Results with Bookmarker: testing

# Random post 2000 reports

Observations:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| report | type | Toc? | Headings? | Bookmark quality? | comments |
| 31649 | Annual report | yes | yes | Complete, but all except summary are classes as subheading because they have a 1.0, 1.1., 2.0 pattern (all are at least cyfra1 punkt cyfra2) | Change heading id intext num2cyfra1   * to transform a trailing zero to nothing ? |
| 32783 | Partial relinquishment report | yes | no | As expected | File is only title page and TOC   * Need to concat with rest of body |
| 33906 | Partial relinquishment report | yes | yes | As in 31649 | See 31469 |
| 39884 | Well proposal | yes | yes | Middling. A few headings are found, but a bit more than half aren't, possibly because they're long and unconventionally named | To improve, expand heading training set with more headings? Maybe that would just be confusing - by being long, not only are they a challenge for CNB, but for BB attributes like wordswidth. Suddenly comparison to TOC headings means a lot more. |
| 40262 | Annual report | Yes, but not recognised | yes | Very middling, probably due to TOC not being recognised | Why wasn't TOC recognised? It doesn't look that strange. |
| 40414 | Partial relinquishment report | yes | yes | Middling. More than half not bookmarked, unclear why except they might be a little long/unseen. | Probably just wants training on reports like it. |
| 47867 | Partial relinquishment report | yes | yes | perfect |  |
| 48118 | Any other report | Yes but extremely small and double columned | Lots, but like article titles and sections | None except for TOC | Queensland mining journal. Date of report is 1919, which my date transform function mistakenly assigned as 2019  I'm going to need to take mining journals through a different process, they can't get restruct page lines - two column pages. In fact they need a reverse of that. |
| 49708 | Annual report | yes | yes | Perfect in terms of catching every real heading, but also catches a lot of noise of the same (second line of) header, which it mistakenly also caught in the TOC. | Need to include second line of a marginal as a marginal too. Need to work with more data. |
| 52224 | Final or end of tenure report | yes | yes | perfect |  |
| 53805 | Partial relinquishment report | yes | yes | perfect |  |
| 54753 | Transmission report | no | no | none | Nothing to bookmark |
| 56794 | Well proposal | yes | yes | Would be perfect, but TOC continues onto a second page, and those headings aren't caught | Need to allow a longer than 1 page TOC |
| 57509 | Well completion report | yes | yes | Lots of sub-sub-headings, so quite messy with it. Not all sub-headings are caught, but all headings are, despite the TOC going onto a second page. Bookmark levels (heading vs subheading) are messy. |  |
| 57514 | Well completion report | Yes, but not recognised | yes | Very messy, second line of header often recognised heading. | Similar report to 57509, but significantly worse result as TOC wasn't caught. |
| 57966 | Annual report | yes | yes | Almost perfect, just one heading marked as sub-heading. |  |
| 62745 | Well completion report | no | Yes, but with no numbering and no TOC they'd be near impossible to catch | Bad. Some bookmarks, almost all wrong. |  |
| 76605 | Annual report | yes | yes | Very sparse, only two headings are caught. Cause unclear. |  |
| 77000 | Annual report | yes | yes | perfect |  |
| 80420 | Well completion report | no | yes | The bookmarker tries, and get some right, but there's a lot of mess among it |  |
| 80685 | Well completion report | no | Yes, but it's a welcom so they're lost amongst the tables | Bad. All noise | Recommend treating WELCOMs *completely* differently to this. |
| 82778 | Well completion report | no | Yes, but it's a welcom |  |  |
| 83936 | Well completion report | no |  |  |  |
| 84284 | Well completion report | yes | yes | Almost perfect | It's a welcom, but it has a TOC, limited tables, and lots of text!  Arrow Energy, 2014 |
| 91779 | Annual report | yes | yes | Very few headings caught, second line of footer caught. | Reason for low quality not clear except for unfamiliar language. |
| 92768 | Well completion report | no | Yes but it's a welcom | bad | Gross tabular welcom by Australia Pacific LNG |
| 103455 | Partial relinquishment report | yes | yes | Almost perfect, just one heading as subheading |  |
| 112814 | Partial relinquishment report | yes | yes | Missing a few heading. Classes subheadings as headings but they have no number. |  |

Out of 28 post-2000 reports:

|  |  |  |
| --- | --- | --- |
| Report type | number | Ratings |
| Annual report | 7 | 5, 5, 5, 4, 2, 1, 1 |
| Final or end of tenure report | 1 | 5 |
| Any other report | 1 | 1 |
| Partial relinquishment report | 7 | None, 5, 5, 5, 5, 3, 2 |
| Transmission report | 1 | None |
| Well completion report | 9 | 5, 4, 2, 1, 1, 1, 1, 1, 1 |
| Well proposal | 2 | 4, 2 |

While WELCOM had some good ratings, most of them were un-bookmarkeable due to the format those reports took. No TOC, mostly tables. That format may be restricted to a company, or it may stem from a time period, when well data cards gained a standardised template. Further analysis is needed to determine that.

