Analyzing Wikipedia Pages

In this project, we'll work with data scraped from Wikipedia. Volunteer content contributors and editors maintain Wikipeda by continuously improving content. Anyone can edit Wikipedia, and because Wikipedia is crowdsourced, it has rapidly assembed a huge library of articles.

We'll implement a simplified version of the grep command-line utility to search for data in 54 megabytes worth of articles. The grep command and the grep utility essentially allows searching for textual data in all files from a given directory.

Articles were saved using the last component of their URLs. A page on Wikipedia has the URL structure https://en.wikipedia.org/wiki/Yarkant_County. When saving the article with the URL provided, we'd save it to the file Yarkant_County.html. All the data files are located in the wiki folder. Note that the files are in raw HTML, but we can treat the files like plain-text.

Our main goals will be the following:

- Search for all occurrences of a string in all of the files.
- Provide a case-insensitive option to the search.
- Refine the result by providing the specific locations of the files.

List all files in the wiki folder

We can create a list with the names of all files in the wiki folder using the os.listdir() function.

In [67]:
```
import os

file_names = os.listdir("wiki")
len(file_names)
```

Out[67]:
```
999
```

Read the first file

Let's read the first file and print its contents. We need to join the name of the file with the wiki folder. We can do this using the os.path.join() function.

In [68]:
```
with open(os.path.join("wiki", file_names[0])) as f:
    print(f.read())
```

