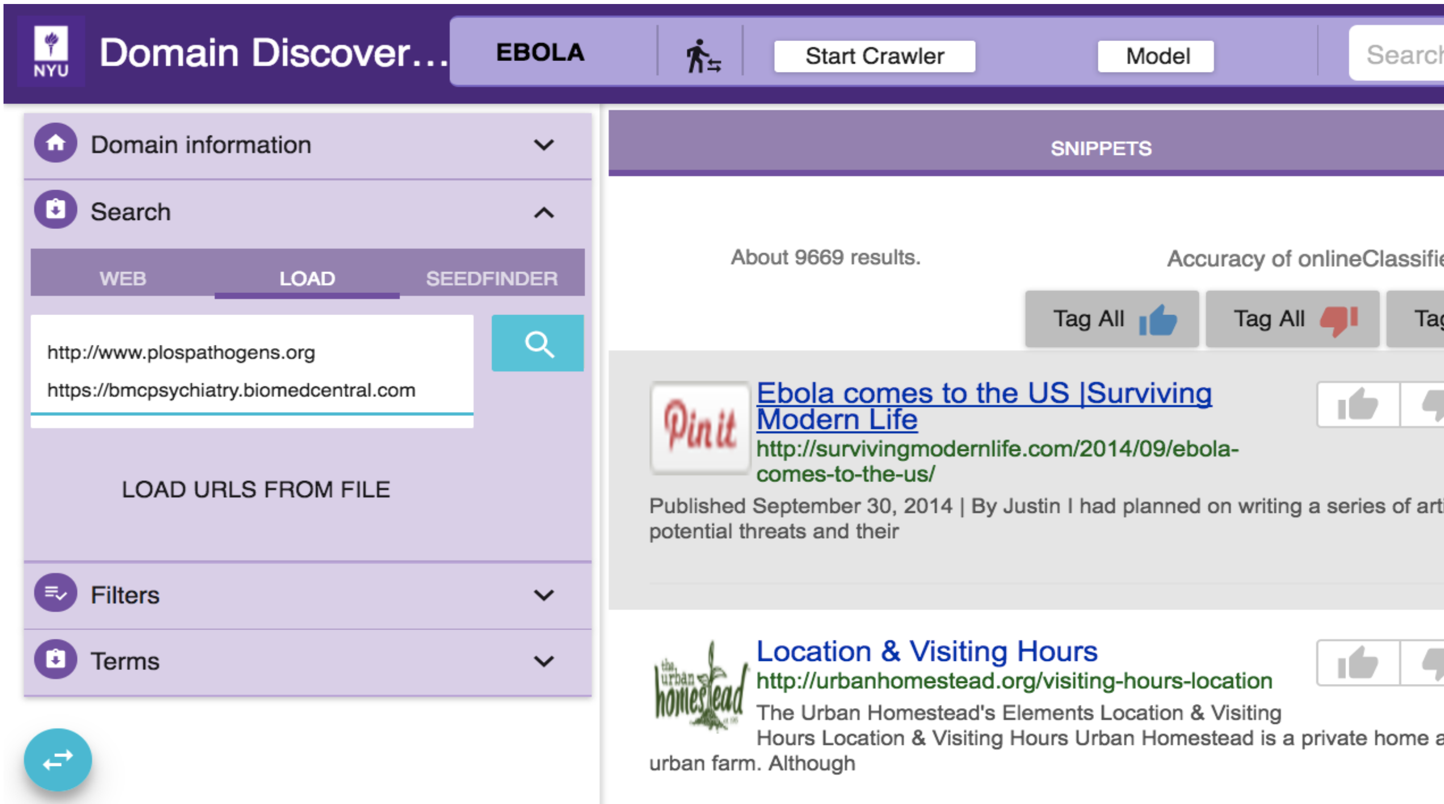
Ebola virus disease - NHS Choices
http://tinyurl.com/Outbreakguide
Find out about Ebola, including what the current threat is, what the symptoms are, what to do if you fall

Symptoms closely associated with Ebola virus disease - The Lancet ...
http://thelancet.com/journals/laninf/article/PIIS1473-3099(15)00470-3/fulltext
We read with interest the article by Marta Lado and colleagues, 1 in which the investigators tried to esta

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**.

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 as shown in figure below:



http://www.plospathogens.org
https://bmcpsychiatry.biomedcentral.com

Ebola comes to the US |Surviving Modern Life
http://survivingmodernlife.com/2014/09/ebola-comes-to-the-us/
Published September 30, 2014 | By Justin I had planned on writing a series of art potential threats and their

Location & Visiting Hours
http://urbanhomestead.org/visiting-hours-location
The Urban Homestead's Elements Location & Visiting Hours Location & Visiting Hours Urban Homestead is a private home a urban farm. Although

