Scientific Publications Mining with Awk and Swift

Ketan C Maheshwari
CADES, ORNL

SMOKY MOUNTAIN Computational Sciences and Engineering Conference | .:DATA:. CHALLENGE

August 29, 2018

`github.com/ketancmaheshwari/SMC18`

Overview

- Data mining millions of scientific publication records
- Useful findings such as influential authors, geographic locations of research, citation networks, and topic trends
- Scalable computation by leveraging simplicity of classical linux tools and power of modern scalable workflow platform

Tools and Techniques

- **`jq`** to convert from **`json`** to tabular format
- **`Awk`** to perform core data crunching
- ANL's **`Swift`** to run awk in data-parallel setup
- Classic Linux tools for post processing: **`sort`**, **`sed`**, **`uniq`**, **`tr`**
- **`D3`**, **`graphviz`**, **`matplotlib`**, **`ffmpeg`** to visualize results

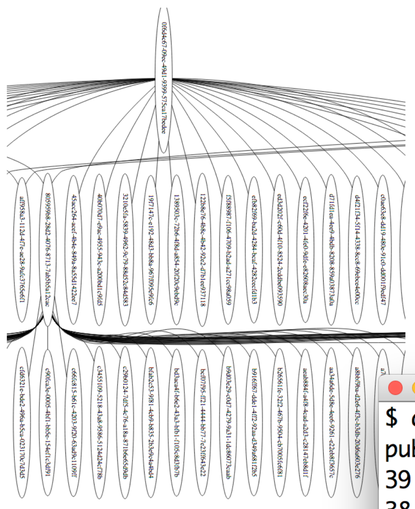
Data

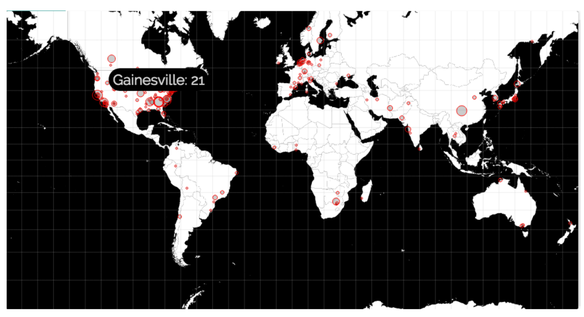
- **256 million** scientific publication records spread across 322 files
- Total data size is 329 GB, format is json
- Auxiliary data used: list of countries, cities, universities and a set of common words to filter from results

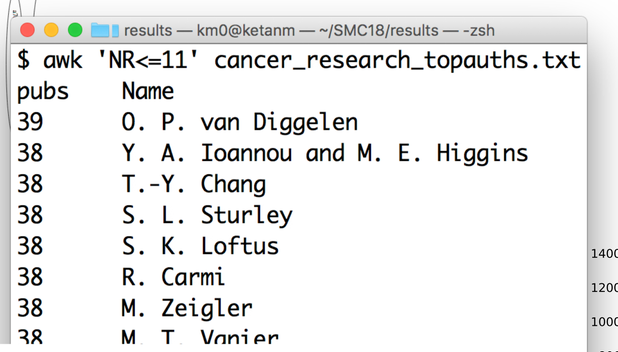
Benefits of using Awk+Swift

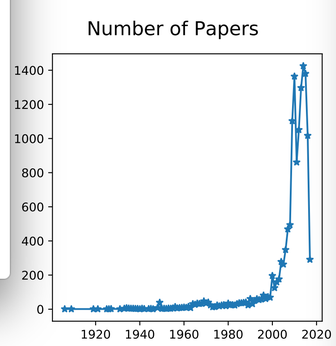
- Awk is highly suitable to text-processing portable to Linux systems
- Swift is portable to shared as well as distributed memory architectures
- Swift is known to work on leadership class supercomputers, eg. Summit