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>Bay of Concepción - 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({"wgCanonicalNamespac
e":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"Bay_of_Concep
ción","wgTitle":"Bay of Concepción","wgCurRevisionId":647460156,"wgRevisionId":64746015
6,"wgArticleId":16044270,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUser
Name":null,"wgUserGroups":["*"],"wgCategories":["Coordinates on Wikidata","All stub articl
es","Landforms of Bío Bío Region","Bays of Chile","Bío Bío Region geography stubs"],"wgBre
akFrames":false,"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgSeparatorT
ransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMont
```

hNames":["","January","February","March","April","May","June","July","August","Septembe
r","October","November","December"],"wgMonthNamesShort":["","Jan","Feb","Mar","Apr","Ma
y","Jun","Jul","Aug","Sep","Oct","Nov","Dec"],"wgRelevantPageName":"Bay_of_Concepción","wg
RelevantArticleId":16044270,"wgRequestId":"WKq3wgpAAEIAAMFPZFwAAABQ","wgIsProbablyEditabl
e":true,"wgRestrictionEdit":[],"wgRestrictionMove":[],"wgFlaggedRevsParams":{"tags":{}},"w
gStableRevisionId":null,"wgWikiEditorEnabledModules":{"toolbar":true,"dialogs":true,"previ
ew":false,"publish":false},"wgBetaFeaturesFeatures":[],"wgMediaViewerOnClick":true,"wgMedi
aViewerEnabledByDefault":true,"wgVisualEditor":{"pageLanguageCode":"en","pageLanguageDi
r":"ltr","usePageImages":true,"usePageDescriptions":true},"wgPreferredVariant":"en","wgMFD
isplayWikibaseDescriptions":{"search":true,"nearby":true,"watchlist":true,"tagline":tru
e},"wgRelatedArticles":null,"wgRelatedArticlesBetaFeatureEnabled":false,"wgRelatedArticles
UseCirrusSearch":true,"wgRelatedArticlesOnlyUseCirrusSearch":false,"wgULSCurrentAutony
m":"English","wgNoticeProject":"wikipedia","wgCentralNoticeCookiesToDelete":[],"wgCentralN
oticeCategoriesUsingLegacy":["Fundraising","fundraising"],"wgCategoryTreePageCategoryOptio
ns":"{\"mode\":0,\"hideprefix\":20,\"showcount\":true,\"namespaces\":false}","wgCoordinate
s":{"lat":-36.683333333333,"lon":-73.033333333333},"wgWikibaseItemId":"Q4874197","wgCentra
lAuthMobileDomain":false,"wgVisualEditorToolbarScrollOffset":0,"wgEditSubmitButtonLabelPub
lish":false});mw.loader.state({"ext.globalCssJs.user.styles":"ready","ext.globalCssJs.sit
e.styles":"ready","site.styles":"ready","noscript":"ready","user.styles":"ready","user":"r
eady","user.options":"loading","user.tokens":"loading","ext.cite.styles":"ready","wikibas
e.client.init":"ready","ext.visualEditor.desktopArticleTarget.noscript":"ready","ext.uls.i
nterlanguage":"ready","ext.wikimediaBadges":"ready","mediawiki.legacy.shared":"ready","med
iawiki.legacy.commonPrint":"ready","mediawiki.sectionAnchor":"ready","mediawiki.skinning.i
nterface":"ready","skins.vector.styles":"ready","ext.globalCssJs.user":"ready","ext.global
CssJs.site":"ready"});mw.loader.implement("",function($,jQuery,require,mo
dule){mw.user.options.set({"variant":"en"});});mw.loader.implement("",fun
ction ($, jQuery, require, module) {
mw.user.tokens.set({"editToken":"+\\","patrolToken":"+\\","watchToken":"+\\","csrfToke
n":"+\\"});/*@nomin*/

});mw.loader.load(["ext.cite.a11y","mediawiki.action.view.postEdit","site","mediawiki.pag
e.startup","mediawiki.user","mediawiki.hidpi","mediawiki.page.ready","mediawiki.legacy.wik
ibits","mediawiki.searchSuggest","ext.gadget.teahouse","ext.gadget.ReferenceTooltips","ex
t.gadget.watchlist-notice","ext.gadget.DRN-wizard","ext.gadget.charinsert","ext.gadget.ref
Toolbar","ext.gadget.extra-toolbar-buttons","ext.gadget.switcher","ext.gadget.featured-art
icles-links","ext.centralauth.centralautologin","mmv.head","mmv.bootstrap.autostart","ext.
visualEditor.desktopArticleTarget.init","ext.visualEditor.targetLoader","ext.eventLogging.
subscriber","ext.wikimediaEvents","ext.navigationTiming","ext.uls.eventlogger","ext.uls.in
it","ext.uls.interface","ext.quicksurveys.init","ext.centralNotice.geoIP","ext.centralNoti
ce.startUp","skins.vector.js"]);});</script>
<link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=ext.cite.styl
es%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimedi
aBadges%7Cmediawiki.legacy.commonPrint%2Cshared%7Cmediawiki.sectionAnchor%7Cmediawiki.skin
ning.interface%7Cskins.vector.styles%7Cwikibase.client.init&only=styles&skin=vecto
r"/>
<script async="" src="/w/load.php?debug=false&lang=en&modules=startup&only=scr
ipts&skin=vector"></script>
<meta name="ResourceLoaderDynamicStyles" content=""/>