The one “Any Other Report” was an instance of the Queensland Mining Journal, which was mistakenly identified as being from 2019 when it was really from 1919. This error couldn’t have been avoided with the current processing, as there’s no indication in the QDEX\_metada\_export.xlsx table of the REPDATE century.

* However, the method could be modified to check if the RECDATE was after the assumed REPDATE. Then a report created in 1919, recorded in 2000, could be dated correctly.

TOC-rating correlation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Count | Mean | Median | Mode |
| Had TOC | 20 | 3.55 | 4 | 5 |
| Didn’t have TOC | 6 | 1 | 1 | 1 |

\* two reports were dropped for not containing headings and thus ratings

Report type ratings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Report type | Count | Mean | Median | Mode |
| Annual report | 7 | 3.2857 | 4 | 5 |
| Final or end of tenure report | 1 | 5 | 5 | 5 |
| Any other report | 1 | 1 | 1 | 1 |
| Partial relinquishment report | 6 | 4.167 | 5 | 5 |
| Well completion report | 9 | 1.89 | 1 | 1 |
| Well proposal | 2 | 3 | 3 | 2, 4 |

\* two reports were dropped for not containing headings and thus ratings

This shows good results for the Partial Relinquishment Report, average results for the Annual Report, and bad for the Well Completion Report.

With these results, I can try to make assumptions of what will work and what won’t:

- Remove WELCOM from sampling as they often take a much different format

- If a TOC is missing, don’t try to find headings?

- Won’t implement this yet to collect more data

# Random post 2000, non WELCOM reports

Looks good.

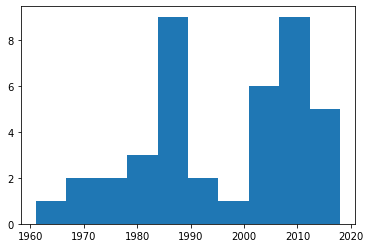
# Random non WELCOM reports [no date cut-off]

Prior to deciding to create new training set from post 1990 or 2000 reports, will test how the bookmarker does without a date cut-off, to perhaps find a better one.

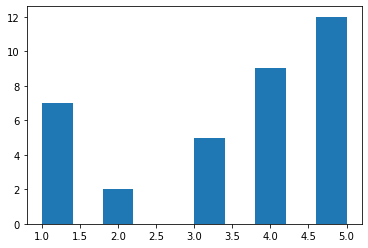
Observations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Report | Type | Repyear | Toc? | Headings? | Bookmark quality? | comments |
| 593 | Any other report | 1961 | No | Yes | 1 | Scan of body of the report, words look splotchy. I wouldn’t be surprised if the OCR was bad too. |
| 41568 | Any other report | 2002 | Yes | Yes | 5 | Unusually, “Appendix” is a single heading in the TOC, meaning it could be recognised, but isn’t due to my current exclusion of appendices. |
| 41982 | Any other report | 1996 | Yes | Yes | 3 | All correct bookmarks are found, but a lot of extra are also found due to a subheading title being used for maps. Heading and subheadings also aren’t numbered, which also increases the difficulty.  Maybe change my definition of figure pages to that they *only* contain figures (no legitimate text), and won’t contain headings? |
| 10189 | Well report other | 1981 | Yes | Yes | 2 | Catches about half of headings, also catches same amount of false positives. |
| 102109 | Annual report | 2017 | Yes | Yes | 4 | Almost perfect catching of headings, only misses 2/14 at the end, 3 minor false positives (very minor headings, not ones in the TOC). |
| 43758 | Partial relinquishment report | 2006 | Yes | Yes | 3 | Misses 3/6 major headings, 2/9 subheadings. 3 footer false positives. |
| 105472 | Partial relinquishment report | 2018 | Yes | Yes | 3 | 1/5 headings missing, 1 false positive. |
| 48907 | Any other report | 1987 | None | None | None | Not textracted |
| 15042 | Seismic survey final report | 1986 | Yes | Yes | 4 | All headings found, no subheadings found because they are neither numbered no paged in the TOC. Instead, they are lettered: a), b), etc. I think this is fine. |
| 41275 | Any other report | 1991 | Yes | Yes | 4 | Most headings are caught, only a few false positives, so the program is working well, however the score is docked from a 5 because the report itself doesn’t number headings or subheadings, so all are id’d as headings. And there are quite a few so the bookmarks bar looks a little long and messy. The only saviour is more major headings being all uppercase.  Improvement: as an alternate to primarily classify sub-headings as such due to their numbering pattern, use the amount of indentation (left) in comparison to the heading? This could be implemented just by adding training data to the set which is thusly indented, and seeing if it was enough for the result to change. |
| 4639 | Annual report | 1973 | No | Yes | 2 | No TOC, but catches 3 out of 17-19 headings. The ones it catches are those with very common heading vocab: introduction, Geology, Geology cont.  Could be improved with more variety in the vocab used for headings. |
| 48670 | Any other report | 1967 | Yes, but very differently formatted. | Yes | 1 | Queensland Mining Journal, so very different format. All bookmarks are of the header. However, this may actually be useful for finding the right page, as the page numbering is lie 279-313. |
| 3051 | Any other report | 1969 | Yes | Yes | 5 | Very good numbered TOC. Only a few false positives and for Summary subheadings (but classed as headings). Otherwise perfect. |
| 24357 | Annual report | 1993 | Yes | Yes | 5 | Perfect. |
| 15568 | Six monthly exploration report | 1986 | Yes | Yes | 4 | 1 false positive, 1 false negative heading. /6 |
| 68677 | Well or bore abandonment report | 2011 | Yes | Yes | 4 | Several footer false positives.  As marginal have a strange but understandable tendency to be false positives for headings, they could be excluded from going to heading identification. This would give them more reason to be identified.  That is, if they are correctly id’d as a marginal. They may still not be, as these are, by being on a secondary line of it. |
| 48897 | Any other report | 1986 | None | None | None | Not textracted. |
| 36490 | Partial relinquishment report | 2004 | Yes | Yes | 3 | Starts off well, but half the headings and about fourth of subheadings are missing at the end. |
| 5261 | Annual report | 1989 | Yes | Yes | 4 | Missing one heading /6. Missing all two subheadings, but they are written a little differently. |
| 44433 | Well proposal | 2006 | No | Yes | 1 | Bad without a TOC. |
| 11148 | Partial relinquishment report | 1982 | Yes | Yes | 5 | One heading mistakenly a subheading. |
| 11527 | Six monthly exploration report | 1983 | Yes | Yes | 5 | Perfect apart from one understandable false positive. |
| 57431 | Partial relinquishment report | 2009 | Yes | Yes | 4 | ¼ headings missing. |
| 68876 | Partial relinquishment report | 2011 | Yes | Yes | 5 | Perfect aside from informative false positive. |
| 45198 | Well report other | 2007 | Yes | Yes | 4 |  |
| 44457 | Well proposal | 2006 | No | No | None | One heading and a lot of figures. |
| 4411 | Final or end of tenure report | 1973 | Yes | Yes | 1 | Has a TOC, but it’s called “INDEX”, and is structured differently. |
| 62180 | Annual report | 2010 | Yes | Yes | 5 | Perfect |
| 92734 | Partial relinquishment report | 2015 | Yes | Yes | 5 | Perfect |
| 47215 | Well proposal | 2007 | No | Yes | 1 | Gets two headings, but there are a lot it doesn’t. |
| 42011 | Any other report | 1989 | Yes | Yes | 5 | Misses only 1 subheading. |
| 72200 | Well production testing report | 2012 | No | No | None | Entire file is one page table. |
| 47314 | Reserves | 2007 | No | No | None | Entire file is one page table. |
| 15648 | Seismic survey final report | 1986 | Yes | Yes | 5 | Perfect |
| 32424 | Partial relinquishment report | 2001 | Yes | Yes | 5 | Perfect |
| 75015 | Final or end of tenure report | 2013 | Yes | Yes | 4 |  |
| 13687 | Final or end of tenure report | 1984 | No | Yes | 1 |  |
| 20170 | Final or end of tenure report | 1989 | No | Yes | 1 |  |
| 79361 | Partial relinquishment report | 2013 | Yes | Yes | 5 | Perfect |
| 58954 | Well proposal | 2009 | Yes | Yes | 3 | Many false positives. Headings are not numbered. |

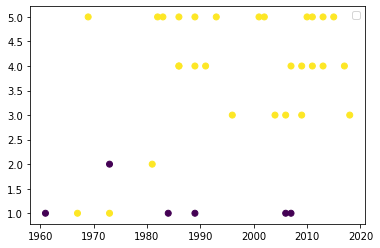
Distribution of years



Distribution of ratings



Scatter plot of Year-to-Rating (with yellow == has TOC, purple == no TOC)



One result of this investigation is staunchly that Queensland Mining Journals should be excluded like Well Completion Reports, and that reports without a TOC should be noted as such in some file. Older reports may not necessarily be excluded, though they have a higher incidence of being low quality, this is variable enough to process them anyway to get the good ones.

Oh yeah, a log file of results should always be output so the command line isn’t the only place metadata of the process exists.