FUSE: A Reproducable, Internet-scale Dataset of Spreadsheets

Titus Barik*†, Kevin Lubick, Justin Smith, Emerson Murphy-Hill
*ABB Corporation Research, Raleigh, North Carolina, USA
North Carolina State University, Raleigh, North Carolina USA

***Abstract*—We submit a corpus consisting of 1.5 million spreadsheets, extracted using the Common Crawl corpus (based on Blekko). This corpus is compared against a proprietary index from a leading search engine company to measure representative against other commercial index. This corpus is intended to replace EUSES.**

I. INTRODUCTION

Spreadsheets are one of the most common end user programming environments [1]. Spreadsheets provide a platform for formula-based and conditional computations. Within a spreadsheet, users can develop complex programs using macros. Like traditional programming environments, researchers have studied code smells, refactoring, and debugging in spreadsheets [2], [3], [4].

Such research is supported by some notable existing corpora, such as the EUSES spreadsheet corpus and the Enron spreadsheet dataset [5]. CITE However, the available spreadsheet corpora exhibit several limitations. Currently spreadsheet corpora exist as snapshots; they do not provide mechanisms for reproduction or updates. Further, with 4498 spreadsheets, the EUSES dataset represents a small sample of all existing spreadsheets and the Enron dataset contains spreadsheets from only one company. Researchers who wish to use these corpora must manually reason about the properties of each spreadsheet as neither corpus provides spreadsheet-level metadata. For example, a researcher only interested in studying spreadsheet macros would have to download and inspect the contents of each spreadsheet from the EUSES corpus even though only 126 contain macros.

Alternatively, we provide a corpus creation technique that overcomes many of these limitations. Not only is our corpus larger, but our technique is open and reproducible. We publicly release the scanning software used to create this corpus. Accordingly, we encourage researchers to validate our data and methods. Further, to help researchers navigate the dataset, we provide metadata for each spreadsheet in our corpus as well as the tool used to compute that metadata.

Columwith is: The text height is 252.0pt

Our contributions are

1) A corpus of ??? spreadsheets pulled from the pubic web
2) A pipeline of tools that allows other researchers to more easily perform spreadsheet extraction and analysis at scale
3) A detailed set of metadata for this corpus and two other corpora [foo], [bar] which allow researchers to make queries without having to download or analyze the entire spreadsheet corpus.

TABLE I
COMMON CRAWL ARCHIVE

Description	TB	Yield
Summer 2013	30.6	42.14%
Winter 2013	35.1	33.46%
March 2014	36.4	
April 2014	41.2	
July 2014	59.2	
August 2014	46.6	
September 2014	48.9	
October 2014	59.1	
November 2014	31.4	
December 2014	35.2	

similar in size to other large corpora representativeness against Google index metadata

helps with tool evaluation

accessible – getting this for yourself is a high cost maybe 5k

this set is reproducable compared with existing data sets, it is new – the others are more than a decade old.

even the enron corpus is 15000 spreadsheets – whoopee. But it's good to compare against since it's a private corpus.

The contribution of this paper is:

- Foo

go to figshare

what metrics?

happy medium between euses and google

deliberate design decision to intentionally spluit the data into metadata and binary phases

II. METHODOLOGY

We selected the Common Crawl[1] index as the primary source for our spreadsheet corpus, because it contains over ???TB of publicly available web crawl data and is regularly updated.

To extract spreadsheets from this index, first we crawled all the available WAT files in the index targeting files that could potentially be spreadsheets, including files tagged with MSDN content types, and files with extensions containing ".xls". This crawl was the most computationally intensive task

[1] http://commoncrawl.org

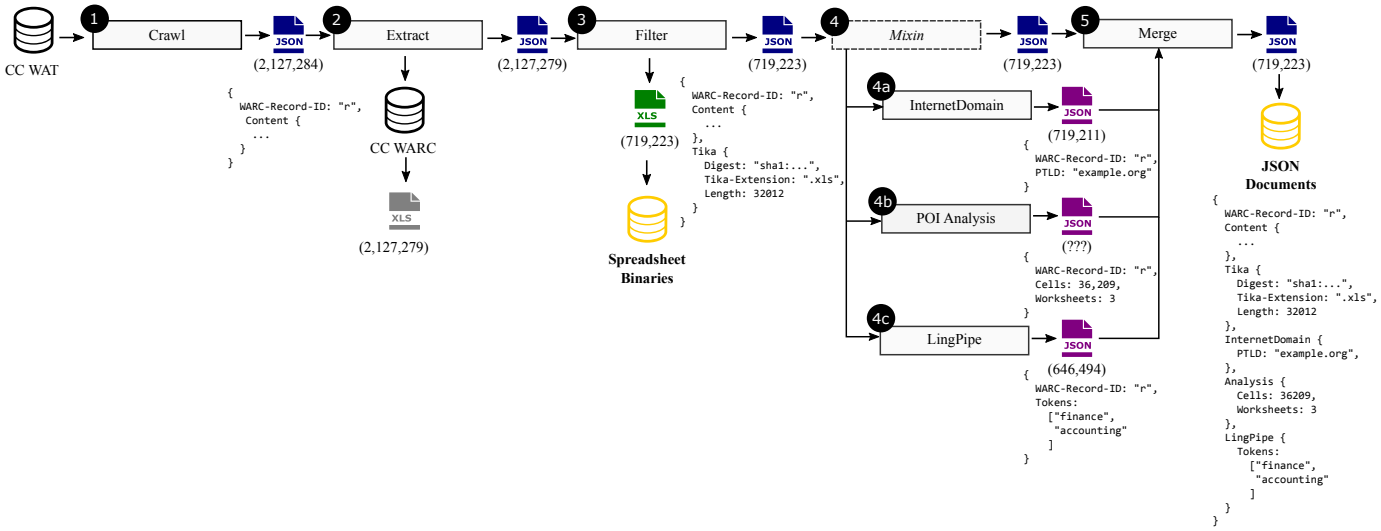

Fig. 1. The MapReduce pipeline for extracting spreadsheets and associated spreadsheet metadata from Common Crawl. In CRAWL (Stage 1), WAT segments containing HTTP headers and offset information into WARC records are parsed. Records that heuristically match spreadsheets (e.g., a `Content-Type` of `application/vnd.ms-excel` are retained. In EXTRACT (Stage 2), ...

in our pipeline, consuming approximately ??? hours of CPU time. The crawl identified ??? candidate spreadsheets which we extracted from their associated WARC files.

Because the WAT files contained some incomplete and incorrect tags, we extracted some invalid spreadsheets in the first phase. We used Tika[2] to identify the valid spreadsheets and then filtered out the invalid files. This stage resulted in ??? valid spreadsheets.

Next we processed the valid spreadsheets using Apache POI[3] to extract metadata from each spreadsheet (see Section V). By computing a sha512 hash we identified and removed duplicate spreadsheets, resulting in a final total of ??? spreadsheets.

+ TODO(tbarik): Copy Mixin and JsonMerge, not shown, but utility tool to easily copy sets and merge them.

A. MapReduce Pipeline

1) Crawl: In this step,

2) Extract: In this step,

3) Filter: In this step,

4) Mixin: In this step,

5) Merge: In this step,

limitation: truncation, 2 MB, 5 MB?

+ TODO(tbarik): Stages or tasks?

needed to create WAT index files

For Summer 2013, Winter 2013, and March 2013, no pre-created WAT path files were available. To manually create this we did:

how did we decide to filter? MSDN content-type

google obtained from proprietary database query that extracted from google index all spreadsheets with content type application/ms-excel.

[2]http://tika.apache.org

[3]http://poi.apache.org

common crawl segments -¿ subproblems

docjoiner

cleaning

put lingpipe put interndomain put analysistool

III. DESCRIPTION OF SPREADSHEET CORPUS

```
db.s.aggregate(
    { "$project":
      {"InternetDomainName.Public-Suffix" : true }
    { "$group":
      {"_id": "$InternetDomainName.Public-Suffix",
    { "$sort": { "count" : -1 } },
    { "$limit": 10 }
)
```

since it's sdescription of corpus, should be descriptive? counts avg etc.

absfrequency, relativefrequency gov, 250211 org, 171845 com, 112573 edu, 63572 gov.au, 43490 pa.us, 7621 net, 7409 mn.us, 4641 ac.uk, 3423 ca.us, 3003

absfreq, relfreq census.gov, 143135 triathlon.org, 106486 amamanualofstyle.com, 45118 abs.gov.au, 42941 utah.gov, 22242 ohio.gov, 16739 usda.gov, 13016 worldbank.org, 11062 theahl.com, 10350 eia.gov, 8216

RQ: How many domains are represented?

4,381 domains. After dedup, 4,342.

RQ what types of headers do people use?

how canonical are urls?

not very

results from this can be used to guide future crawls to increase the diversity of the spreadsheet corpus.

Actually looking at binaries, so this looks at dedup:

application/vnd.ms-excel, 238673

application/vnd.openxmlformats-officedocument.spreadsheetml.sheet,

TABLE II
COMPARISON OF FUSE AND OTHER SPREADSHEET CORPORA

	FUSE	**EUSES**	**Enron**
Size (n)	249,376	6,000	15,570
Space (GB)	b	0.64	23.3
Research access	All	Researchers	All
Unique formulas	894361	693266	84004
Extendable	Yes	Not scalable	No
Framework	Hadoop	Excel + VBA	Scantool
TIme Period	2006	2006	2006

TABLE III
SPREADSHEET CLASSIFICATION

Category	**FUSE**	**EUSES**
Database	4518	720
Finance	3058	780
Grade	2915	731
Homework	61	682
Inventory	2243	756
Model	2143	966

10555 application/vnd.openxmlformats-officedocument.spreadsheetml.template, 148

all zero: application/vnd.ms-excel.sheet.macroEnabled.12 application/vnd.ms-excel.template.macroEnabled.12 application/vnd.ms-excel.addin.macroEnabled.12 application/vnd.ms-excel.sheet.binary.macroEnabled.12

RQ: How much can you trust HTTP headers?

RQ: Train a text classifier to identify topicality.

RQ: Evolution of spreadsheets

Breakdown of analysis:

"OK", 112421 "CORRUPT",87574 "OTHER", 31676 "BIFF5", 17643 "ENCRYPTED", 62

place a non-shitty graph here

A summary of your spreadsheet corpus can be found in Table II.

IV. RESULTS

A. Classification of spreadsheets

The results in this section are intended to demonstrate the essential properties of the corpus.

B. RQ2: How diverse is the common common crawl corpus?

C. RQ1: How stable are URIs?

D. RQ3: NLP Extraction of Spreadsheet?

E. RQ4: What types of formulas ar eused by end-user software programmers?

Top two formulas are the same across the three corpora: 1. Add up the three cells to my left 2. Add up the two cells to my left And most of the top 10 are Add up n cells to my left. Fuse #3 HYPERLINK("http://www.eia.doe.gov/totalenergy/data/monthly/dataunits.cfm", "Notes: Information about data precision and revisions.") Enron #3 NOW()

give warc record ids

V. DATA SCHEMA

In this section, we describe the schema. The metadata we collected for the indices was largely influenced by the summary statistics presented in [5]. For each spreadsheet, there are over 450 entries, so we will not list them all here. In general, the entries summarize the contents of the cells. To list a few examples, the number of times a given Excel function (such as SUM or VLOOKUP) is used, the total number of input or data cells, the number of numeric input cells, the number of formulas used more than 50 times, the most common formula used, etc.

render json

```
asdfsdf
```

VI. MIXINS

how do you get it?

csv file mongodb recordobject warc extracts

VII. DATASET CHALLENGES AND LIMITATIONS

relevance is dependent on common crawl definition of relevant

VIII. CONCLUSION

The conclusion goes here.

IX. RELATED WORK

A. Why use spreadsheets

[6] Use spreadsheet corpus + interviews to determine which features end-users use.

[2] Detecting code smells in spreadsheets. Analyze EUSES to study occurrence of smells.

[3] Refactor spreadsheet formula. Perform case study using EUSES dataset.

[4] Support debugging spreadsheets

B. What other corpora?

EUSES [5]

[7] Automatically extract relational data from spreadsheets. Extracted 410,554 spreadsheets from clue09 web crawl.

ENRON find citation ¡– icse seip 2015

[7] Automatically extract relational data from spreadsheets. Extracted 410,554 spreadsheets from clue09 web crawl.

Existing corpora. Spreadsheet tools.

ACKNOWLEDGMENT

This material is based upon work supported in whole or in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.

Just kidding, I did this all by myself.

REFERENCES

[1] C. Scaffidi, M. Shaw, and B. Myers, "Estimating the numbers of end users and end user programmers," in *Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on*. IEEE, 2005, pp. 207–214.

[2] M. Pinzger, F. Hermans, and A. van Deursen, "Detecting code smells in spreadsheet formulas," in *Proceedings of the 2012 IEEE International Conference on Software Maintenance (ICSM)*, ser. ICSM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 409–418. [Online]. Available: http://dx.doi.org.prox.lib.ncsu.edu/10.1109/ICSM.2012.6405300

[3] S. Badame and D. Dig, "Refactoring meets spreadsheet formulas," in *Proceedings of the 2012 IEEE International Conference on Software Maintenance (ICSM)*, ser. ICSM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 399–409. [Online]. Available: http://dx.doi.org.prox.lib.ncsu.edu/10.1109/ICSM.2012.6405299

[4] R. Abraham and M. Erwig, "Goaldebug: A spreadsheet debugger for end users," in *Proceedings of the 29th International Conference on Software Engineering*, ser. ICSE '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 251–260. [Online]. Available: http://dx.doi.org.prox.lib.ncsu.edu/10.1109/ICSE.2007.39

[5] M. Fisher and G. Rothermel, "The EUSES spreadsheet corpus," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4. ACM, Jul. 2005, p. 1.

[6] C. Chambers and C. Scaffidi, "Struggling to excel: A field study of challenges faced by spreadsheet users," in *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, ser. VLHCC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 187–194. [Online]. Available: http://dx.doi.org.prox.lib.ncsu.edu/10.1109/VLHCC.2010.33

[7] Z. Chen, M. Cafarella, J. Chen, D. Prevo, and J. Zhuang, "Senbazuru: A prototype spreadsheet database management system," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1202–1205, 2013.