500k User Session Collection

This collection is distributed for NON-COMMERCIAL RESEARCH USE ONLY.
Any application of this collection for commercial purposes is STRICTLY PROHIBITED.

Brief description:

This collection consists of ~20M web queries collected from ~650k users over three months.
The data is sorted by anonymous user ID and sequentially arranged.

The goal of this collection is to provide real query log data that is based on real users. It could be used for personalization, query reformulation or other types of search research.

The data set includes {AnonID, Query, QueryTime, ItemRank, ClickURL}.

```
        AnonID - an anonymous user ID number.
        Query  - the query issued by the user, case shifted with
                 most punctuation removed.
        QueryTime - the time at which the query was submitted for search.
        ItemRank  - if the user clicked on a search result, the rank of the
                    item on which they clicked is listed.
        ClickURL  - if the user clicked on a search result, the domain portion of
                    the URL in the clicked result is listed.
```

Each line in the data represents one of two types of events:

1. A query that was NOT followed by the user clicking on a result item.
2. A click through on an item in the result list returned from a query.

In the first case (query only) there is data in only the first three columns/fields -- namely AnonID, Query, and QueryTime (see above).
In the second case (click through), there is data in all five columns. For click through events, the query that preceded the click through is included. Note that if a user clicked on more than one result in the list returned from a single query, there will be TWO lines in the data to represent the two events. Also note that if the user requested the next "page" or results for some query, this appears as a subsequent identical query with a later time stamp.

CAVEAT EMPTOR -- SEXUALLY EXPLICIT DATA! Please be aware that these queries are not filtered to remove any content. Pornography is prevalent on the Web and unfiltered search engine logs contain queries by users who are looking for pornographic material. There are queries in this collection that use SEXUALLY EXPLICIT LANGUAGE. This collection of data is intended for use by mature adults who are not easily offended by the use of pornographic search terms. If you are offended by sexually explicit language you should not read through this data. Also be aware that in some states it may be illegal to expose a minor to this data. Please understand that the data represents REAL WORLD USERS, un-edited and randomly sampled, and that AOL is not the author of this data.

Basic Collection Statistics

```
Dates:
  01 March, 2006 - 31 May, 2006

Normalized queries:
  36,389,567 lines of data
  21,011,340 instances of new queries (w/ or w/o click-through)
   7,887,022 requests for "next page" of results
  19,442,629 user click-through events
  16,946,938 queries w/o user click-through
  10,154,742 unique (normalized) queries
     657,426 unique user ID's
```

Please reference the following publication when using this collection:

G. Pass, A. Chowdhury, C. Torgeson, "A Picture of Search" The First
International Conference on Scalable Information Systems, Hong Kong, June,
2006.

Copyright (2006) AOL