2010 Crowdsourced Web Relevance Judgments Data

==

Source: the NIST TREC Relevance Feedback Track 2010:
http://trec.nist.gov/data/relevance.feedback10.html

Data: https://www.ischool.utexas.edu/~ml/data/trec-rf10-crowd.tgz

Release Date: April 25, 2013

Contributors: Chris Buckley, Matthew Lease, Mark D. Smucker, Hyun Joon Jung, and Catherine Grady

Contact: Matthew Lease, <*myinitials*>@ischool.utexas.edu

Citation: the official paper to cite when using this dataset is

```
@inproceedings{Buckley10-notebook,
  author={Chris Buckley and Matthew Lease and Mark D. Smucker},
  title={{Overview of the TREC 2010 Relevance Feedback Track (Notebook)}},
  booktitle={{The Nineteenth Text Retrieval Conference (TREC) Notebook}},
  institute = {{National Institute of Standards and Technology (NIST)}},
  year={2010}
}
```

This paper is available from the authors, upon request.

==

Data Description

Mechanical Turk workers judged relevance of English Web pages from the ClueWeb09 collection (http://lemurproject.org/clueweb09/) for English search queries drawn from the TREC 2009 Million Query track (http://ir.cis.udel.edu/million).

Relevance was judged on a ternary scale: highly relevant, relevant, and non-relevant. A fourth judgment option indicated a broken link which could not be judged. For quality assurance, we also intentionally included URLs for non-existent Web pages (broken links) to be judged, which were expected to be reported by this 4th option.

There are 20,232 total (topic,document) examples (noisily) judged by 766 workers, who produced a total of 98,453 judgments. 3277 of the examples have prior "gold" labels by NIST.

Worker IDs have been anonymized.

==

Data Format

```
topicID workerID        docID                                   gold    label
20002           w1              clueweb09-en0000-66-24091-1     0
20002           w1              clueweb09-en0001-31-15410-1     0
20002           w1              clueweb09-en0000-05-22942-1     0
20002           w1              clueweb09-en0000-05-22943-1     0
...
```

Gold labels were produced by NIST. Crowd judgments are in the last column.

Judgment categories:

 2: highly relevant
 1: relevant
 0: non-relevant
-1: unknown (no gold label)
-2: broken link

==

Reference R code for computing descriptive statistics

```
> data <- read.delim('trec-rf10-data.txt',fill=TRUE,header=TRUE,sep="\t")

> names(data)
[1] "topicID"  "workerID" "docID"    "gold"     "label"

> nrow(data)
[1] 98453

> unique_examples <- unique(data[c("topicID","docID","gold")])
> nrow(unique_examples)
[1] 20232

> table(unique_examples$gold)
   -2    -1     0     1     2
 1183 15772  1501   863   913

> length(which(unique_examples$gold>-1))
[1] 3277
```

```
> length(unique(data$workerID))
[1] 766
```

==

A different, simplified version of this dataset was later released August 19, 2011, for use in a different NIST TREC Track:

The NIST TREC 2011 Crowdsourcing Track, Task 2
https://sites.google.com/site/treccrowd/home

data: https://sites.google.com/site/treccrowd/home/trec11-cs-task2-test.tar.bz2

The simplified version differed in several respects:
* documentIDs were anonymized
* broken links were removed
* highly relevant and relevant categories were conflated to produce binary judgments and ground truth

==

NAACL AMT 2010 Earlier, Smaller dataset

http://www.ischool.utexas.edu/~ml/data/naacl-amt-2010.zip

The companion paper and official citation for this earlier dataset is:

```
@inproceedings{Grady10,
  author = {Grady, Catherine  and  Lease, Matthew},
  title = {Crowdsourcing Document Relevance Assessment with Mechanical Turk},
  booktitle = {Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and
Language Data with Amazon's Mechanical Turk},
  month = {June},
  year = {2010},
  address = {Los Angeles},
  publisher = {Association for Computational Linguistics},
  pages = {172--179},
  url = {http://www.aclweb.org/anthology/W10-0727}
}
```

This paper described the method on Mechanical Turk data collection that was largely adopted for producing the larger dataset for the TREC RF 2010 track.

===

Additional References

The papers below further discuss methods of statistical quality assurance for this dataset:

```
@inproceedings{Jung12-hcomp,
  author = {Hyun Joon Jung and Matthew Lease},
  title = {{Improving Quality of Crowdsourced Labels via Probabilistic Matrix Factorization}},
  booktitle = {{Proceedings of the 4th Human Computation Workshop (HCOMP) at AAAI}},
  year = {2012},
  url = {http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/download/5258/5609}
}

@inproceedings{Jung12-sigir,
  author = {Hyun Joon Jung and Matthew Lease},
  title = {{Inferring Missing Relevance Judgments from Crowd Workers via Probabilistic Matrix Factorization}},
  booktitle = {{Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval}},
  year = {2012},
  url = {https://www.ischool.utexas.edu/~ml/papers/jung-sigir12.pdf}
}

@inproceedings{Jung11-hcomp,
  author = {Hyun Joon Jung and Matthew Lease},
  title = {{Improving Consensus Accuracy via Z-score and Weighted Voting}},
  booktitle = {{Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI}},
  year = {2011},
  pages = {88--90},
  url = {http://www.ischool.utexas.edu/~ml/papers/jung-hcomp11.pdf}
}

@inproceedings{Tang11-cir,
  author = {Wei Tang and Matthew Lease},
  title = {Semi-Supervised Consensus Labeling for Crowdsourcing},
  booktitle = {{ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR)}},
  year = {2011},
  url = {http://www.ischool.utexas.edu/~ml/papers/tang-cir11.pdf}
}
```