What are the “extracted features”? We have extracted a number of useful features, including part-of-speech (POS) -tagged token counts, reader and footer identification, and various line-level information. Providing token information at the page level makes it possible to separate paratext from text (e.g. identify pages of publishers’ ads at the back of a book).

We break each page into three parts: header, body, and footer. We perform correction of hyphenation of tokens at end of lines, but not any additional data cleaning or OCR correction.

How are the extracted features delivered to the user? Features are provided to users in the form of files in the JSON (JavaScript Object Notation) format — a commonly used lightweight data interchange format. The EF datasets are transferred via rsync, a robust file synchronization/transfer utility.

The EF dataset for any HTRC workset can be retrieved as follows. A user can create a workset (or choose an existing workset) from the HTRC Portal, and execute the EF rsync script generator algorithm with that workset. This will produce a script that the user can download and execute on their own machine. When executed on the user’s machine, the script transfers the EF data files for that workset from the HTRC’s server to the user’s hard disk, resulting in two zipped files containing EF data in JSON format — “basic” and “advanced” — for each volume in the selected workset.

An EF data file for a volume consists of the volume-level metadata and of the data for each page in the volume, in JSON format. Each page is an array element containing the features (basic or advanced) for that page.

Here is a simplified EF data file for basic features, with metadata and features for a single page:

```json
{  "id": "t1fj34w02",  "metadata": {    "schemaVersion": "1.1",    "tokenPosCount": {      "postCount": 1,      "preCount": 1,      "INTRODUCTION": 1,      "DRAMATIS": 1,      "SHAKESPEARE": 1,      "ENGLISH": 1,      "AND": 1    },    "dateCreated": "2015-02-12T13:30",    "title": "Shakespeare's Romeo and Juliet",    "pubDate": "1920",    "languages": [      {"en": 1.00}    ],    "sentenceCount": 6,    "emptyLineCount": 9,    "lineCount": 35,    "tokenCount": 212,    "header": {      "languages": [        {"en": 1.00}      ],      "sentenceCount": 7,      "emptyLineCount": 10,      "lineCount": 35,      "tokenCount": 212    },    "body": {      "languages": [        {"en": 1.00}      ],      "sentenceCount": 5,      "emptyLineCount": 8,      "lineCount": 28,      "tokenCount": 195    },    "footer": {      "languages": [        {"en": 1.00}      ],      "sentenceCount": 3,      "emptyLineCount": 3,      "lineCount": 3,      "tokenCount": 31    }  }
```

References


Acknowledgments

The work described here is generously supported by the HathiTrust Consortium.