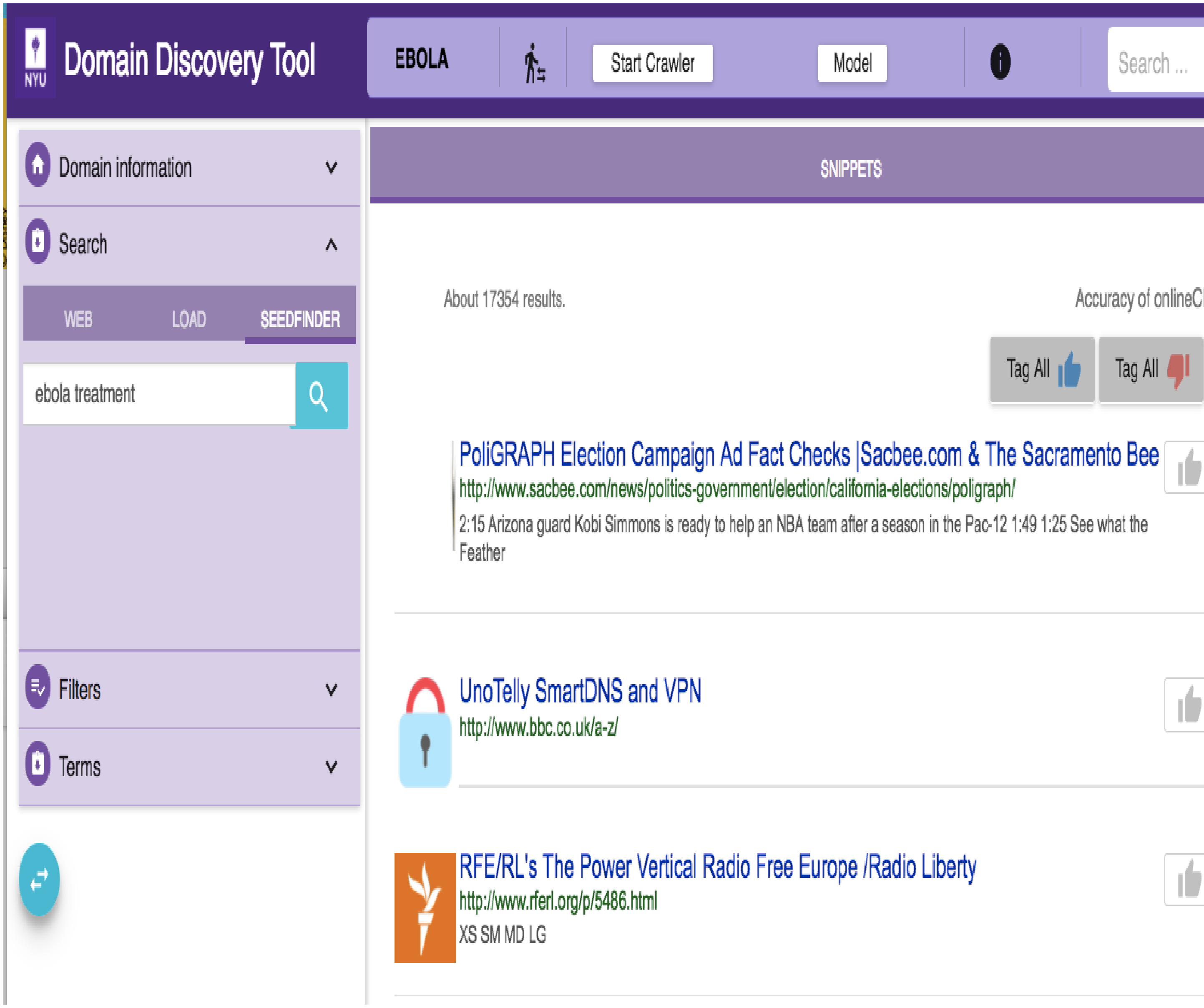
Enter one URL per line.

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 URLs should be entered one per line in the file. Download an example URLs list file for ebola domain HERE. The uploaded URLs are listed 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 annoated 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.



ebola treatment

About 17354 results.

Accuracy of onlineC

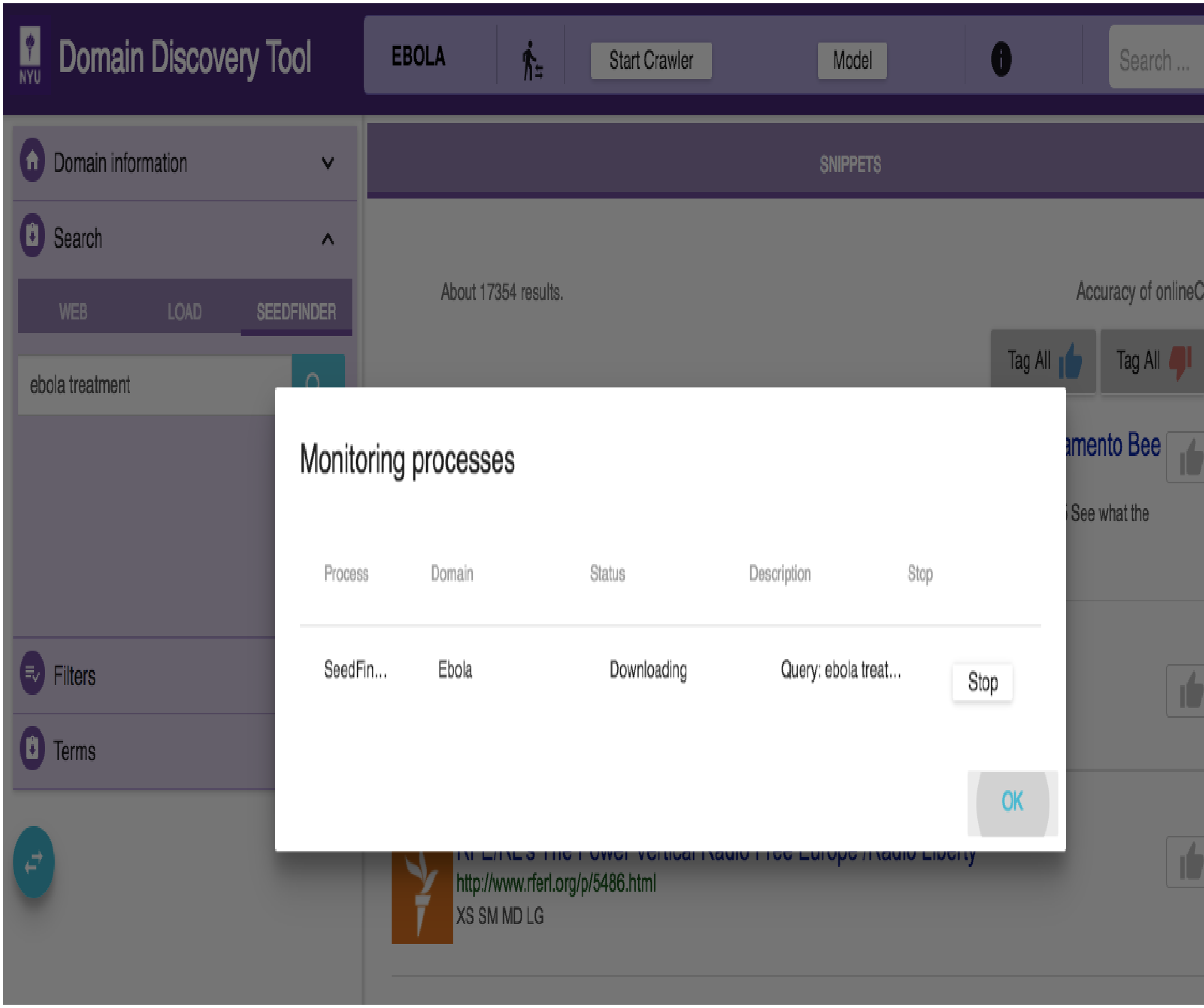
PoliGRAPH Election Campaign Ad Fact Checks |Sacbee.com & The Sacramento Bee
http://www.sacbee.com/news/politics-government/election/california-elections/poligraph/
2:15 Arizona guard Kobi Simmons is ready to help an NBA team after a season in the Pac-12 1:49 1:25 See what the Feather

