

Introduction to Web Scraping with R

Inspecting the HTML Tree

A word cloud visualization where the size of each word represents its frequency or importance. The most prominent words are 'elements', 'document', 'the', 'xhtml', 'element', 'markup', and 'attributes'. Other visible words include 'tag', 'instead', 'css', 'attribute', 'browser', 'text', 'may', 'use', 'can', 'attributes', 'type', 'content', 'xml', 'web', 'based', 'sgml', 'deprecated', 'used', 'tags', 'example', 'documents', and 'specification'.

Simon Munzert | IPSDS

Browsing vs. scraping

Browsing vs. scraping the web

Browsing vs. scraping the web

Using your browser to access webpages

1. you click on a link, enter a URL, run a Google query, etc.
2. browser/your machine sends request to server that hosts website
3. server returns resource (often an HTML document)
4. browser interprets HTML and renders it in a nice fashion



Browsing vs. scraping the web

Using your browser to access webpages

1. you click on a link, enter a URL, run a Google query, etc.
2. browser/your machine sends request to server that hosts website
3. server returns resource (often an HTML document)
4. browser interprets HTML and renders it in a nice fashion



Using R to access webpages

1. you manually specify a resource
2. R/your machine sends request to server that hosts website
3. server returns resource
4. R parses HTML, but does not render it in a nice fashion
5. it's up to you to tell R what content to extract



Interacting with your browser

On web browsers

- modern browsers are complex pieces of software that take care of multiple operations while you browse the web
- common operations: retrieve resources, render and display information, provide interface for user-webpage interaction
- although your goal is to automate web data retrieval, your browser is an important tool in web scraping workflow



Interacting with your browser

The use of browsers for web scraping

- give you an intuitive impression of the architecture of a webpage
- allow you to inspect the source code
- let you construct XPath/CSS selector expressions with plugins
- render dynamic web content (JavaScript interpreter)

Interacting with your browser

The use of browsers for web scraping

- give you an intuitive impression of the architecture of a webpage
- allow you to inspect the source code
- let you construct XPath/CSS selector expressions with plugins
- render dynamic web content (JavaScript interpreter)

A note on browser differences

- inspecting the source code (as shown on the following slides) works more or less identically in **Chrome** and **Firefox**
- in **Safari**, go to → **Preferences**, then → **Advanced** and select "Show Develop menu in menu bar". This unlocks the "Show Page Source" option and the Web Developer Tools

Inspecting the HTML source code

Inspecting the HTML source code

Example

- browser: Google Chrome
- source: https://en.wikipedia.org/wiki/List_of_tallest_buildings

The screenshot shows a web browser window displaying the Wikipedia article 'List of tallest buildings'. The page title is 'List of tallest buildings' and it is described as 'From Wikipedia, the free encyclopedia'. The main content discusses the ranking of skyscrapers by height, noting that non-building structures like towers are excluded. A sidebar contains a 'Contents' table of contents with 10 items, including 'Ranking criteria and alternatives' at the top. Below the table of contents is a section titled 'Ranking criteria and alternatives' which provides information about the Council on Tall Buildings and Urban Habitat (CTBUH) and its standards for building measurement. To the right of the main content area is a large image of the Burj Khalifa in Dubai, labeled as the world's tallest building at 828 meters (2,717 ft). The image also includes a caption stating that the Burj Khalifa has been classified as a 'Megatall'.

Click to go forward, hold to see history | Log in

Secure | https://en.wikipedia.org/wiki/List_of_tallest_buildings

Not logged in | [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#) Read View source View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book

List of tallest buildings

From Wikipedia, the free encyclopedia

This list of tallest buildings in the world ranks skyscrapers by height. Only buildings with continuously occupiable floors are included, thus non-building structures, including towers, are not included. (See [List of tallest buildings and structures](#).)

Contents [hide]

- 1 Ranking criteria and alternatives
- 2 Tallest buildings in the world (350 m+)
- 3 Photo gallery
- 4 Alternative measurements
 - 4.1 Height to pinnacle (highest point)
- 5 Buildings under construction
- 6 List by continent
- 7 See also
- 8 Notes
- 9 References
- 10 External links

Ranking criteria and alternatives

The international non-profit organization [Council on Tall Buildings and Urban Habitat](#) (CTBUH) was formed in 1969 and announces the title of 'The World's Tallest Building' and sets the standards by which buildings are measured. It maintains a list of the 100 tallest completed buildings in the world.^[3] The organization currently ranks [Burj Khalifa](#) in Dubai as the tallest at 828 m (2,717 ft).^[3] The CTBUH only recognizes buildings that are complete, however, and some buildings listed within these list articles are not considered complete by the CTBUH.

The 828-metre (2,717 ft) tall [Burj Khalifa](#) in [Dubai](#) has been the world's tallest building since 2008.^[1] The Burj Khalifa has been classified as a [Megatall](#).^[2]

Inspecting the HTML source code

Example

- browser: Google Chrome
- source: https://en.wikipedia.org/wiki/List_of_tallest_buildings
- right-click on page

W List of tallest buildings - Wikipedia

Secure | https://en.wikipedia.org/wiki/List_of_tallest_buildings

Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search Wikipedia

List of tallest buildings

From Wikipedia, the free encyclopedia

This list of tallest buildings in the world ranks skyscrapers by height. Only buildings with continuously occupiable floors are included, thus non-building structures, including towers, are not included. (See List of tallest buildings and structures.)

Contents [hide]

- 1 Ranking criteria and alternatives
- 2 Tallest buildings in the world (350 m+)
- 3 Photo gallery
- 4 Alternative measurements
 - 4.1 Height to pinnacle (highest point)
- 5 Buildings under construction
- 6 List by continent
- 7 See also
- 8 Notes
- 9 References
- 10 External links

Back
Forward
Reload
Save As...
Print...
Cast...
Translate to English
View Page Source
Inspect

