Merck medical papers challenge project documentation

Maximilian Balthasar Mansky, Richard Schulz

August 8, 2020

1 Introduction

The Merck Medical Papers challenge is a public request to develop an algorithm to best predict future citations from a given article. Dataset for this challenge is the open access PubMed bulk package, around 30 GB of medical papers in XML format.

1.1 Data format

In the XML format, the data is organised in a tree structure, with meta data and abstract located under `article/front` and the main text under `article/body`. The `article/back` section contains information regarding funding and citations for the body.

Each article file is named with its PMC ID, a unique 7 digit number prefixed by 'PMC', file ending `.nxml`.

2 First Steps

Before delving into the depths of NLP, we first check whether auxiliary information can already predict citations to any accuracy. For this, prepare a table from the files, including the following headers:

PMC ID	Journal name	Publication date	Number of citations
`numerical id`	`string`	`date`	`integer`

The first three can be extracted from the files, the last one needs to captured from the website. This table will also serve as a master table for connecting citations to PMC ID. Trying to predict the citation count from

meta data is not without merit, it has been shown to be highly correlated with the citation count.[1]

Much of the required information is available straight from the index file.[2] From there only the citations need to be retrieved.

A line in the index file has the structure:

```
oa_package/08/e0/PMC13900.tar.gz\tBreast Cancer Res. 2001 Nov 2; 3(1):55-60\tPMC13900
```

Splitting by tabs, the name of the publication and the date can be retrieved by the following regex:

```
([\w\s]+.?)\s(\d{4}\s\w{3}\s\d+);
```

and for those papers where the day is missing or a month-range is given:

```
([\w\s()]+.?)\s(\d{4}\s\w{3})[-;\s]
```

[1] https://github.com/closingsquarebracket/merck_medical_papers

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#index_files