UnoTelly SmartDNS and VPN
http://www.bbc.co.uk/a-z/

RFE/RL's The Power Vertical Radio Free Europe /Radio Liberty
http://www.rferl.org/p/5486.html
XS SM MD LG

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 be accessed by clicking on the ⓘ on the top as shown below:



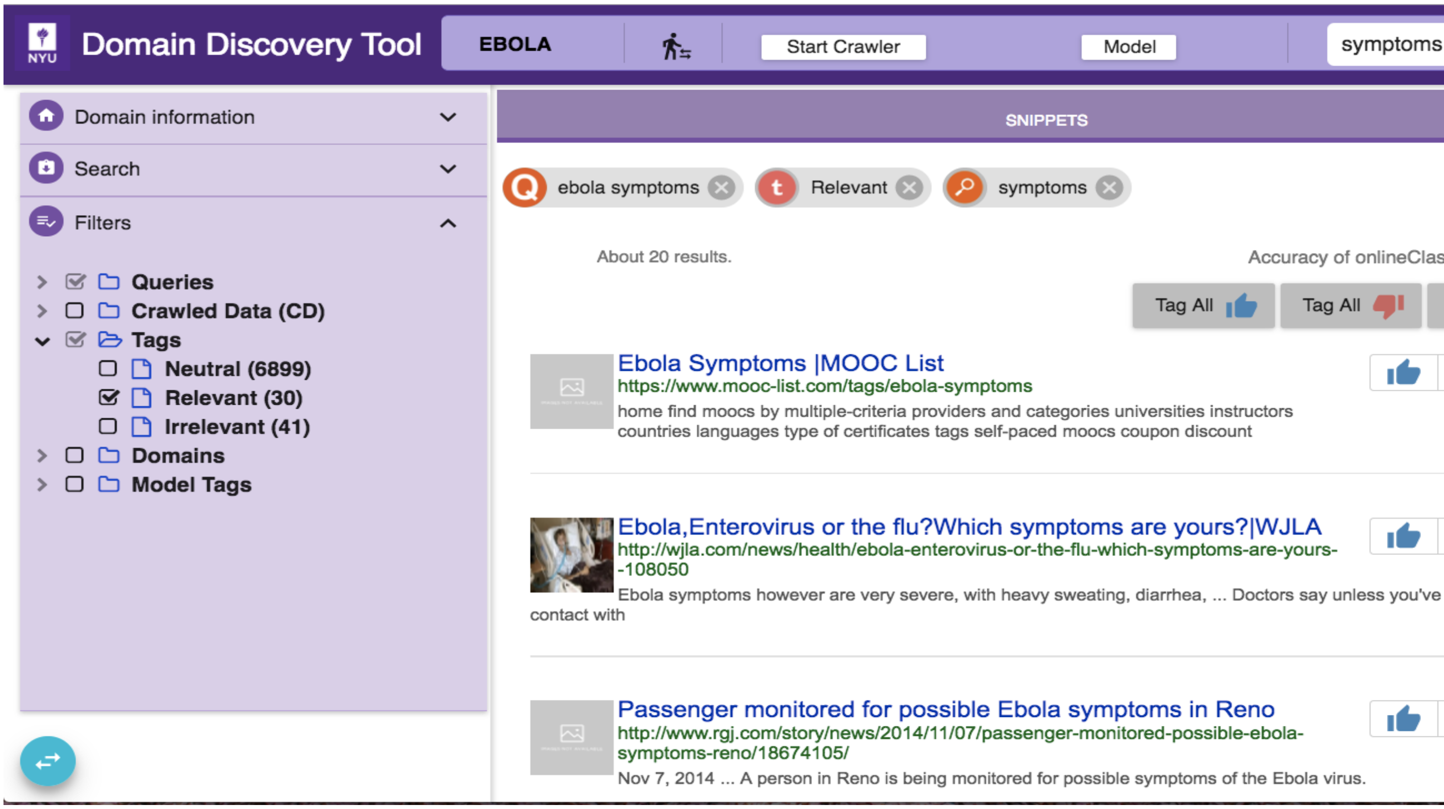
Monitoring processes

Process	Domain	Status	Description	Stop
SeedFin...	Ebola	Downloading	Query: ebola treat...	Stop

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

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.

1.2.3 Explore Data (Filters)



☑ Queries
□ Crawled Data (CD)
☑ Tags
□ Neutral (6899)
☑ Relevant (30)
□ Irrelevant (41)
□ Domains
□ Model Tags

About 20 results.

Accuracy of onlineClas

Ebola Symptoms |MOOC List
https://www.mooc-list.com/tags/ebola-symptoms
home find moocs by multiple-criteria providers and categories universities instructors countries languages type of certificates tags self-paced moocs coupon discount

Ebola,Enterovirus or the flu?Which symptoms are yours?|WJLA
http://wjla.com/news/health/ebola-enterovirus-or-the-flu-which-symptoms-are-yours--108050
Ebola symptoms however are very severe, with heavy sweating, diarrhea, ... Doctors say unless you've contact with

Passenger monitored for possible Ebola symptoms in Reno
http://www.rgj.com/story/news/2014/11/07/passenger-monitored-possible-ebola-symptoms-reno/18674105/
Nov 7, 2014 ... A person in Reno is being monitored for possible symptoms of the Ebola virus.

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, uploaded URLs and seedfinder queries made to date in the domain. You can select one or more of these queries to get pages for those specific queries.

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.

Tags

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

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.

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**.

