2016-06

# Data mining of extremely large ad hoc data sets to produce inverted indices

Coudray, Aaron D.

Monterey, California: Naval Postgraduate School


# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

## READ ME: SUPPLEMENTAL TO THESIS

**DATA MINING OF EXTREMELY LARGE AD HOC DATA SETS TO PRODUCE INVERTED INDICES**

by

Aaron D. Coudray

June 2016

Thesis Advisor: Frank Kragh
Co-Advisor: Jim Scrofani

**Supplemental to this thesis are large files of computer code. They were provided to the Dudley Knox Library via CD. The Supplemental files are for unlimited public release. See pp. 2–3 of this READ ME for a complete list of files.**

| REPORT DOCUMENTATION PAGE | | | *Form Approved OMB No. 0704-0188* |
|---|---|---|---|
| Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503. | | | |
| **1. AGENCY USE ONLY** *(Leave blank)* | **2. REPORT DATE** June 2016 | **3. REPORT TYPE AND DATES COVERED** Master's Thesis | |
| **4. TITLE AND SUBTITLE** DATA MINING OF EXTREMELY LARGE AD HOC DATA SETS TO PRODUCE INVERTED INDICES | | | **5. FUNDING NUMBERS** |
| **6. AUTHOR(S)** Aaron D. Coudray | | | |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)** Naval Postgraduate School Monterey, CA 93943-5000 | | | **8. PERFORMING ORGANIZATION REPORT NUMBER** |
| **9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)** N/A | | | **10. SPONSORING / MONITORING AGENCY REPORT NUMBER** |
| **11. SUPPLEMENTARY NOTES** The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____. | | | |
| **12a. DISTRIBUTION / AVAILABILITY STATEMENT** Approved for public release; distribution is unlimited | | | **12b. DISTRIBUTION CODE** |
| **13. ABSTRACT (maximum 200 words)** The purpose of this study is to leverage existing Internet-sized ad hoc data sets by creating an inverted index that will enable a robust search capability. In particular, this study is focused on the Common Crawl web corpus. This involves exploring the tools and techniques necessary to effectively traverse this data set, as well as producing the tools to create an inverted index relationship between the terms and websites found within web archive files. The primary tools utilized in this process are Apache Hadoop, Apache MapReduce, Amazon Web Services, and Java. Additionally, methods to enhance this relationship with other information of interest are investigated in this thesis. Specifically, an index was developed that contains the added component of term relative location. This inverted index relationship is an essential component of—and the first step in—creating a robust search capability for a very large ad hoc data set. | | | |
| **14. SUBJECT TERMS** big data, Common Crawl, Hadoop, inverted index, inverted indices, Java, MapReduce | | | **15. NUMBER OF PAGES** 167 |
| | | | **16. PRICE CODE** |
| **17. SECURITY CLASSIFICATION OF REPORT** Unclassified | **18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified | **19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified | **20. LIMITATION OF ABSTRACT** UU |

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

# APPENDIX K. SUPPLEMENTAL

Included in this section is a list of all supplemental source code and project files that are available upon request. To obtain a copy of these files, contact the Dudley Knox Library located on the campus of the Naval Postgraduate School in Monterey, California. A copy of this information is included in the project files.

## A. ORIGINAL CODE FILES

- basicInvertedIndexCount.java
- basicInvertedLocation.java
- basicWordCount.java
- termLocationCustomWritable.java
- WARCCompressURL.java
- WARCCompressURLMap.java
- WARCExtractContentType.java
- WARCExtractContentTypeMap.java
- WARCIndexLocationCustom.java
- WARCIndexLocationCustomMap.java
- WARCWordCount.java
- WARCWordCountMap.java
- WETIndexCount.java
- WETIndexCountMap.java
- WETIndexLocation.java
- WETIndexLocationMap.java
- WETIndexLocationCustom.java
- WETIndexLocationCustomMap.java
- WETIndexRecordLocation.java

- WETIndexRecordLocationMap.java
- WETWordCount.java
- WETWordCountMap.java
- pom.xml

## B. CODE FILES FROM ANOTHER SOURCE

The following files were authored by mathijs.kattenberg@surfsara.nl and obtained from github.com.

- WarcInputFormat.java
- WarcIOConstants.java
- WarcRecordReader.java
- WarcSequenceFileInputFormat.java
- WarcSequenceFileRecordReader.java