<link rel="stylesheet" href="/w/load.php?debug=false&lang=en&modules=site.styles&a
mp;only=styles&skin=vector"/>
<meta name="generator" content="MediaWiki 1.29.0-wmf.12"/>
<meta name="referrer" content="origin-when-cross-origin"/>
<meta property="og:image" content="https://upload.wikimedia.org/wikipedia/commons/9/9d/Txu
-oclc-224571178-sj18-04-quiriquina.jpg"/>
<link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Bay_o
f_Concepci%C3%B3n"/>
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?
title=Bay_of_Concepci%C3%B3n&action=edit"/>
<link rel="edit" title="Edit this page" href="/w/index.php?title=Bay_of_Concepci%C3%B3n&am
p;action=edit"/>
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.p
hp" title="Wikipedia (en)"/>
<link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=r
sd"/>
<link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="canonical" href="https://en.wikipedia.org/wiki/Bay_of_Concepci%C3%B3n"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />

</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-Bay_of_Concepción rootpage-Bay_of_Concepción skin-vector action-view">

<!-- CentralNotice -->

<h1 id="firstHeading" class="firstHeading" lang="en">Bay of Concepción</h1>

From Wikipedia, the free encyclopedia

Jump to: navigation, search

Region of BioBio

The Bay of Concepción is a natural bay on the coast of the Province of Concepción in the Bío Bío Region of Chile. Within the bay are many of the most important ports of the region and the country, among them Penco, Talcahuano, and Lirquén.
Quiriquina Island, located to the north in the mouth of the bay provides a windbreak. The island creates two entrances to the bay: Boca Chica and Boca Grande. Boca Chica, between Quriquina Island and the Peninsula of Tumbes, measures 2 km wide and in its narrower part 1,500 metres, with shoals to the sides and although water depth is 15 metres, the passage of large ships is reduced to 400 metres.^{[1]} Boca Grande, is 5 km wide, with depths of 35 metres, which makes it comodious for large vessels.^{[1]}
The sector of the bay where the Port of Talcahuano is located is known as the Bay of Talcahuano, and is protected by the Peninsula of Tumbes and Quiriquina Island.^{[1]}
<h2>References[edit]</h2>
<ol class="references">
<li id="cite_note-Espinoza.2C_Enrique_1897-1">^ ^a ^b <a href="#cite_ref-Espinoza.2C_

Enrique_1897_1-2">^c Espinoza, Enrique; 1897. Geografía Descriptiva de la República de Chile. Cuarta edición, Imprenta y encuadernación Barcelona, Santiago, Chile.

Coordinates: 36°41'S 73°02'W / 36.683°S 73.033°W / -36.683; -73.033

<table class="metadata plainlinks stub" role="presentation" style="background:transparent">
<tr>
<td></td>
<td>This Bío Bío Region location article is a stub. You can help Wikipedia by expanding it.

<li class="nv-view"><abbr title="View this template">v</abbr>
<li class="nv-talk"><abbr title="Discuss this template">t</abbr>
<li class="nv-edit"><abbr title="Edit this template">e</abbr>

</td>
</tr>
</table>

<!--
NewPP limit report
Parsed by mw1251
Cached time: 20170208034214
Cache expiry: 2592000
Dynamic content: false
CPU time usage: 0.040 seconds
Real time usage: 0.057 seconds
Preprocessor visited node count: 98/1000000
Preprocessor generated node count: 0/1500000
Post-expand include size: 4419/2097152 bytes
Template argument size: 0/2097152 bytes
Highest expansion depth: 3/40
Expensive parser function count: 0/500
Lua time usage: 0.016/10.000 seconds
Lua memory usage: 813 KB/50 MB
-->
<!--
Transclusion expansion time report (%,ms,calls,template)
100.00%   38.984   1 -total
 73.51%   28.658   1 Template:Coord
 26.14%   10.189   1 Template:Biobío-geo-stub
 21.36%   8.328   1 Template:Asbox
-->

<!-- Saved in parser cache with key enwiki:pcache:idhash:16044270-0!*!0!*!*!4!* and timestamp 20170208034214 and revision id 647460156
 -->
<noscript></noscript>

Retrieved from "https://en.wikipedia.org/w/index.php?title=Bay_of_Concepción&oldid=647460156"

Categories: Landforms of Bío Bío Region
Bays of Chile
Bío Bío Region geography stubs Hidden categories: Coordinates on Wikidata
All stub articles

<h2>Navigation menu</h2>

<h3 id="p-personal-label">Personal tools</h3>

<li id="pt-anonuserpage">Not logged in
<li id="pt-anontalk">Talk
<li id="pt-anoncontribs">Contributions
<li id="pt-createaccount">Create account
<li id="pt-login">Log in

<h3 id="p-namespaces-label">Namespaces</h3>

<li id="ca-nstab-main" class="selected">Article

<li id="ca-talk">Talk

<h3 id="p-variants-label">
Variants

</h3>

