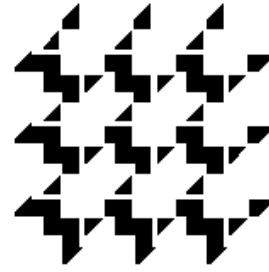

DIMACS

*Center for Discrete Mathematics & Theoretical Computer Science
Founded as a National Science Foundation Science and
Technology Center*



Adaptive Filtering Software: Rocchio, Centroid, and kNN with inverted file heuristic.

David D. Lewis and Vladimir Menkov

Last update: November 1, 2005

[New in this version](#)

[Overview](#)

[Installation Guide](#)

[User Guide](#)

[Future Work](#)

[Appendix 1: Lemur Guide](#)

[Appendix 2: Additional Data Sets](#)

[Acknowledgments](#)

[References](#)

[Legal Notice](#)

[Feedback](#)

New in this version

This version adds new capabilities to Index2SVM program for converting Lemur Index data to vector format, and allows classification with multiple linear models (pre-trained). The details and in appropriate sections of the User Guide.

Acknowledgments

The work was partially supported under funds provided by the KD-D group for a project at DIMACS on Monitoring Message Streams, funded through National Science Foundation grant EIA-0087022 to Rutgers University. The NSF also partially supported the work through ITR grant DMS-0113236.

Source code used:

- The Lemur Toolkit for Language Modeling and Information Retrieval: <http://www-2.cs.cmu.edu/~lemur/>
- Numerical Recipes: <http://www.library.cornell.edu/nr/bookcpdf.html>

References

Project Reports at the main MMS Working Group page (<http://www.stat.rutgers.edu/~madigan/mms/>):

- Andrei Anghelescu, David D. Lewis, Vladimir Menkov, and Paul Kantor. "Training a Rocchio classifier for adaptive filtering". (TREC 2002).
- David D. Lewis and Vladimir Menkov. "A Closer Look at Nearest Neighbor Text Classification". May 2003.
- David D. Lewis, S. Muthukrishnan, R. Ostrovsky, M. Strauss. "Towards Fast, Effective Nearest Neighbor Text Classification". November 2002. (A number of randomized projection methods referred to in that paper are implemented by this software, but as they are not particularly effective, they are not documented here).

Internal reports on the results of kNN experiments (<http://aplab.rutgers.edu/vmenkov/>):

- [kNN runs on RCV1-v2](#) (January 2003)
- [kNN runs on WIPO-alpha](#) (Summer 2003)
- [kNN with Inverted List Heuristics \(ILH\)](#) (Summer-Fall 2003)
- [On the optimal neighborhood size k for individual topics](#) (September 2003)

Information on the RCV1-v2 corpus (LYRL):

- David D. Lewis, Yiming Yang, Tony Rose, and Fan Li. RCV1: "A New Benchmark Collection for Text Categorization Research". Journal of Machine Learning Research, 2004. It is expected to appear in early 2004 on the JMLR web site, <http://www.jmlr.org/>. A draft version of the article is available as a local copy in [draft7-pdflatex.pdf](#).
- Data at David Lewis' site:
<http://www.daviddlewis.com/resources/testcollections/>.
- T. Rose, M. Stevenson, M. Whitehead: "The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources".
http://about.reuters.com/researchandstandards/corpus/LREC_camera_ready.pdf

Information on the WIPO-alpha corpus:

- WIPO site: <http://www.wipo.int/ibis/datasets/index.html>
- Our internal report on corpus preparation:
<http://aplab.rutgers.edu/vmenkov/wipo/index.html>

TREC:

- [Text REtrieval Conference \(TREC\)](#)

Legal Notice

Software is free for non-commercial use.

Software is provided as is, without any guarantee or warranty. DIMACS or authors are not responsible for implications from the use of this software.

Feedback

Questions, comments, bug reports welcome by email: vmenkov@aplab.rutgers.edu