Scraping Tumblr

Jason Crome

Disclaimer

- This works well for most any site with an API

- You can use other modules than the ones I present. I happen to be fond of WWW::Mechanize

Background

- Once upon a time, I kept a blog on Tumblr
- Now, I hate Tumblr, but I didn’t want to lose the content
- Might have been quicker to copy/paste (there wasn’t that much there), but where’s the fun in that?!
- So, I wrote a program to do it for me.
- Read HTML from Tumblr, then wrote Markdown files for use by Hugo.

Step 1: Get Posts

```
sub get_posts {
  my $mech = WWW::Mechanize→new;

  my $response  = $mech→get($blog_url)→decoded_content;
  my $post_json = decode_json( $response );

  my @post_list = $post_json→{ response }{ posts };

  return @{ $post_list[0] };
}
```

```
\ {
    blog                    {
        description   "Where Jason spouts off about technology, consulting projects, adventures in small business, Perl, and life.",
        name          "crome-plated",
        title         "the crome-plated blog",
        updated       1514948977,
        url           "https://crome-plated.tumblr.com/",
        uuid          "t:dYHoT_zyXamW8D1uFBxGXg"
    },
    blog_name               "crome-plated",
    body                    "<p>Welcome to 2018, everyone! Here’s hoping to a happier, healthier, less chaotic year, filled with more kindness, love, and understanding than the world saw in 2017.</p><p>Airline flights are a good source of downtime, and have historically been a
great time for me to brainstorm and reflect, and today’s flight from Chicago-Midway to Charlotte was no exception. I now have an ambitious and aggressive roadmap for this year, both personally and professionally, and it includes everything from Dancer improvements to new service
offerings and products. Most of 2017 felt wasted to me, and it’s important to me to make amends to myself in the coming year.</p><p>So until my next post, keep on rocking in the free world, and wherever else you may be! Good fortune to you in 2018.</p>",
    can_like                JSON::PP::Boolean  {
        public methods (0)
        private methods (1) : __ANON__
        internals: 0
    },
    can_reblog              var{can_like},
    can_reply               var{can_like},
    can_send_in_message     JSON::PP::Boolean  {
        public methods (0)
        private methods (1) : __ANON__
        internals: 1
    },
    date                    "2018-01-03 03:09:37 GMT",
    display_avatar          var{can_send_in_message},
    format                  "html",
    id                      169249807334,
    is_blocks_post_format   var{can_like},
    note_count              0,
    post_url                "https://crome-plated.tumblr.com/post/169249807334/happy-new-year",
    reblog                  {
        comment     "<p>Welcome to 2018, everyone! Here’s hoping to a happier, healthier, less chaotic year, filled with more kindness, love, and understanding than the world saw in 2017.</p><p>Airline flights are a good source of downtime, and have historically been a great time
for me to brainstorm and reflect, and today’s flight from Chicago-Midway to Charlotte was no exception. I now have an ambitious and aggressive roadmap for this year, both personally and professionally, and it includes everything from Dancer improvements to new service offerings
and products. Most of 2017 felt wasted to me, and it’s important to me to make amends to myself in the coming year.</p><p>So until my next post, keep on rocking in the free world, and wherever else you may be! Good fortune to you in 2018.</p>",
        tree_html   ""
    },
    reblog_key              "FkYf2fNe",
    recommended_color       undef,
    recommended_source      undef,
    short_url               "https://tmblr.co/ZBorxw2Te4nlc",
    slug                    "happy-new-year",
    state                   "published",
    summary                 "Happy New Year!",
    tags                    [
        [0] "crome-plated"
    ],
    timestamp               1514948977,
    title                   "Happy New Year!",
    trail                   [
        [0] {
            blog              {
                active            var{can_send_in_message},
                can_be_followed   var{can_send_in_message},
                name              "crome-plated",
                share_following   var{can_like},