```
                                                <div class="menu">
                                                        <ul>

</ul>
                                                </div>
                                        </div>
                                                                        </div>
                                <div id="right-navigation">
                                                                                <div id="p
-views" role="navigation" class="vectorTabs" aria-labelledby="p-views-label">
                                                <h3 id="p-views-label">Views</h3>
                                                <ul>

<li id="ca-view" class="selected"><span><a href="/wiki/Bay_of_Concepci%C3%B3n" >Read</a></
span></li>

<li id="ca-edit"><span><a href="/w/index.php?title=Bay_of_Concepci%C3%B3n&amp;action=edit"
title="Edit this page [e]" accesskey="e">Edit</a></span></li>

<li id="ca-history" class="collapsible"><span><a href="/w/index.php?title=Bay_of_Concepci%
C3%B3n&amp;action=history"  title="Past revisions of this page [h]" accesskey="h">View his
tory</a></span></li>

</ul>
                                        </div>
                                                                                <div id="p
-cactions" role="navigation" class="vectorMenu emptyPortlet" aria-labelledby="p-cactions-l
abel">
                                                <h3 id="p-cactions-label"><span>More</span
><a href="#"></a></h3>

                                                <div class="menu">
                                                        <ul>

</ul>
                                                </div>
                                        </div>
                                                                                <div id="p
-search" role="search">
                                                <h3>
                                                        <label for="searchInput">Search</l
abel>
                                                </h3>

                                                <form action="/w/index.php" id="searchfor
m">
                                                        <div id="simpleSearch">
                                                        <input type="search" name="search"
placeholder="Search Wikipedia" title="Search Wikipedia [f]" accesskey="f" id="searchInpu
t"/><input type="hidden" value="Special:Search" name="title"/><input type="submit" name="f
ulltext" value="Search" title="Search Wikipedia for this text" id="mw-searchButton" class
="searchButton mw-fallbackSearchButton"/><input type="submit" name="go" value="Go" title
="Go to a page with this exact name if it exists" id="searchButton" class="searchButton"/>
</div>
                                                </form>
                                        </div>
                                                                        </div>
                        </div>
                        <div id="mw-panel">
                                <div id="p-logo" role="banner"><a class="mw-wiki-logo" hre
f="/wiki/Main_Page"  title="Visit the main page"></a></div>
                                                <div class="portal" role="navigation" id
='p-navigation' aria-labelledby='p-navigation-label'>
                        <h3 id='p-navigation-label'>Navigation</h3>

                        <div class="body">
                                                                        <ul>
                                                <li id="n-mainpage-description"><a href="/
wiki/Main_Page" title="Visit the main page [z]" accesskey="z">Main page</a></li><li id="n-
contents"><a href="/wiki/Portal:Contents" title="Guides to browsing Wikipedia">Contents</a
```

```
></li><li id="n-featuredcontent"><a href="/wiki/Portal:Featured_content" title="Featured c
ontent – the best of Wikipedia">Featured content</a></li><li id="n-currentevents"><a href
="/wiki/Portal:Current_events" title="Find background information on current events">Curre
nt events</a></li><li id="n-randompage"><a href="/wiki/Special:Random" title="Load a rando
m article [x]" accesskey="x">Random article</a></li><li id="n-sitesupport"><a href="http
s://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm_source=donate&amp;utm_mediu
m=sidebar&amp;utm_campaign=C13_en.wikipedia.org&amp;uselang=en" title="Support us">Donate
 to Wikipedia</a></li><li id="n-shoplink"><a href="//shop.wikimedia.org" title="Visit the
 Wikipedia store">Wikipedia store</a></li>                                    </ul>
                                                    </div>
                </div>
                        <div class="portal" role="navigation" id='p-interaction' aria-labe
lledby='p-interaction-label'>
                        <h3 id='p-interaction-label'>Interaction</h3>

