go.ncsu.edu/fuse

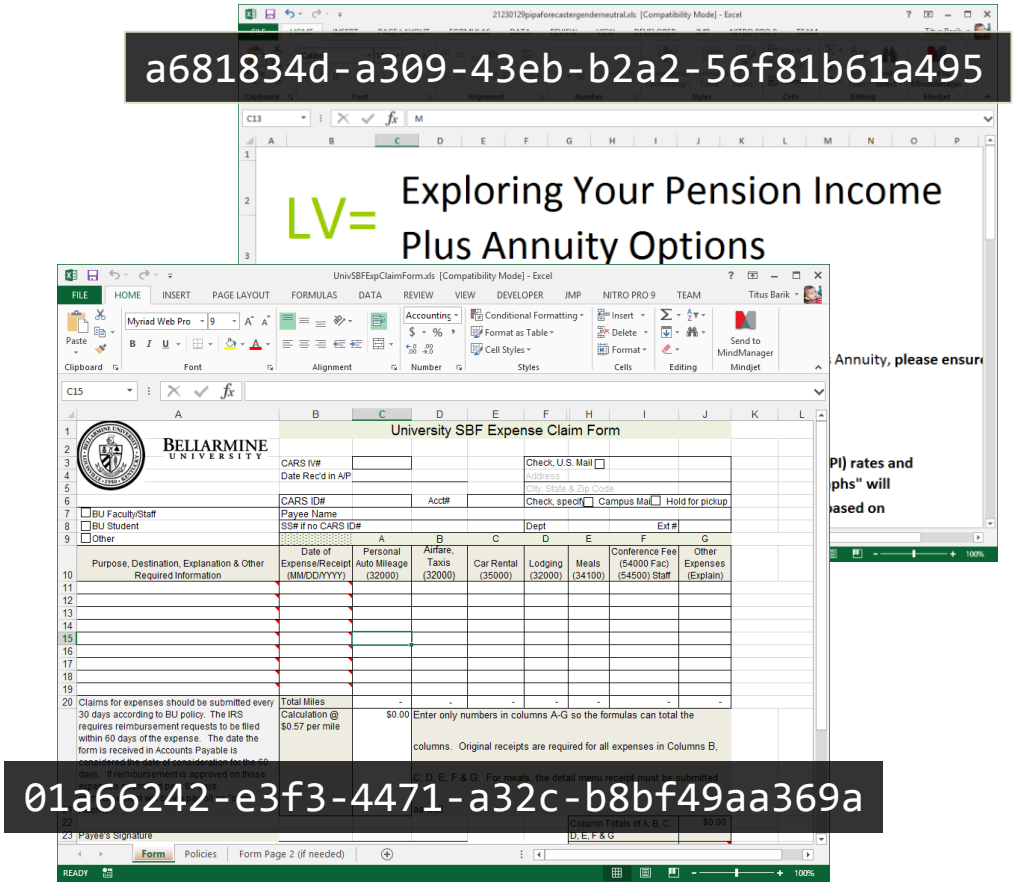
@barik

FUSE: A Reproducible, Extendable, Internet-scale Corpus of Spreadsheets

Titus Barik, Kevin Lubick, Justin Smith, John Slankas, Emerson Murphy-Hill

Why Spreadsheets?

Spreadsheets are perhaps the most ubiquitous form of **end-user programming software**.




```
db.spreadsheets.find( { "LingPipe.Tokens":
  {$all : ["travel", "request", "university",
           "reimbursement"]} } )
```

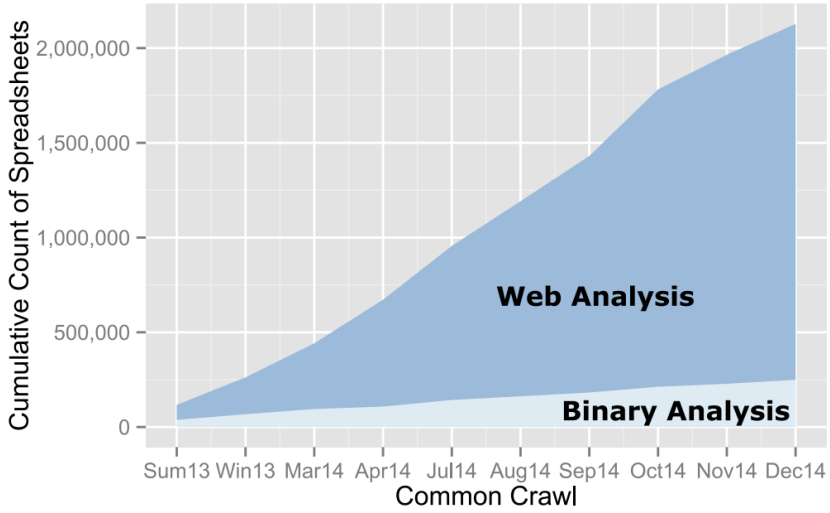
Description of FUSE

Extracted from Common Crawl archive of 26 billion web pages.

Over **2 million URLs** that return spreadsheets.

Over **249 thousand unique spreadsheets**.

JSON metadata for each spreadsheet.




```
{
  "WARC-Target-URI":
      "http://www.example.org/res
  "InternetDomainName": {
    "Host": "www.example.org",
    "Top-Private-Domain": "exampl
    "Public-Suffix": "org"
  }
}
```

```
{
  "LingPipe": {
    "Tokens": [
      "finance",
      "city"
    ]
  }
}
```

FUSE Benefits

Reproducible.

Independently obtain identical corpus.

Extendable.

Common Crawl archives released monthly.

Internet-scale.

Contains spreadsheets from diverse domains.

Thanks!

FUSE is unencumbered by any license agreements, available to all, and intended for wide usage by researchers.

go.ncsu.edu/fuse

@barik