HW9 indexer results

pt-index entry structure:

key: urlHash

urls column: URL1:TARGET\_WORD\_FREQUENCY1, URL2:TARGET\_WORD\_FREQUENCY2, …

example:

banana

http://foo.com/page1.html:2, http://foo.com/page2.html:3

banana appears 2 times in <http://foo.com/page1.html>, appears 3 times in <http://foo.com/page2.html>

pt-crawl entry:

For each html processed by indexer, indexer will add "totalWordCount" (non-stop-word valid word count), “macFreq” (the number of times the most frequent word appears), “title” , “h1”, “h2”, “h3”, “h4”, “h5”, “h6”, “keywords” and “description” columns (words joint by “,”) into the url’s corresponding pt-crawl row if information is available.

Example output results:

url: https://en.wikipedia.org:443/wiki/Climate\_of\_Svalbard

title: climate,of,svalbard,wikipedia

h1:

h2: historical,data,edit

h3: permafrost,edit

h4: terrestrial,ecology,edit

h5: terrestrial,invasive,species,edit

h6:

description: