Title: Assignment 2: Building a Corpus

Author: Zhiheng Wang

Date: 2/11/2019

Description: This program extracts all pages from Wikipedia's Category:2018 films, clean unstructured data, and store the results in json format. i.e.

- Title

- Director

- Starring

- Running_time (in minutes)

- Country - Language

- Time (what year(s) the story takes place)

- Location (city/country(s) where the story takes place)

- Categories (list of Wikipedia categories that the page belongs to)

- Text (the full text of the page, excluding the data extracted from templates)

Dependencies:

wikipediaapi https://pypi.org/project/Wikipedia-API/

wptools https://pypi.org/project/wptools/

regex https://pypi.org/project/regex/

beautifulsoup4 https://pypi.org/project/beautifulsoup4/

Build Instructions: You can install these packages in any sequence, and it is recommended to install them via pip.

Run Instructions: You can simply run it in terminal or cmd. You do not need to input anything since the input is extracted from wiki pages. The output is json

file with 10 fields as the format shows in the description above.