README

Anna Čermáková
Centre for Corpus Research
College of Arts and Law
University of Birmingham
a.cermakova@bham.ac.uk

Anthony Hennessey
Statistics and Probability
School of Mathematical Sciences
University of Nottingham
anthony.hennessey@nottingham.ac.uk

Jamie Lentin
Shuttle Thread
Manchester
jamie.lentin@shuttlethread.com

Michaela Mahlberg
Centre for Corpus Research
College of Arts and Law
University of Birmingham
m.a.mahlberg@bham.ac.uk

Viola Wiegand
Centre for Corpus Research
College of Arts and Law
University of Birmingham
v.wiegand@bham.ac.uk

Contents

1 Corpora

1.1 ChiLit - Children's Literature

From work by Anna Čermáková.

1.1.1 Included texts

2 Cleaning of corpora texts

The sources were the Gutenberg plain text UTF-8 files.

1. Convert to unix line endings.
2. Remove non-authorial text.
3. Reformat the book title and author to make consistent across all texts.
4. Reformat chapter headings to make consistent across all texts.
5. Manual corrections
6. Convert to 7-bit ASCII.

Steps 2, 3 and 4 were done manually.
Step 1 was achieved using the following command

```
for f in ChiLit/*.txt; do dos2unix -m $f; done
```

Some specifics of step 2:

- Tables of content are removed.
- Lists of illustrations are removed.

- Any preface text attributed to a person other than the author is removed. When attribution is unclear the text is left.
- Any postface text attributed to a person other than the author is removed. When attribution is unclear the text is left.
- Transcriber notes are removed.
- In the texts illustrations are usually indicated by text enclosed in square brackets. Where this text includes a caption the caption is kept, for example

  ```
  [Illustration: THE WONDERSTONE.]
  ```

  becomes

  ```
  [THE WONDERSTONE.]
  ```

  Where there is no authorial caption the construct is deleted. All the following example would be deleted

  ```
  [Illustration]

  [Illustration: Chapter Seventeen]

  [Illustration: Page 91]
  ```

- The book title is put on the first line of the file, without any newlines.
- The book author is put on the second line of the file, without any newlines.
- Chapter headings are formatted as follows: If the chapter heading begins with 'CHAPTER' or 'BOOK' it must be followed by a number or roman numerals and then a dot. The chapter or book number cannot be written in word form. The heading can optionaly be followed by a chapter title; the chapter title must not break onto a new line. Here are some examples

  ```
  CHAPTER 1. The Old Sea-dog at the Admiral Benbow

  CHAPTER 2. TRAVELLING COMPANIONS.

  CHAPTER 3.

  CHAPTER IV. Little Meg's Treat to Her Children

  CHAPTER V.

  BOOK 1.

  BOOK II. Jessica's Mother
  ```

  Sections beginning with 'INTRODUCTION', 'PREFACE', 'CONCLUSION', 'PROLOGUE', 'PRELUDE' or 'MORAL' are also be treated as seperate chapters. These do not require numbers, but do require the dot. Again the heading can optionaly be followed by a title; the title must not break onto a new line. Here are some examples

  ```
  PREFACE.

  INTRODUCTION.

  PROLOGUE. THE OLYMPIANS
  ```

```
 MORAL.--_There is no moral to this chapter._
```

In all cases there must be no space at the beginning of the line.

- Part headings are on a line before the first chapter of that part, in the same format. Blank lines are allowed between the part heading and the chapter heading. The following example is from Treasure Island

```
PART TWO. The Sea-cook

CHAPTER 7. I Go to Bristol

IT was longer than the squire imagined ere we were ready for the sea,
and none of our first plans--not even Dr. Livesey's, of keeping me
```

Step 6 unifies the use of hyphens, apostrophes and quotes across the texts; 7-bit ASCII[1] is a subset of UTF-8[2] and so the files may be treated as UTF-8.
Step 6 was achieved using Version 1.30 of the Perl module `Text::Unidecode`[3] with the following command

```
perl -C -MText::Unidecode -n -i -e'print unidecode($_)' *.txt
```

3 Maintaining this corpora repository

3.1 `.bib` file

We currently manage the bibliography in a shared zotero folder. The important fields in the bib entries are:

- The `shorttitle` field must match the filename of the relevant text file in the corpus folder.
- The `keywords` field must contain the name of the corpus.
- The `title`, `author` and `date` fields must be present.

Example entry:

```
@book{grahame_wind_1908,
    title = {The Wind in the Willows},
    url = {https://www.gutenberg.org/ebooks/289},
    shorttitle = {willows},      <<===  filename willows.txt
    author = {Grahame, Kenneth},
    editor = {Lough, Mike},
    urldate = {2017-06-28},
    date = {1908},
    keywords = {{ChiLit}}        <<===  corpus id
}
```

3.2 Adding a new text to a corpus

1. Clean the text as described in Section 2.
2. Add entry to the `.bib` file; see Section 3.1.
3. Generate a new `coorpora.json` file.

   ```
   cd scripts
   R --no-restore --no-save <./bib2json.R
   ```

4. Update repository tags; see Section 3.4.

[1] https://tools.ietf.org/html/rfc20
[2] https://tools.ietf.org/html/rfc3629
[3] http://search.cpan.org/perldoc?Text::Unidecode

3.3 Adding a new corpus

1. Edit the `scripts/bib2json.R` script to include the new corpus.
2. For each new corpus file
   (a) Clean the text as described in Section 2.
   (b) Add entry to the `.bib` file; see Section 3.1.
3. Generate a new `coorpora.json` file.

```
cd scripts
R --no-restore --no-save <./bib2json.R
```

4. Update repository tags; see Section 3.4.

3.4 Repository Tags

TODO