Newspaper Title Routine

The routine can automatically detect newspaper titles (the “leads”), add a period after all titles and subtitles, and separate titles from the article corpus. Adding a period at the end of newspaper article titles is a crucial operation for the proper analysis of newspaper articles with NLP tools. These tools (e.g., the Stanford CoreNLP) split sentences on the basis of common sentence stoppers (e.g., . ! ?). Newspaper titles are generally typeset without sentence stoppers at the end. As such, titles would be processed by NLP tools as part of the first sentence.

INPUT

In input
The routine will ask you to type in a number for the expected maximum-length title in **number of characters** in your corpus; this will help the routine detect titles with higher accuracy. The default number is 20. For example, the title “MAN SHOT DEAD” will have length 13.

OUTPUT

In output, the routine will create a subfolder named ExtractTitles under the default output directory. Under the ExtractTitles folder the routine will create three subfolders: articlesWithoutTitles, articlesWithTitles, titles.

The folder ‘articlesWithoutTitles’ contains the original articles files excluding the titles. Only the article corpus is preserved.

The folder ‘articlesWithTitles’ contains the original files with a period added at the end of each title and subtitle.

The folder ‘titles’ contains one file called titles.txt, in which titles extracted for each input file are listed in numerical order.