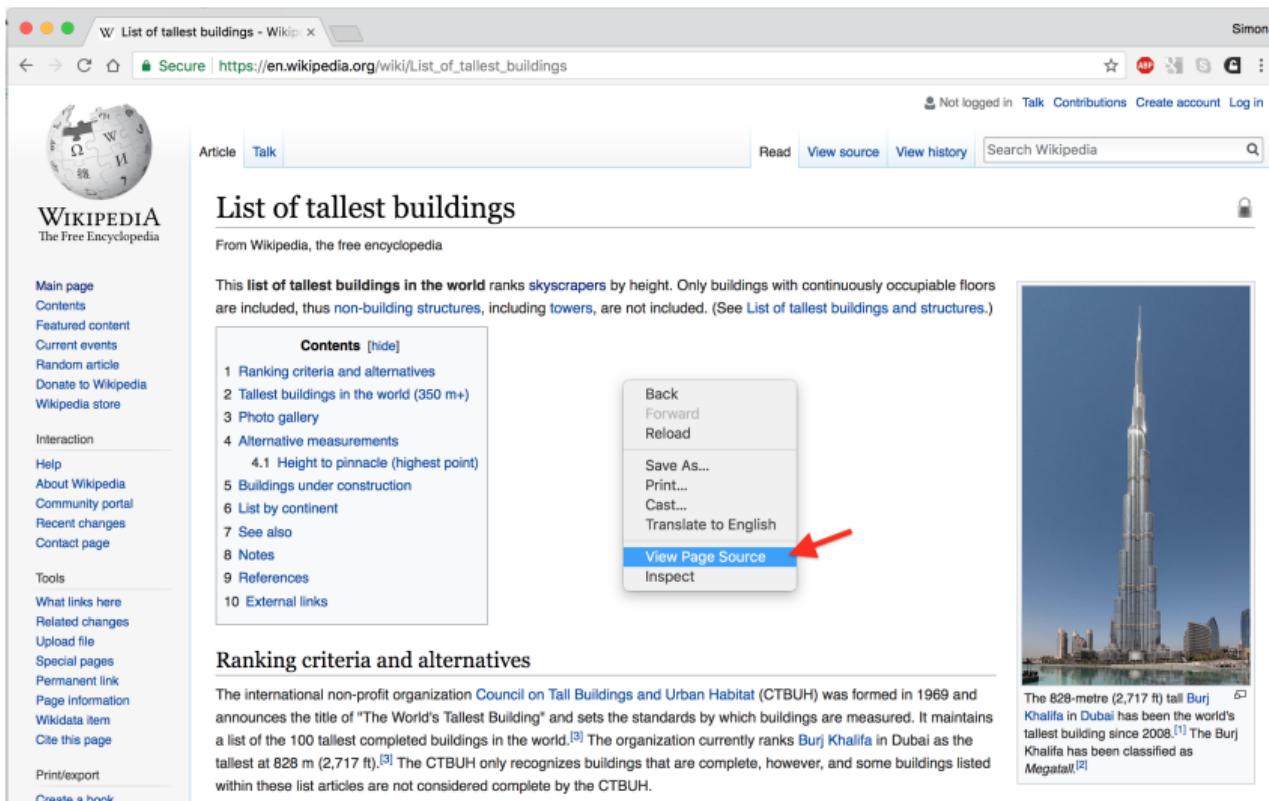


The 828-metre (2,717 ft) tall Burj Khalifa in Dubai has been the world's tallest building since 2008.^[1] The Burj Khalifa has been classified as Megatall.^[2]

Inspecting the HTML source code

Example

- browser: Google Chrome
- source: https://en.wikipedia.org/wiki/List_of_tallest_buildings
- right-click on page
- select "View Page Source"



The screenshot shows a web browser window displaying the Wikipedia article 'List of tallest buildings'. The browser's address bar shows the URL https://en.wikipedia.org/wiki/List_of_tallest_buildings. The page content includes a sidebar with navigation links like 'Main page', 'Contents', and 'Tallest buildings in the world (350 m+)'. To the right is a large image of the Burj Khalifa. A context menu is open, with the 'View Page Source' option highlighted in blue and marked with a red arrow.

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book

Article Talk Read View source View history Search Wikipedia

List of tallest buildings

From Wikipedia, the free encyclopedia

This list of tallest buildings in the world ranks skyscrapers by height. Only buildings with continuously occupiable floors are included, thus non-building structures, including towers, are not included. (See [List of tallest buildings and structures](#).)

Contents [hide]

- 1 Ranking criteria and alternatives
- 2 Tallest buildings in the world (350 m+)
- 3 Photo gallery
- 4 Alternative measurements
 - 4.1 Height to pinnacle (highest point)
- 5 Buildings under construction
- 6 List by continent
- 7 See also
- 8 Notes
- 9 References
- 10 External links

