課題２　本文抽出

1. ソースコード

output.php に記載

2. アルゴリズム説明

① output.html（ブログの html 情報）の中身を取得する。

② 取得した html から本文とは関係なさそうなタグ、html のコメントを削除する。
（<head>,<script>,<noscript>,<style>,<header>,<footer>,<form>,<aside>）

③ プロフィールらしきものを削除。

④ div,td タグで囲まれた要素を１つずつ確認していき、句読点や記号（、。,．　！？）の数と文字数を数える。

⑤ 句読点や記号の数が多くなる場合や、同じ数であっても文字数が少なくなる場合はその要素を保存する（本文中には高確率で句読点や記号が含まれる。また、文字数だけが少なくなる場合には、無駄な要素が含まれているものと考えた。）

⑥ 取得できた要素から更に次のように思われるものを削除する。
コメント、bottom に関連する要素、サイドバー、ヘッダー、フッター、ウィジェット、プラグイン、バナー、ボタン、カレンダー、トラックバック、メニュー

⑦ もう一度④⑤の処理を繰り返す。

⑧ html タグや空白を削除して、得られた文字列を出力する。

3. 実行結果

result ディレクトリに input と output を記載。
完全に本文のみを抜き出せたのは４件程度。その他についても基本的に本文付近を抜き出せた。

4. 簡単な実行環境構築手順、プログラム実行手順

http://donatu33.sakura.ne.jp/kadai2/
にアクセスし、取得したいブログの URL を入力しダウンロード（input.html が取得できる）。次にそのファイルを選択し、抽出（output.txt）が取得できる。

5. 工夫した点、苦労した点、取り組み総時間

- **工夫した点**

  html を削除する際に、要素毎に削除すること。例えば単純に正規表現だけで削除しようとすると、 文章 などといった場合に のみが残ることが考えられる。

  句読点や記号だけでなく文字数をカウントし、文字数が短い方を優先した。例えば タイトル 本文 などの構成になっている場合に、自動的にタイトルなど不要なものが省かれると考えた。

  一度大まかに抜き取ってから、再度削除を行う。初めから全てを削除してしまうと想定以上に削除されてしまうことが多かった。

- **苦労した点**

  dom を扱える simple_html_dom.php の使い方を学習するのに苦労した。また、正規表現による要素の抽出もなかなか思うようにいかずに苦労した。削除するものを少なくすれば無駄なものが残り、多くすれば必要なものが消えるという状況で調整が難しかった。

- **取り組み総時間**

  ６０時間程度

6. 参考 URL

- ブログやニュースの本文を抽出する方法 - 僕のススメ。

http://d.hatena.ne.jp/steel-plate/20090528/1243514514

- ブログの本文抽出にチャレンジ - Ceekz Logs (Move to y.ceek.jp)

http://private.ceek.jp/archives/002039.html

- PHP Simple HTML DOM Parser の使用方法

http://so-zou.jp/web-app/tech/programming/php/library/simplehtmldom/