                        <div class="body">
                                                                <ul>
                                                    <li id="n-help"><a href="/wiki/Help:Conten
ts" title="Guidance on how to use and edit Wikipedia">Help</a></li><li id="n-aboutsite"><a
href="/wiki/Wikipedia:About" title="Find out about Wikipedia">About Wikipedia</a></li><li
 id="n-portal"><a href="/wiki/Wikipedia:Community_portal" title="About the project, what y
ou can do, where to find things">Community portal</a></li><li id="n-recentchanges"><a href
="/wiki/Special:RecentChanges" title="A list of recent changes in the wiki [r]" accesskey
="r">Recent changes</a></li><li id="n-contactpage"><a href="//en.wikipedia.org/wiki/Wikipe
dia:Contact_us" title="How to contact Wikipedia">Contact page</a></li>
</ul>
                                                    </div>
                </div>
                        <div class="portal" role="navigation" id='p-tb' aria-labelledby='p
-tb-label'>
                        <h3 id='p-tb-label'>Tools</h3>

                        <div class="body">
                                                                <ul>
                                                    <li id="t-whatlinkshere"><a href="/wiki/Sp
ecial:WhatLinksHere/Bay_of_Concepci%C3%B3n" title="List of all English Wikipedia pages con
taining links to this page [j]" accesskey="j">What links here</a></li><li id="t-recentchan
geslinked"><a href="/wiki/Special:RecentChangesLinked/Bay_of_Concepci%C3%B3n" rel="nofollo
w" title="Recent changes in pages linked from this page [k]" accesskey="k">Related changes
</a></li><li id="t-upload"><a href="/wiki/Wikipedia:File_Upload_Wizard" title="Upload file
s [u]" accesskey="u">Upload file</a></li><li id="t-specialpages"><a href="/wiki/Special:Sp
ecialPages" title="A list of all special pages [q]" accesskey="q">Special pages</a></li><l
i id="t-permalink"><a href="/w/index.php?title=Bay_of_Concepci%C3%B3n&amp;oldid=647460156"
title="Permanent link to this revision of the page">Permanent link</a></li><li id="t-inf
o"><a href="/w/index.php?title=Bay_of_Concepci%C3%B3n&amp;action=info" title="More informa
tion about this page">Page information</a></li><li id="t-wikibase"><a href="https://www.wi
kidata.org/wiki/Q4874197" title="Link to connected data repository item [g]" accesskey
="g">Wikidata item</a></li><li id="t-cite"><a href="/w/index.php?title=Special:CiteThisPag
e&amp;page=Bay_of_Concepci%C3%B3n&amp;id=647460156" title="Information on how to cite this
page">Cite this page</a></li>                                    </ul>
                                                    </div>
                </div>
                        <div class="portal" role="navigation" id='p-coll-print_export' ari
a-labelledby='p-coll-print_export-label'>
                        <h3 id='p-coll-print_export-label'>Print/export</h3>

                        <div class="body">
                                                                <ul>
                                                    <li id="coll-create_a_book"><a href="/w/in
dex.php?title=Special:Book&amp;bookcmd=book_creator&amp;referer=Bay+of+Concepci%C3%B3n">Cr
eate a book</a></li><li id="coll-download-as-rdf2latex"><a href="/w/index.php?title=Specia
l:Book&amp;bookcmd=render_article&amp;arttitle=Bay+of+Concepci%C3%B3n&amp;returnto=Bay+of+
Concepci%C3%B3n&amp;oldid=647460156&amp;writer=rdf2latex">Download as PDF</a></li><li id
="t-print"><a href="/w/index.php?title=Bay_of_Concepci%C3%B3n&amp;printable=yes" title="Pr
intable version of this page [p]" accesskey="p">Printable version</a></li>
</ul>
                                                    </div>
                </div>
                        <div class="portal" role="navigation" id='p-lang' aria-labelledby
='p-lang-label'>
```