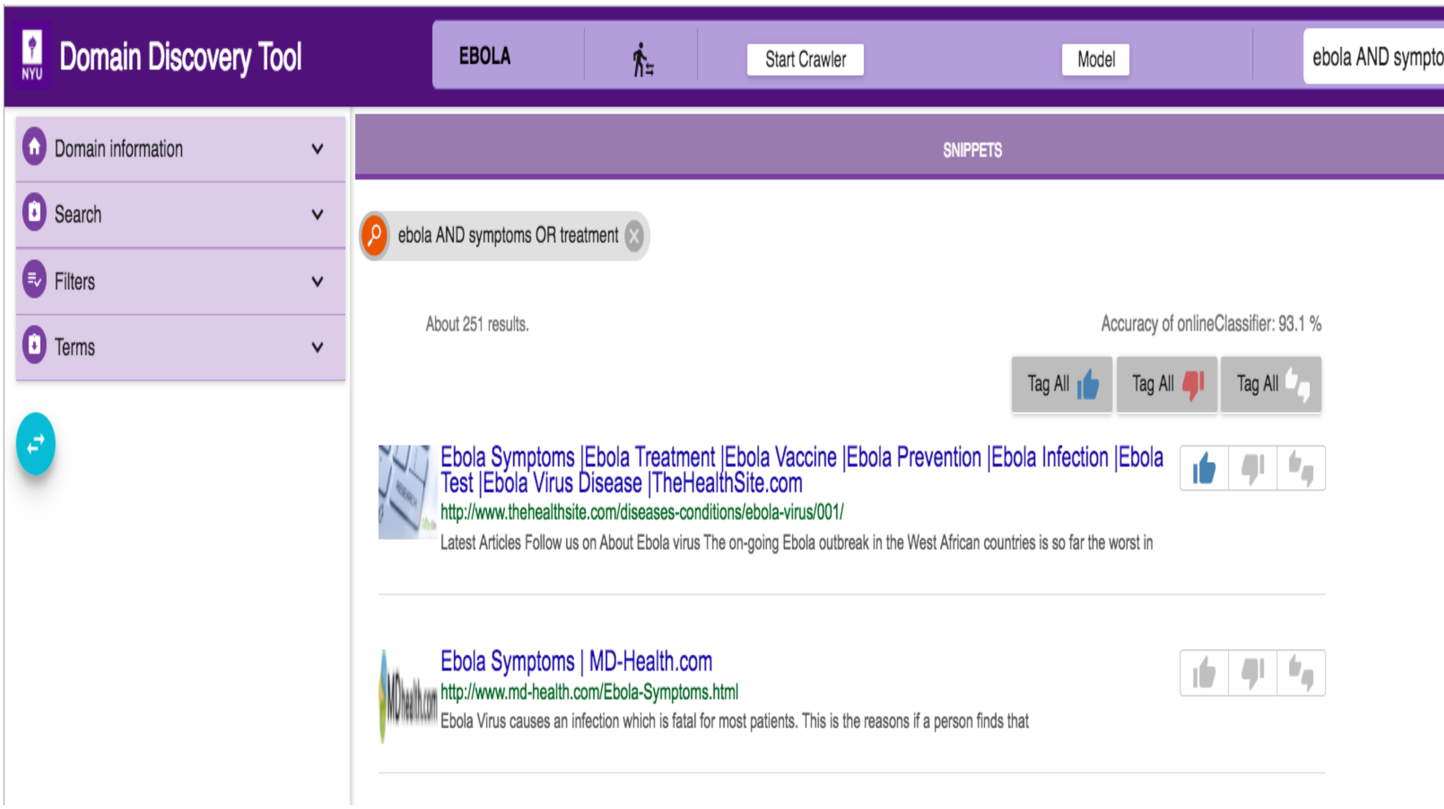
Model Tags

You can expand the **Model Tags** and click the **Upate 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.

Search



ebola AND symptoms OR treatment

About 251 results.

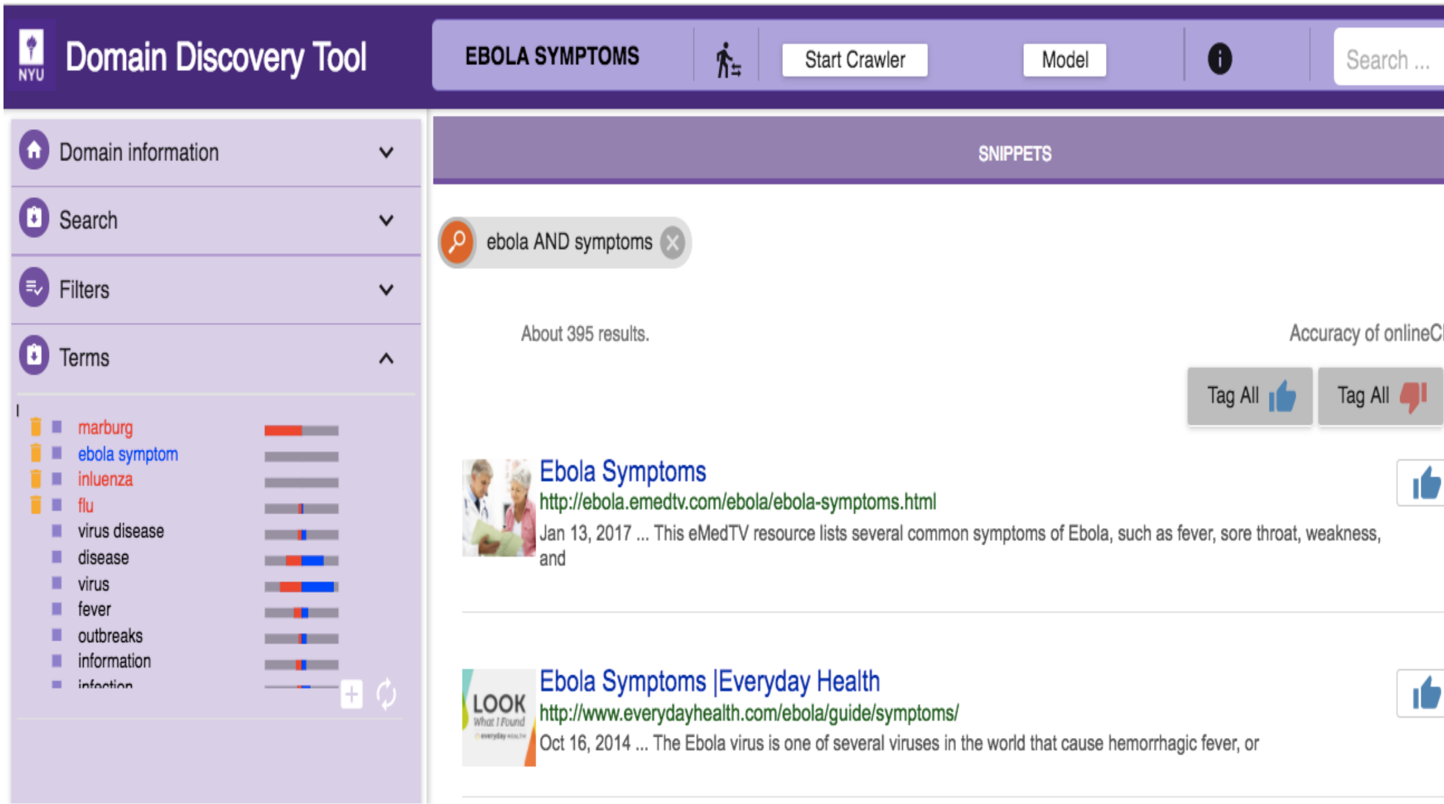
Accuracy of onlineClassifier: 93.1 %

Ebola Symptoms |Ebola Treatment |Ebola Vaccine |Ebola Prevention |Ebola Infection |Ebola Test |Ebola Virus Disease |TheHealthSite.com
http://www.thehealthsite.com/diseases-conditions/ebola-virus/001/
Latest Articles Follow us on About Ebola virus The on-going Ebola outbreak in the West African countries is so far the worst in