Back
Forward
Reload

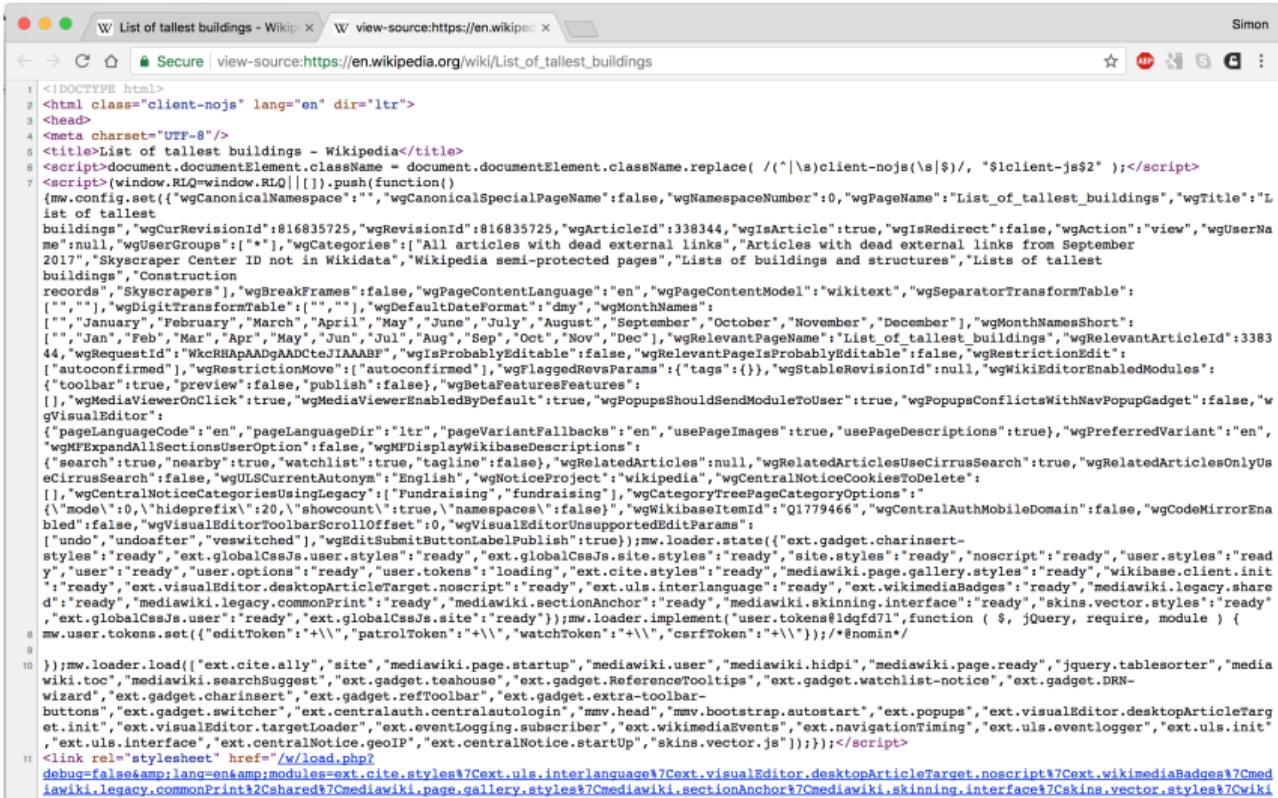
Save As...
Print...
Cast...
Translate to English
View Page Source (highlighted)
Inspect

Ranking criteria and alternatives

The international non-profit organization [Council on Tall Buildings and Urban Habitat](#) (CTBUH) was formed in 1969 and announces the title of "The World's Tallest Building" and sets the standards by which buildings are measured. It maintains a list of the 100 tallest completed buildings in the world.^[3] The organization currently ranks [Burj Khalifa](#) in Dubai as the tallest at 828 m (2,717 ft).^[3] The CTBUH only recognizes buildings that are complete, however, and some buildings listed within these list articles are not considered complete by the CTBUH.

The 828-metre (2,717 ft) tall [Burj Khalifa](#) in [Dubai](#) has been the world's tallest building since 2008.^[3] The Burj Khalifa has been classified as [Megatall](#).^[2]

Inspecting the HTML source code



The screenshot shows a browser window with two tabs: "List of tallest buildings - Wikipedia" and "view-source:https://en.wikipedia.org/wiki/List_of_tallest_buildings". The main content area displays the raw HTML source code of the Wikipedia page. The code is extensive, starting with the DOCTYPE declaration and including HTML, head, and body sections, along with various script and style elements. The code is color-coded for syntax highlighting.

