第3章: 正規表現

Wikipediaの記事を以下のフォーマットで書き出したファイルjawiki-country.json.gzがある．

- 1行に1記事の情報がJSON形式で格納される
- 各行には記事名が”title”キーに，記事本文が”text”キーの辞書オブジェクトに格納され，そのオブジェクトがJSON形式で書き出される
- ファイル全体はgzipで圧縮される

以下の処理を行うプログラムを作成せよ．

20. JSONデータの読み込み

Wikipedia記事のJSONファイルを読み込み，「イギリス」に関する記事本文を表示せよ．問題21-29では，ここで抽出した記事本文に対して実行せよ．

21. カテゴリ名を含む行を抽出

記事中でカテゴリ名を宣言している行を抽出せよ．

22. カテゴリ名の抽出

記事のカテゴリ名を（行単位ではなく名前で）抽出せよ．

23. セクション構造

記事中に含まれるセクション名とそのレベル（例えば”== セクション名 ==”なら1）を表示せよ．

24. ファイル参照の抽出

記事から参照されているメディアファイルをすべて抜き出せ．

25. テンプレートの抽出

記事中に含まれる「基礎情報」テンプレートのフィールド名と値を抽出し，辞書オブジェクトとして格納せよ．

26. 強調マークアップの除去

25の処理時に，テンプレートの値からMediaWikiの強調マークアップ（弱い強調，強調，強い強調のすべて）を除去してテキストに変換せよ（参考: マークアップ早見表 (http://ja.wikipedia.org/wiki/Help:%E6%97%A9%E8%A6%8B%E8%A1%A8)）．

27. 内部リンクの除去

26の処理に加えて，テンプレートの値からMediaWikiの内部リンクマークアップを除去し，テキストに変換せよ（参考: マークアップ早見表 (http://ja.wikipedia.org/wiki/Help:%E6%97%A9%E8%A6%8B%E8%A1%A8)）．

28. MediaWikiマークアップの除去

27の処理に加えて，テンプレートの値からMediaWikiマークアップを可能な限り除去し，国の基本情報を整形せよ．

29. 国旗画像のURLを取得する

テンプレートの内容を利用し，国旗画像のURLを取得せよ．（ヒント: MediaWiki API (http://www.mediawiki.org/wiki/API:Main_page/ja)のimageinfo (https://www.mediawiki.org/wiki/API:Imageinfo)を呼び出して，ファイル参照をURLに変換すればよい）

Updated: April 7, 2020