```
                        <h3 id='p-lang-label'>Languages</h3>

                        <div class="body">
                                                                        <ul>
                                                    <li class="interlanguage-link interwiki-es"><a href="https://es.wikipedia.org/wiki/Bah%C3%ADa_de_Concepci%C3%B3n" title="Bahía de Concepción – Spanish" lang="es" hreflang="es" class="interlanguage-link-target">Español</a></li><li class="interlanguage-link interwiki-fr"><a href="https://fr.wikipedia.org/wiki/Baie_de_Concepci%C3%B3n" title="Baie de Concepción – French" lang="fr" hreflang="fr" class="interlanguage-link-target">Français</a></li><li class="interlanguage-link interwiki-ru"><a href="https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%81%D0%B5%D0%BF%D1%81%D1%8C%D0%BE%D0%BD_(%D0%B7%D0%B0%D0%BB%D0%B8%D0%B2)" title="Консепсьон (залив) – Russian" lang="ru" hreflang="ru" class="interlanguage-link-target">Русский</a></li><li class="interlanguage-link interwiki-uk"><a href="https://uk.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%81%D0%B5%D0%BF%D1%81%D1%8C%D0%B9%D0%BE%D0%BD_(%D0%B7%D0%B0%D1%82%D0%BE%D0%BA%D0%B0)" title="Консепсьйон (затока) – Ukrainian" lang="uk" hreflang="uk" class="interlanguage-link-target">Українська</a></li>                                                </ul>
                                <div class='after-portlet after-portlet-lang'><span class="wb-langlinks-edit wb-langlinks-link"><a href="https://www.wikidata.org/wiki/Q4874197#sitelinks-wikipedia" title="Edit interlanguage links" class="wbc-editpage">Edit links</a></span></div>                        </div>
                </div>
                                <div id="footer" role="contentinfo">
                                                                <ul id="footer-info">
                                                                                                <li id="footer-info-lastmod"> This page was last modified on 16 February 2015, at 22:18.</li>
                                                                                                <li id="footer-info-copyright">Text is available under the <a rel="license" href="//en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License">Creative Commons Attribution-ShareAlike License</a><a rel="license" href="//creativecommons.org/licenses/by-sa/3.0/" style="display:none;"></a>;
additional terms may apply.  By using this site, you agree to the <a href="//wikimediafoundation.org/wiki/Terms_of_Use">Terms of Use</a> and <a href="//wikimediafoundation.org/wiki/Privacy_policy">Privacy Policy</a>. Wikipedia® is a registered trademark of the <a href="//www.wikimediafoundation.org/">Wikimedia Foundation, Inc.</a>, a non-profit organization.</li>
                                                                                                                </ul>
                                                                <ul id="footer-places">
                                                                                                <li id="footer-places-privacy"><a href="https://wikimediafoundation.org/wiki/Privacy_policy" class="extiw" title="wmf:Privacy policy">Privacy policy</a></li>
                                                                                                <li id="footer-places-about"><a href="/wiki/Wikipedia:About" title="Wikipedia:About">About Wikipedia</a></li>
                                                                                                <li id="footer-places-disclaimer"><a href="/wiki/Wikipedia:General_disclaimer" title="Wikipedia:General disclaimer">Disclaimers</a></li>
                                                                                                <li id="footer-places-contact"><a href="//en.wikipedia.org/wiki/Wikipedia:Contact_us">Contact Wikipedia</a></li>
                                                                                                <li id="footer-places-developers"><a href="https://www.mediawiki.org/wiki/Special:MyLanguage/How_to_contribute">Developers</a></li>
                                                                                                <li id="footer-places-cookiestatement"><a href="https://wikimediafoundation.org/wiki/Cookie_statement">Cookie statement</a></li>
                                                                                                <li id="footer-places-mobileview"><a href="//en.m.wikipedia.org/w/index.php?title=Bay_of_Concepci%C3%B3n&amp;mobileaction=toggle_view_mobile" class="noprint stopMobileRedirectToggle">Mobile view</a></li>
                                                                                                                </ul>
                                                                                                <ul id="footer-icons" class="noprint">
                                                                                                <li id="footer-copyrightico">
                                                                <a href="https://wikimediafoundation.org/"><img src="/static/images/wikimedia-button.png" srcset="/static/images/wikimedia-b
```

utton-1.5x.png 1.5x, /static/images/wikimedia-button-2x.png 2x" width="88" height="31" alt
="Wikimedia Foundation"/>
                                                                                <l
i id="footer-poweredbyico">
                                                        <im
g src="/static/images/poweredby_mediawiki_88x31.png" alt="Powered by MediaWiki" srcset="/s
tatic/images/poweredby_mediawiki_132x47.png 1.5x, /static/images/poweredby_mediawiki_176x6
2.png 2x" width="88" height="31"/>

                <script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgPage
ParseReport":{"limitreport":{"cputime":"0.040","walltime":"0.057","ppvisitednodes":{"valu
e":98,"limit":1000000},"ppgeneratednodes":{"value":0,"limit":1500000},"postexpandincludesi
ze":{"value":4419,"limit":2097152},"templateargumentsize":{"value":0,"limit":2097152},"exp
ansiondepth":{"value":3,"limit":40},"expensivefunctioncount":{"value":0,"limit":500},"enti
tyaccesscount":{"value":1,"limit":400},"timingprofile":["100.00%   38.984   1 -total","
73.51%   28.658   1 Template:Coord"," 26.14%   10.189   1 Template:Biobío-geo-stu
b"," 21.36%   8.328   1 Template:Asbox"]},"scribunto":{"limitreport-timeusage":{"valu
e":"0.016","limit":"10.000"},"limitreport-memusage":{"value":832932,"limit":52428800}},"ca
chereport":{"origin":"mw1251","timestamp":"20170208034214","ttl":2592000,"transientconten
t":false}}});});</script><script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set
({"wgBackendResponseTime":51,"wgHostname":"mw1271"});});</script>
        </body>
</html>

Adding the MapReduce function to this project

We start by adding the MapReduce function so that we can use it throughout the project. We explore the data a little bit more and count the total number of lines in all files stored in the wiki folder using the MapReduce function.

In [69]:
```
import math
import functools
from multiprocessing import Pool

