TREC Dynamic Domain Track
2015 Guidelines

The goal of the dynamic domain (DD) track is to support research in dynamic, exploratory search of complex information domains. DD systems receive relevance feedback as they explore a space of subtopics within the collection in order to satisfy a user's information need.

1. Participation in TREC

In order to take part in the DD track, you need to be a registered participant of TREC. The TREC Call for Participation, at http://trec.nist.gov/pubs/call2015.html, includes instructions on how to register. **You must register before May 1, 2015 in order to participate.**

The datasets and relevance judgments will be made generally available to non-participants after the TREC 2015 cycle, in February 2016. So register to participate if you want early access.

2. Domains and datasets for 2015

For 2015 there are three domains, each with different data:

Illicit Goods: this data is related to how illicit and counterfeit goods such as fake viagra are made, advertised, and sold on the Internet. The dataset comprises 5,000,000 million posts from underground hacking forums, arranged into threads.

Ebola: this data is related to the Ebola outbreak in Africa in 2014-2015. The dataset comprises 30 million tweets relating to the outbreak, and 500,000 web pages from sites hosted in the affected countries and designed to provide information to citizens and aid workers on the ground.

Local Politics: this data is related to regional politics in the Pacific Northwest and the small-town politicians and personalities that work it. The dataset comprises 1,000,000 web news items from the TREC 2014 KBA Stream Corpus.

All the datasets are formatted using the streamcorpus format described at http://streamcorpus.org. That site includes documentation and tools for working with the data. For the DD track, you will be working with the datasets as static corpora, meaning that for purposes of the task you can index and search the datasets without processing them in any particular stream order.

DD systems are expected to be able to use all three domain datasets simultaneously, and the same system must be used for all three domains.

3. Task description

Within each domain, there will be 25-50 *topics* that represent user search needs. An example topic for the Ebola domain might be titled, "foreign aid health workers." The topics will be developed by the NIST assessors. Each topic will have a number of subtopic interests that make up the various pieces that the user wants to know. Following the above example, subtopics might include "how many foreign health workers are in Freetown" and "what clinics are primarily staffed by foreign health workers".

DD systems will receive an initial query for each topic, where the query is two to four words and additionally indicates the domain by a number 1, 2, or 3. In response to that query, systems may return up to five documents to the user. The simulated user will respond by indicating which of the retrieved documents are relevant to their interests in the topic, and to which subtopic the document is relevant to. Additionally, the simulated user will identify passages from the relevant documents and assign the passages to the subtopics with an integer relevance rating of 1, 2, or 3. (Guidelines on the relevance rating are in preparation.) The system may then return another five documents for more feedback. Systems should stop when they believe they have covered all the user's subtopics sufficiently. The subtopics are not known the system in advance; systems must discover the subtopics from the user’s responses.

The system interactions with the user will be simulated using a jig that the track coordinators will provide to you. This jig will be implemented in Python, will run on Windows, Mac OS, and Linux, and will provide a local shell-based API for your DD system to operate against. The API jig will include all relevance information for the task, but your system may only interact with this information through the jig. The jig produces an output file documenting the interaction that you send to NIST as your formal track participation.

The test harness is under development here:
https://github.com/trec-dd/trec-dd-simulation-harness

and a two-minute video of interacting with the initial version (2015-01-16) is available here:
https://www.dropbox.com/s/d2h666pddjqxnsb/2015-01-16-TREC-DD-simulation-harness-demo.mov?dl=0

4. Task measures

The primary measures will be Cube Test, μ-ERR, and session NDCG. Reference scripts will be provided by the track coordinators. We will also likely report other diagnostic measures such as basic precision and recall.

The CubeTest proposes that a search process can be understood by analogy to water filling a compartmentalized task cube. The time-sensitivity is represented by the fact that searchers want the multi-faceted components of the cube filled as quickly as possible.
Reference: Jiyun Luo, Christopher Wing, Hui Yang, Marti Hearst. The Water Filling Model and The Cube Test: Multi-Dimensional Evaluation for Professional Search. CIKM 2013.
http://cs-sys-1.uis.georgetown.edu/~xd47/InfoSense/publication/cikm2013.pdf

u-ERR is an extension to the cascade model of ERR by defining an outer loop in which the user changes the entity profile each time they find a useful piece of new information.
Reference: Oliver Chappelle, Don Metzler, Y Zhang, ACM 2009 Expected reciprocal rank for graded relevance. http://dl.acm.org/citation.cfm?id=1646033

session-nDCG generalizes the nDCG scoring function to multi-query session evaluation.
Reference: Evangelos Kanoulas, Ben Carterette, Paul D. Clough, Mark Sanderson. Evaluating Multi-Query Sessions. SIGIR 2011.
http://dl.acm.org/citation.cfm?id=2009916.2010056

5. Run Format

In TREC, a "run" is the output of a search system over all topics. Participating groups typically submit more than one run corresponding to different parameter settings or algorithmic choices. The maximum number of runs allowed for DD 2015 is five from each team.

The jig outputs a correctly-formatted run file, but for reference (or if you are performing manual searches), the format is as follows. The run is a plain text line-oriented format:

```
<topic> <step> <document-id> <score> <runtag>
```

where <topic> is the topic number, <step> is the step number, starting from 1, <document-id> is a document identifier from the datasets, <score> is a floating-point number indicating your systems retrieval score for that document, and <runtag> is an identifier such as "NISTrun1" that identifies the team and the specific run. Whitespace separates each field.

6. Requirements

Participants are expected to submit at least one run by the deadline.

Runs may be fully automatic, or manual. Manual indicates intervention by a person at any stage of the retrieval. We welcome unusual approaches to the task including human-in-the-loop searching, as this helps us set upper performance bounds.

7. Timeline

Domain collections available: May 1
Topics available to participants: May 1
Runs due from participants: Aug 1
Evaluation results returned: Sep 15
TREC 2015 notebook paper deadline: October
TREC 2015 conference: November 17-20, 2015