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>List of tallest buildings - Wikipedia</title>
<script>document.documentElement.className = document.documentElement.className.replace( /(^\s)client-nojs(\s$/), '$1client-js$2' );</script>
<script>(window.RLQ>window.RLQ||[]).push(function()
{mw.config.set({wgCanonicalNamespace:"",wgCanonicalSpecialPageName:false,"wgNamespaceNumber":0,"wgPageName":"List_of_tallest_buildings","wgTitle":"List of tallest
buildings","wgCurRevisionId":816835725,"wgRevisionId":816835725,"wgArticleId":338344,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["All articles with dead external links","Articles with dead external links from September 2017","Skyscrapers Center ID not in Wikidata","Wikimedia semi-protected pages","Lists of buildings and structures","Lists of tallest
buildings","Construction
records","Skyscrapers"],"wgBreakFrames":false,"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","May","June","July","August","September","October","November","December"],"wgMonthNamesShort":["","Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"],"wgRelevantPageName":"List_of_tallest_buildings","wgRelevantArticleId":338344,"wgRequestPageID":"WkcRHApAAdqAADcteJIAAAE","wgIsProbablyEditable":false,"wgRelevantPageIsProbablyEditable":false,"wgRestrictionEdit":false,"wgRestrictionMove":false,"wgFlaggedRevParams":{"tags":{}}, "wgStableRevisionId":null,"wgWikiEditorEnabledModules":[]}, "wgMediaViewerOnClick":true, "wgMediaViewerEnabledByDefault":true, "wgPopupsShouldSendModuleToUser":true, "wgPopupsConflictsWithNavPopupGadget":false, "wgVisualEditor":true, "wgPageLanguageCode": "en", "pageLanguageDir": "ltr", "pageVariantFallbacks": "en", "usePageImage": true, "usePageDescriptions": true, "wgPreferredVariant": "en", "wgFPExcludeAllSectionsUserOption": false, "wgFDIsDisplayWikibaseDescriptions": false, "wgSearch": true, "nearby": true, "watchlist": true, "tagline": false, "wgRelatedArticles": null, "wgRelatedArticlesUseCirrusSearch": true, "wgRelatedArticlesOnlyUseCirrusSearch": false, "wgULCurrentAutonym": "English", "wgNoticeProject": "wikipedia", "wgCentralNoticeCookiesToDelete": [], "wgCentralNoticeCategoriesUsingLegacy": ["Fundraising", "fundraising"], "wgCategoryFreePageCategoryOptions": "(\\\"mode\\\":\\\"hideprefix\\\":20,\\\"showcount\\\":true,\\\"namespaces\\\":false)", "wgWikibaseItemID": "Q177946", "wgCentralAuthMobileDomain": false, "wgCodeMirrorEnabled": false, "wgVisualEditorToolbarScrollOffset": 0, "wgVisualEditorUnsupportedEditParams": ["undo", "undoafter", "weswitched"], "wgEditSubmitButtonLabel": true, "mw.loader.state": {"ext.gadget.charinserter": "ready", "ext.globalCssJs.user.styles": "ready", "ext.globalcssjs.site.styles": "ready", "site.styles": "ready", "noscript": "ready", "user.styles": "ready", "user": "ready", "user.options": "ready", "user.tokens": "loading", "ext.cite.styles": "ready", "mediawiki.page.gallery.styles": "ready", "wikibase.client.init": "ready", "ext.visualEditor.desktopArticleTarget.noscript": "ready", "ext.uls.interlanguage": "ready", "ext.wikimediaBadges": "ready", "mediawiki.legacy.shareD": "ready", "mediawiki.legacy.commonPrint": "ready", "mediawiki.sectionAnchor": "ready", "mediawiki.skinning.interface": "ready", "skins.vector.styles": "ready", "ext.globalCssJs.user": "ready", "ext.globalcssjs.site": "ready"}, "mw.loader.implement": "user.tokens#lcfgd71", "function": "$, jQuery, require, module" } }, mw.user.tokens.set({ "editToken": "+\\\", "patrolToken": "+\\\", "watchToken": "+\\\", "csrfToken": "+\\\" }), /*nominate*/ ); mw.loader.load(['ext.cite.ally', 'site', 'mediawiki.page.startup', 'mediawiki.user', 'mediawiki.hippi', 'mediawiki.page.ready', 'jquery.tablesorter', 'mediawiki.toc', 'mediawiki.searchSuggest', 'ext.gadget.teahouse', 'ext.gadget.ReferenceTooltips', 'ext.gadget.watchlist-notice', 'ext.gadget.DRN-wizard', 'ext.gadget.charinserter', 'ext.gadget.refToolbar', 'ext.gadget.extra-toolbar-buttons', 'ext.gadget.switcher', 'ext.centralauth.centralautologin', 'mmv.head', 'mmv.bootstrap.autostart', 'ext.popups', 'ext.visualEditor.desktopArticleFarest.init', 'ext.visualEditor.targetLoader', 'ext.eventLogging.subscriber', 'ext.wikimediaEvents', 'ext.navigationTiming', 'ext.uls.init', 'ext.uls.interface', 'ext.centralNotice.geoIP', 'ext.centralNotice.startUp', 'skins.vector.js' ]);});</script>
<link rel="stylesheet" href="/load.php?
debug=false&lang=en&modules=ext.cite.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimediaBadges%7Cmediawiki.legacy.commonPrint%2Cshared%7Cmediawiki.page.gallery.styles%7Cmediawiki.sectionAnchor%7Cmediawiki.skinning.interface%7Cskins.vector.styles%7Cwiki
```

Example

- HTML (and JavaScript) code can be ugly...

Inspecting the HTML source code

Example

- HTML (and JavaScript) code can be ugly...
- but looking more closely, we can find the displayed information

```
W List of tallest buildings - Wikipedia W view-source:https://en.wikipedia.org/... Secure | view-source:https://en.wikipedia.org/wiki/List_of_tallest_buildings
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-List_of_tallest_buildings rootpage-List_of_tallest_buildings skin-vector action-view">
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
    <div id="content" class="mw-body" role="main">
        <a id="top"></a>
        <div id="siteNotice" class="mw-body-content"><!-- CentralNotice --></div><div class="mw-indicators mw-body-content">
<div id="mw-indicator-pp-default" class="mw-indicator"><a href="/wiki/Wikipedia:Protection_policy#semi" title="This article is semi-protected until March 17, 2019."></a></div>
</div>
<h1 id="firstHeading" class="firstHeading" lang="en">List of tallest buildings</h1>
    <div id="bodyContent" class="mw-body-content">
        <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia</div>
        <div id="jump-to-nav" class="mw-jump">
            Jump to: <a href="#mw-head">navigation</a>, <a href="#p-search">search</a>
        </div>
        <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr"><div class="mw-parser-output"><div class="thumb tright">
<div class="tumbinner" style="width:222px;"><a href="/wiki/File:Burj_Khalifa.jpg" class="image"></a>
<div class="thumbcaption">
<div class="magnify"><a href="/wiki/File:Burj_Khalifa.jpg" class="internal" title="Enlarge"></a></div>
The 828-metre (2,7176/160 ft) tall <a href="/wiki/Burj_Khalifa" title="Burj Khalifa">Burj Khalifa</a> in <a href="/wiki/Dubai" title="Dubai">Dubai</a> has been the world's tallest building since 2008.<sup id="cite_ref-1" class="reference"><a href="#cite_note-1" title="Reference">[1]</a></sup> The Burj Khalifa has been classified as <i>Megatall</i>. <sup id="cite_ref-CTBUH-2" class="reference"><a href="#cite_note-CTBUH-2" title="Reference">[2]</a></sup></div>
</div>
<div class="thumb tright">
<div class="tumbinner" style="width:487px;"><a href="/wiki/File:Tallest_buildings_in_the_world.png" class="image"></a>
<div class="thumbcaption">
<div class="magnify"><a href="/wiki/File:Tallest_buildings_in_the_world.png" class="internal" title="Enlarge"></a></div>
Schematic of the tallest buildings in the world in 2015.</div>
</div>
</div>
<p>This <b>list of tallest buildings in the world</b> ranks <a href="/wiki/Skyscraper" title="Skyscraper">skyscrapers</a> by height. Only buildings with continuously occupiable floors are included, thus <a href="/wiki/Nonbuilding_structure" title="Nonbuilding structure">non-building structures</a>, including <a href="/wiki/Tower" title="Tower">towers</a>, are not included. (See <a href="/wiki/List_of_tallest_buildings_and_structures" title="List of tallest buildings and structures">List of tallest buildings and structures</a>.)</p>
```

Inspecting the live HTML tree

Inspecting individual elements and the live HTML tree

Example

- another way of inspecting the HTML code is to use the **Web Developer Tools**
- to that end, do:
- right-click on element of interest
- select "Inspect"

The screenshot shows a web browser window displaying the 'List of tallest buildings' page on Wikipedia. The URL is https://en.wikipedia.org/wiki/List_of_tallest_buildings. The page title is 'List of tallest buildings'. A context menu is open over the first item in the list, with the 'Inspect' option highlighted by a red arrow. The menu also includes options like 'View Page Source' and 'Contents [hide]'. To the right of the menu, there is a large image of the Burj Khalifa and some text about it.

This list of tallest buildings in the world ranks skyscrapers by height. Only buildings with continuously occupiable floors are included, thus non-building structures, including towers, are not included. (See List of tallest buildings and structures.)

Contents [hide]

- 1 Ranking criteria and alternatives
- 2 Tallest buildings in the world
- 3 Photo gallery
- 4 Alternative measurements
 - 4.1 Height to pinnacle (highest point)
- 5 Buildings under construction
- 6 List by continent
- 7 See also
- 8 Notes
- 9 References
- 10 External links

Back
Forward
Reload
Save As...
Print...
Cast...
Translate to English
View Page Source
Inspect

Ranking criteria and alternatives

The international non-profit organization Council on Tall Buildings and Urban Habitat (CTBUH) was formed in 1969 and announces the title of "The World's Tallest Building" and sets the standards by which buildings are measured. It maintains a list of the 100 tallest completed buildings in the world.^[3] The organization currently ranks Burj Khalifa in Dubai as the tallest at 828 m (2,717 ft).^[3] The CTBUH only recognizes buildings that are complete, however, and some buildings listed within these list articles are not considered complete by the CTBUH.

In 2008, as a response to the dispute as to whether the Defense Tower or

The 828-metre (2,717 ft) tall Burj Khalifa in Dubai has been the world's tallest building since 2008.^[1] The Burj Khalifa has been classified as Megatall.^[2]

Inspecting individual elements and the live HTML tree

Example

- the Web Developer Tools window pops up
 - corresponding part in the HTML tree is highlighted
 - you can hover over other parts of the code to get an instant view of the page at that position
 - click on arrows to expand/hide tags

Summary

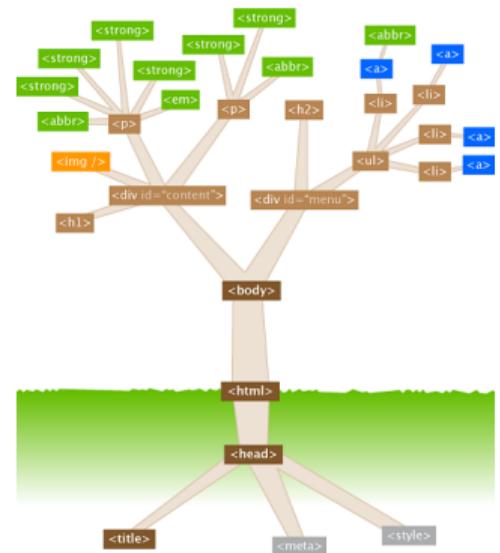
Summary

When to inspect the complete page source

- check whether data is in static source code
- for small HTML files: understand structure
- count tables

When to inspect individual elements using the Web Inspector Tools

- almost always
- particularly useful to construct XPath/CSS selector expressions
- to monitor dynamic changes in the DOM tree (see later)



Source:
<http://watershedcreative.com/naked/html-tree.html>