                share_likes       var{can_send_in_message},
                theme             {
                    avatar_shape          "square",
                    background_color      "#FAFAFA",
                    body_font             "Helvetica Neue",
                    header_bounds         "34,1873,1073,26",
                    header_focus_height   1039,
                    header_focus_width    1847,
                    header_full_height    1200,
                    header_full_width     1900,
                    header_image          "https://static.tumblr.com/6308c232392caa1eceefa06e47d16e04/c7qkugd/fVMofo85e/tumblr_static_aaro2fxe9hs8ssskc8o48ks8o.png",
                    header_image_focused  "https://static.tumblr.com/ed1e3165891f3acd325b12f32e625864/c7qkugd/J1Uofo85n/tumblr_static_tumblr_static_aaro2fxe9hs8ssskc8o48ks8o_focused_v3.jpg",
                    header_image_scaled   "https://static.tumblr.com/6308c232392caa1eceefa06e47d16e04/c7qkugd/fVMofo85e/tumblr_static_aaro2fxe9hs8ssskc8o48ks8o_2048_v2.png",
                    header_stretch        var{can_send_in_message},
                    link_color            "#529ECC",
                    show_avatar           var{can_send_in_message},
                    show_description      var{can_send_in_message},
                    show_header_image     var{can_send_in_message},
                    show_title            var{can_send_in_message},
                    title_color           "#444444",
                    title_font            "Gibson",
                    title_font_weight     "bold"
                }
            },
            content           "<p>Welcome to 2018, everyone! Here&rsquo;s hoping to a happier, healthier, less chaotic year, filled with more kindness, love, and understanding than the world saw in 2017.</p><p>Airline flights are a good source of downtime, and have historically
been a great time for me to brainstorm and reflect, and today&rsquo;s flight from Chicago-Midway to Charlotte was no exception. I now have an ambitious and aggressive roadmap for this year, both personally and professionally, and it includes everything from Dancer improvements to
new service offerings and products. Most of 2017 felt wasted to me, and it&rsquo;s important to me to make amends to myself in the coming year.</p><p>So until my next post, keep on rocking in the free world, and wherever else you may be! Good fortune to you in 2018.</p>",
            content_raw       "<p>Welcome to 2018, everyone! Here’s hoping to a happier, healthier, less chaotic year, filled with more kindness, love, and understanding than the world saw in 2017.</p><p>Airline flights are a good source of downtime, and have historically been a
great time for me to brainstorm and reflect, and today’s flight from Chicago-Midway to Charlotte was no exception. I now have an ambitious and aggressive roadmap for this year, both personally and professionally, and it includes everything from Dancer improvements to new service offerings
and products. Most of 2017 felt wasted to me, and it’s important to me to make amends to myself in the coming year.</p><p>So until my next post, keep on rocking in the free world, and wherever else you may be! Good fortune to you in 2018.</p>",
            is_current_item   var{can_send_in_message},
            is_root_item      var{can_send_in_message},
            post              {
                id   169249807334
            }
        }
    ],
    type                    "text"
```

Step 2: Transform Content

```
sub get_body( $post ) {
  my $body = $post→{ body };

  $body =~ s/<a href="(.+?)" target="_blank">(.+?)<\/a>/\[$2\]\
($1\)/g;

  $body = HTML::FormatText→new(
    leftmargin ⇒ 0, rightmargin ⇒ 72 )→format_string( $body );

  return unidecode( $body );
}
```

Step 3: Write Markdown

```
foreach my $post ( @posts ) {
  # Create blog file
  open my $blog_fh, '>', "scraped_content/$filename"
    or die "Can't open $filename: $!";
  print $blog_fh qq{---
title: $title
hero_image: "hero.jpg"
tags: $tags
date: $date
---

$body};

    close $blog_fh;
    say "done!";
}
```

Step 4: Profit!

(there are a few more steps than this in reality)

All the Code!

https://gist.github.com/cromedome/1d4546da87ff4f6acb9ead2c514ce7a3

The End Result

https://cromedome.net/

Questions?

Thank you!