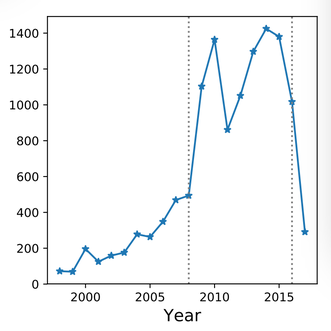
Results

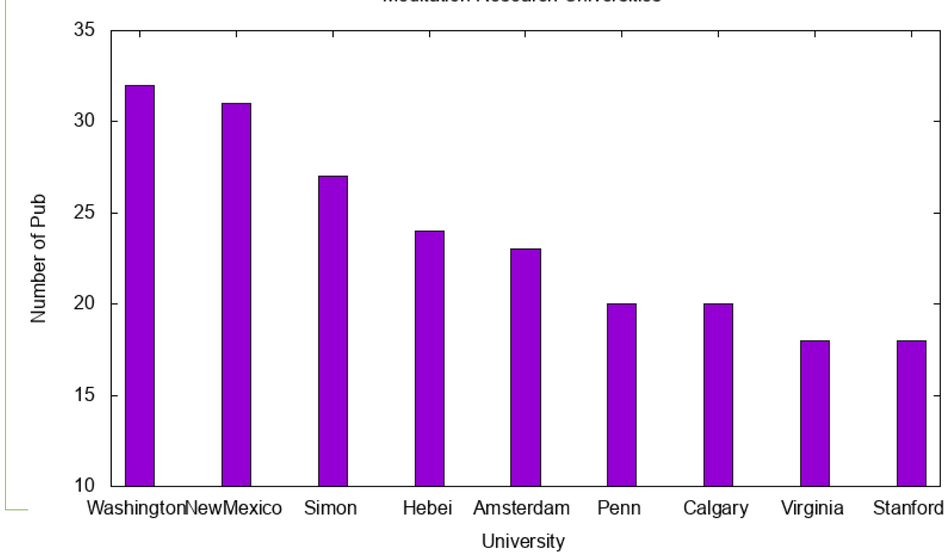


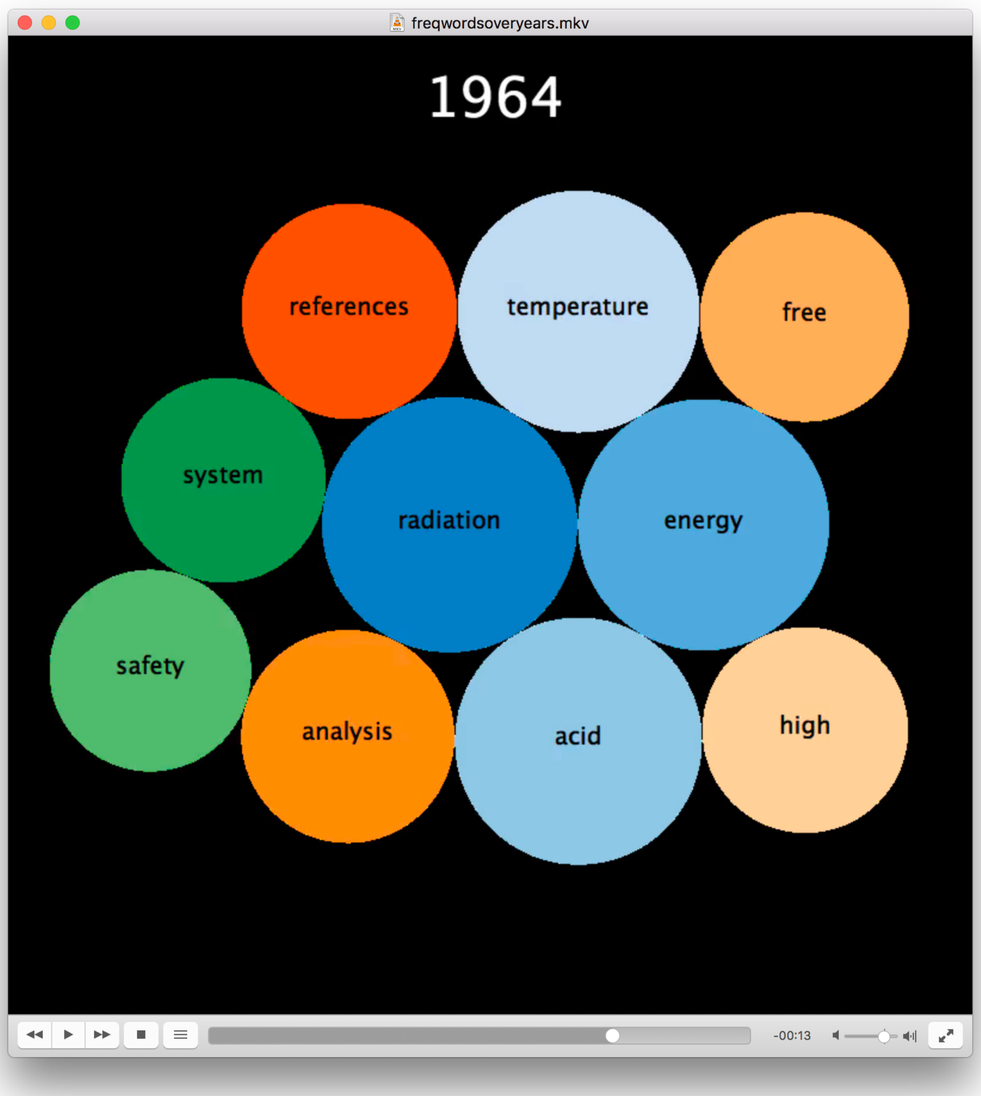












Performance

- Three out of five solutions takes **less than a minute** to finish
- Remaining two solutions takes less than an hour to finish
- Swift radically improves run time by concurrently running awk over **in-memory data** across the input

Bottomline

- Awk crunches massive amounts of data
- Swift parallelizes hundreds of awk calls
- In-memory computation for speedy runs
- Portable across legacy and modern architectures
- Scalable from laptop to leadership class systems

Acknowledgments

CADES for resources. Suhas Somnath, Brian Zachary, Drew Schmidt for valuable inputs. SMC organizers for hosting.

OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY