FUSE: A Reproducable, Internet-scale Dataset of Publicly-indexed Spreadsheets for Science

Titus Barik*†, Kevin Lubick†, Justin Smith†, Emerson Murphy-Hill†
*ABB Corporate Research, Raleigh, USA
†North Carolina State University, Raleigh, USA
titus.barik@us.abb.com, {kjlubick, jssmit11}@ncsu.edu, emerson@csc.ncsu.edu

***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. CALL FOR PAPERS

Data papers. We want to encourage researchers to share their data. Data papers should describe data sets curated by their authors and made available to others. They are expected to be at most 4 pages long and should address the following: description of the data, including its source; methodology used to gather it; description of the schema used to store it, and any limitations and/or challenges of this data set. The data should be made available at the time of submission of the paper for review, but will be considered confidential until publication of the paper. Further details about data papers are available on the conference website.

1) Description of the data.
2) Methodology used to gather it
3) Description of the schema used to store it
4) Limitations and Challenges of data set

II. INTRODUCTION

spreadsheets as end-user programmers; focus on how spreadsheets are useful for software engineerig

all the problems euses sucks [1]

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

A. Subsection Heading Here

Subsection text here.

need a fancy table

what metrics?

happy medium between euses and google

1) Subsubsection Heading Here: Subsubsection text here.

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

III. DATA EXTRACTION

needed to create WAT index files

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.

common crawl segments -¿ subproblems

docjoiner

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet,

consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

cleaning

IV. Description of Spreadsheet Corpus

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

V. Data Schema

how do you get it?
csv file mongodb recordobject warc extracts

VI. Dataset Limitations

VII. Conclusion

The conclusion goes here.

VIII. Related Work

why do people use spreadsheets?
what other corpora exists?
Existing corpora. Spreadsheet tools.

Acknowledgment

I did this all by myself.

References

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

TABLE I
A COMPARISON OF THE EUSES, COMMON CRAWL INDEX, AND PROPRIETARY INDEX

	EUSES	**Common**	**Prop. Search Index**
n	6,000	600,000	$>$ 1,000,000
size	23.3	32	2
Five	Six		