Ebola Symptoms | MD-Health.com
http://www.md-health.com/Ebola-Symptoms.html
Ebola Virus causes an infection which is fatal for most patients. This is the reasons if a person finds that

Search by keywords within the 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.

1.2.4 Extracted Terms Summary



Ebola Symptoms
http://ebola.emedtv.com/ebola/ebola-symptoms.html
Jan 13, 2017 ... This eMedTV resource lists several common symptoms of Ebola, such as fever, sore throat, weakness, and

Ebola Symptoms |Everyday Health
http://www.everydayhealth.com/ebola/guide/symptoms/
Oct 16, 2014 ... The Ebola virus is one of several viruses in the world that cause hemorrhagic fever, or

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.

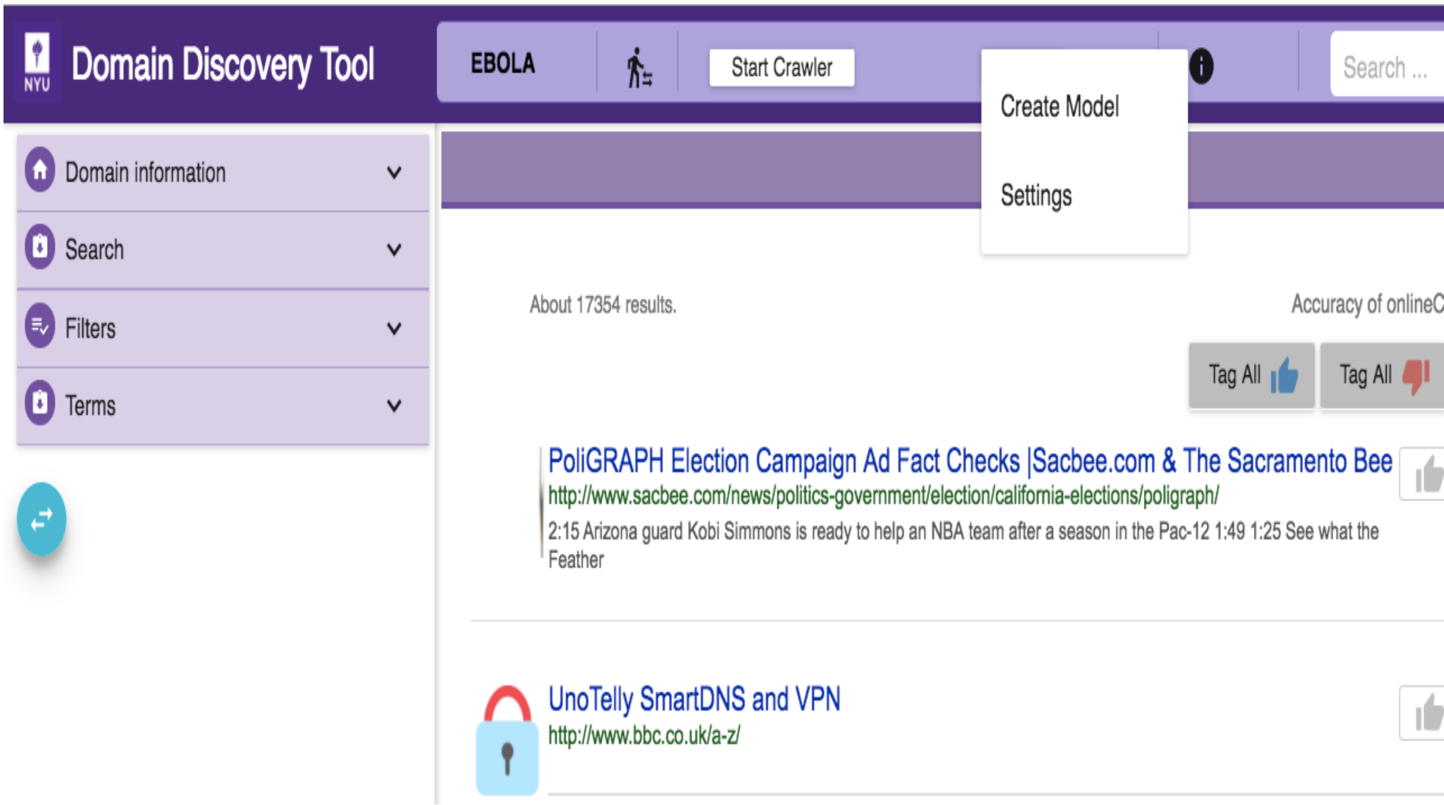
Custom relevant and irrelevant terms can be added by clicking the + button to boost the extraction of more relevant terms. These custom terms are distinguised by the delete icon before them which can be clicked to delete the custom term.

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.

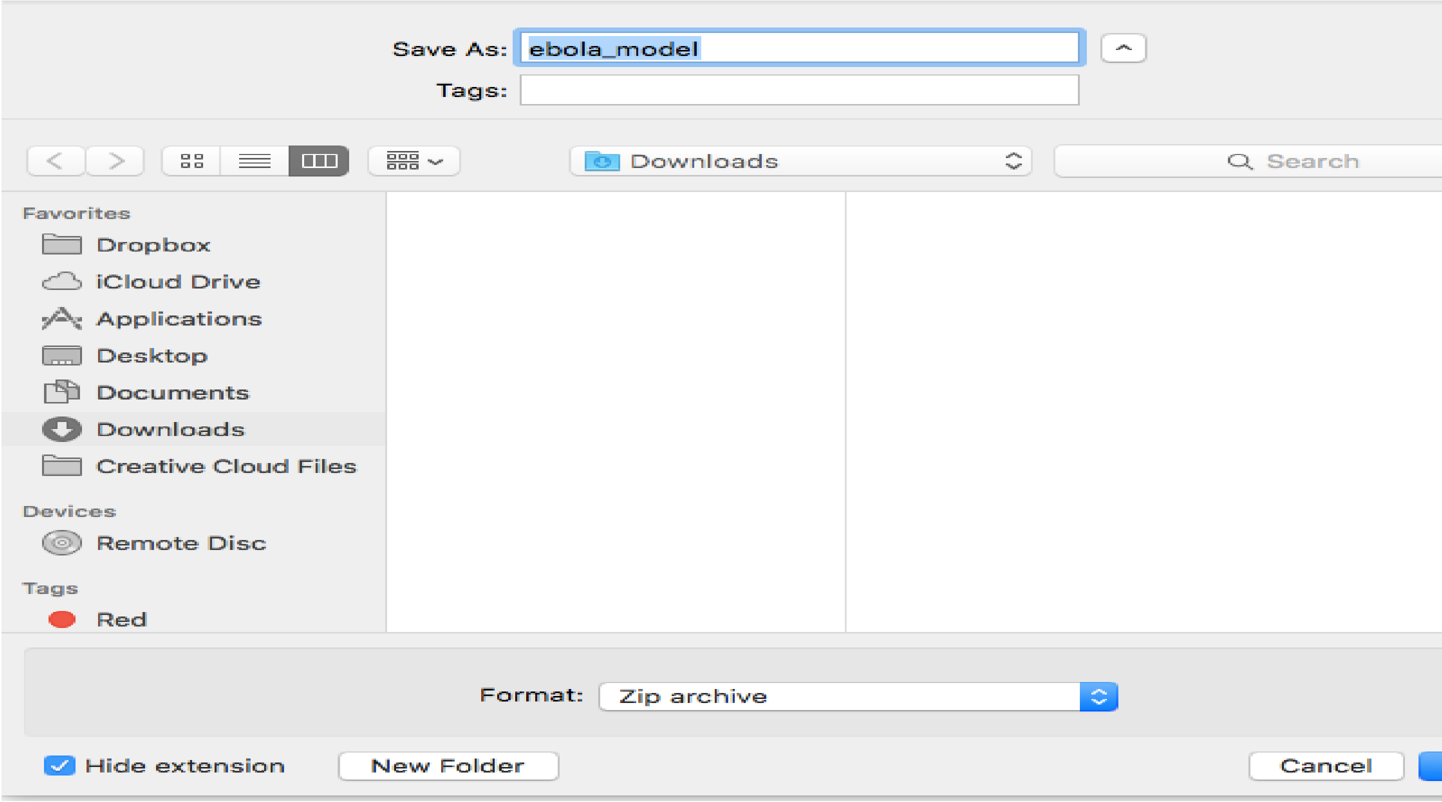
1.2.5 Create Model

DDT incrementally builds a model as the user annotates the retrieved pages. The accuracy of the domain model is displayed on the top right corner. It provides an indication of the model coverage of the domain and how it is influenced by annotations.

The domain model can be exported by clicking on the **Model** button on the top (this button will be dsiabled when there are no sufficient annotations to build the model and the model **Accuracy of onlineClassifier: 0 %**). This will show a drop down as shown in figure below:


Domain Discovery Tool
EBOLA
Start Crawler
Create Model
Settings
Search ...
Domain information
Search
Filters
Terms
About 17354 results.
Accuracy of onlineC
Tag All
Tag All
PoliGRAPH Election Campaign Ad Fact Checks |Sacbee.com & The Sacramento Bee
http://www.sacbee.com/news/politics-government/election/california-elections/poligraph/
2:15 Arizona guard Kobi Simmons is ready to help an NBA team after a season in the Pac-12 1:49 1:25 See what the Feather
UnoTelly SmartDNS and VPN
http://www.bbc.co.uk/a-z/

Click on **Create Model** to export the model. This should bring up a file explorer pop-up (makes sure you enable pop-up on your browser) as shown below. Save the compressed model file.



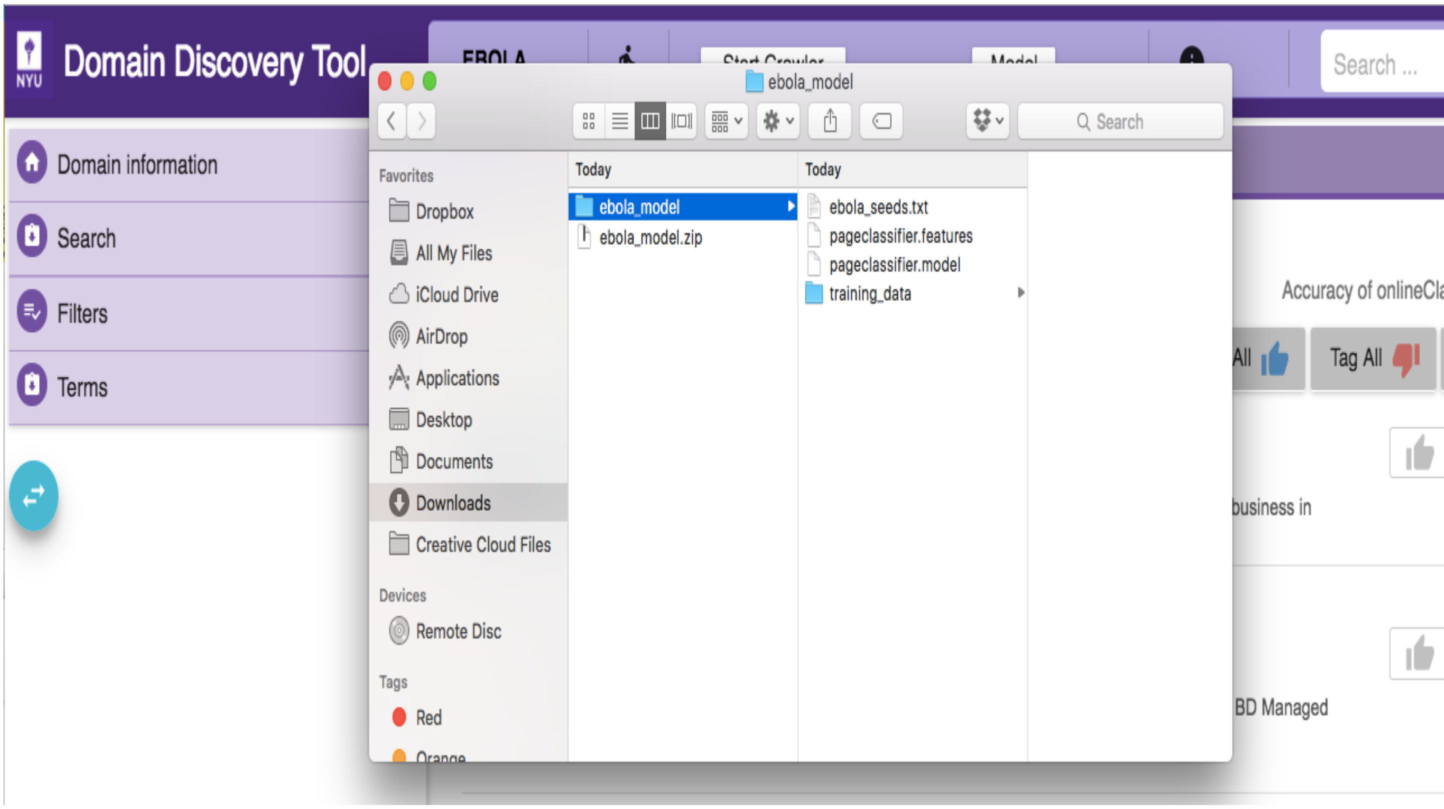
Save As: ebola_model

Tags:

Format: Zip archive

☑ Hide extension

This saved model file contains the ACHE classifier model, the training data for the model and the initial seed list required for focused crawling as shown in figure below:



Annotation

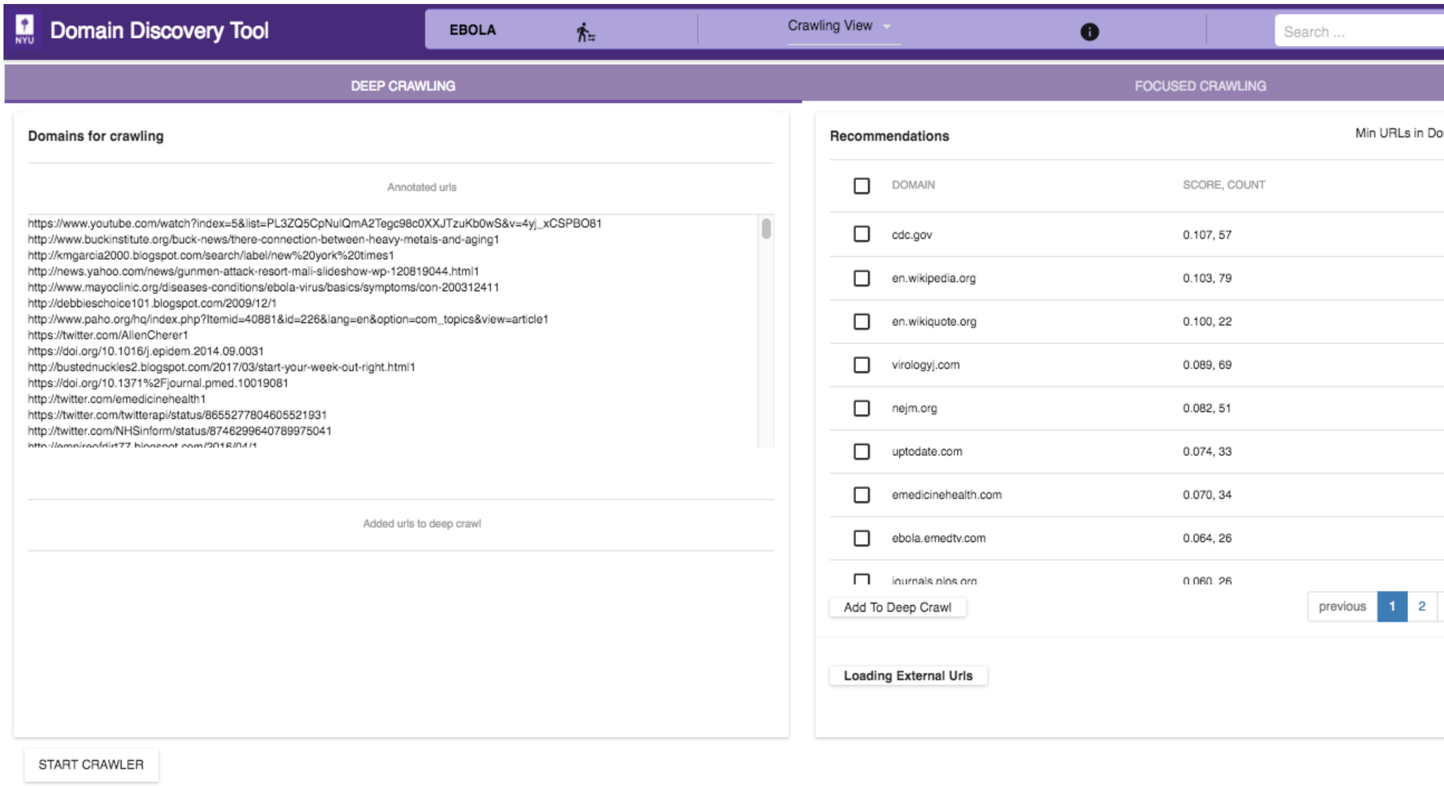
Currently, pages can be annotated as Relevant, Irrelevant or Neutral using the Tag All 👍 Tag All 👎 Tag All 👍👎 buttons respectively to tag all pages in the current view. 👍 👎 👍👎 buttons can be used to tag individual pages. Annotations are used to build the domain model.

Note:

- At least 10 pages each of relevant and irrelevant pages should be annotated to build the model. The more the annotations, hence the better coverage of the domain, the better the domain model.
- Ensure that the relevant and irrelevant page annotations are balanced for a better model.

1.2.6 Run Crawler

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



Domains for crawling

Annotated urls

```
https://www.youtube.com/watch?index=5&list=PL3ZQ5CpNulQmA2Tegc98c0XXJTzuKb0wS&v=4yj_xCSPBO81
http://www.buckinstitute.org/buck-news/there-connection-between-heavy-metals-and-aging1
http://kmgarcia2000.blogspot.com/search/label/new%20york%20times1
http://news.yahoo.com/news/gunmen-attack-resort-mali-slideshow-wp-120819044.html1
http://www.mayoclinic.org/diseases-conditions/ebola-virus/basics/symptoms/con-200312411
http://debbieschoice101.blogspot.com/2009/12/1
http://www.paho.org/hq/index.php?Itemid=40881&id=226&lang=en&option=com_topics&view=article1
https://twitter.com/AllenCherer1
https://doi.org/10.1016/j.epidem.2014.09.0031
http://bustednuckles2.blogspot.com/2017/03/start-your-week-out-right.html1
https://doi.org/10.1371%2Fjournal.pmed.10019081
http://twitter.com/emedicinehealth1
https://twitter.com/twitterapi/status/8655277804605521931
http://twitter.com/NHSinform/status/8746299640789975041
http://empireofdirt77.blogspot.com/2016/04/1
```

Added urls to deep crawl

START CRAWLER

Recommendations

Min URLs in Do

☐	DOMAIN	SCORE, COUNT
☐	cdc.gov	0.107, 57
☐	en.wikipedia.org	0.103, 79
☐	en.wikiquote.org	0.100, 22
☐	virologyj.com	0.089, 69
☐	nejm.org	0.082, 51
☐	uptodate.com	0.074, 33
☐	emedicinehealth.com	0.070, 34
☐	ebola.emedtv.com	0.064, 26
☐	journals.plos.org	0.060, 26

Add To Deep Crawl

previous 1 2

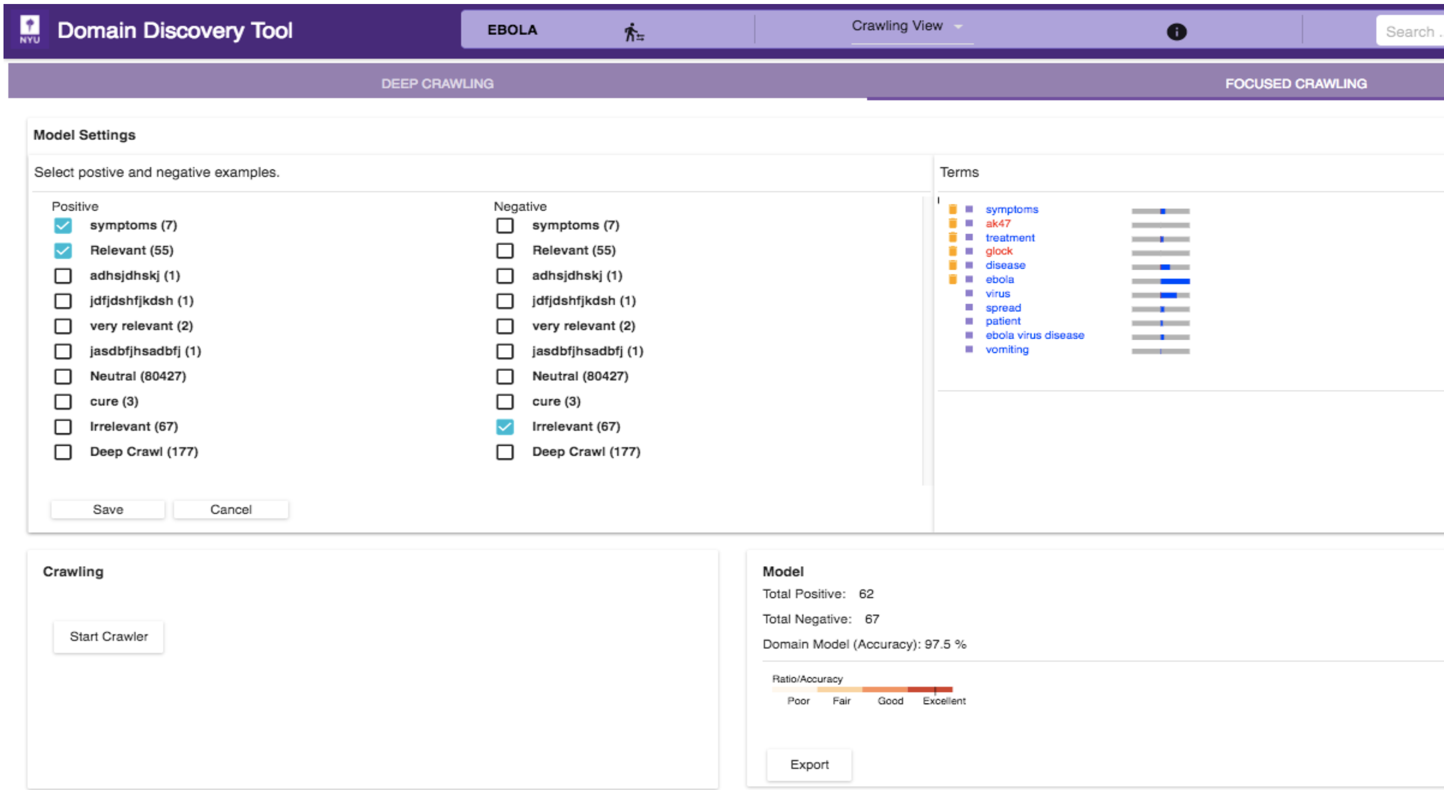
Loading External Urls

The crawler view support a deep and focused crawl. The figure above shows the Deep Crawl View. The list on the left shows all pages annoated 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 have. Otherwise the domains are reocommended by the number of pages they contain.

The ACHE 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.

The figure below shows the Focused Crawler View:



Model Settings

Select postive and negative examples.

Positive	Negative
☑ symptoms (7)	□ symptoms (7)
☑ Relevant (55)	□ Relevant (55)
□ adhsjdhskj (1)	□ adhsjdhskj (1)
□ jdfjdshfjkdsh (1)	□ jdfjdshfjkdsh (1)
□ very relevant (2)	□ very relevant (2)
□ jasdbfjhsadbfj (1)	□ jasdbfjhsadbfj (1)
□ Neutral (80427)	□ Neutral (80427)
□ cure (3)	□ cure (3)
□ Irrelevant (67)	☑ Irrelevant (67)
□ Deep Crawl (177)	□ Deep Crawl (177)

Model

Total Positive: 62

Total Negative: 67

Domain Model (Accuracy): 97.5 %

First, 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. If there are 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 not annotated terms are available then the top 50 terms are used to build a regular expression model.

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 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).

1.4 Contact

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

CHAPTER 2

Links

- GitHub repository

CHAPTER 3

Indices and tables

- genindex
- modindex
- search