def make_chunks(data, num_chunks):
    chunk_size = math.ceil(len(data) / num_chunks)
    return [data[i:i+chunk_size] for i in range(0, len(data), chunk_size)]

def map_reduce(data, num_processes, mapper, reducer):
    chunks = make_chunks(data, num_processes)
    pool = Pool(num_processes)
    chunk_results = pool.map(mapper, chunks)
    pool.close()
    pool.join()
    return functools.reduce(reducer, chunk_results)
```

Counting the total number of lines on all files

We will count the total number of lines on all files in the wiki folder using the MapReduce function.

In [70]:
```
def map_line_count(file_names):
    total = 0
    for fn in file_names:
        with open(os.path.join("wiki", fn)) as f:
            total += len(f.readlines())
    return total

def reduce_line_count(count1, count2):
    return count1 + count2
```

```
target = "data"
map_reduce(file_names, 8, map_line_count, reduce_line_count)
```

Out[70]: 499797

Grep string function

A mapreduce_grep_string() function was defined that takes two arguments as input:

1. A path to the folder. We will use it on the wiki folder but having the argument makes the function easier to reuse.
2. The string that we want to find.

The mapper function receives a chunk of filenames and calculates all occurences of the target string on them. If a file contains no occurences, we chose to not include an entry for that file in the dictionary result.

The reducer function uses the dict.update() method to merge the result dictionaries. MapReduce is used to create a function that, given a string, creates a dictionary where the keys are the file names and the values are lists with all line indexes that contain the given string.

Notice that the target variable will be defined outside and will be the string that we are looking for.

In [71]:
```
# The target variable is defined outside and contains the string
def map_grep(file_names):
    results = {}
    for fn in file_names:
        with open(fn) as f:
            lines = [line for line in f.readlines()]
        for line_index, line in enumerate(lines):
            if target in line:
                if fn not in results:
                    results[fn] = []
                results[fn].append(line_index)
    return results

def reduce_grep(lines1, lines2):
    lines1.update(lines2)
    return lines1

def mapreduce_grep(path, num_processes):
    file_names = [os.path.join(path, fn) for fn in os.listdir(path)]
    return map_reduce(file_names, num_processes,  map_grep, reduce_grep)
```

Finding the occurences of "data"

Using the function created using MapReduce, find all occurrences of the string "data" in the files stored in the wiki folder.

In [72]:
```
target = "data"
data_occurrences = mapreduce_grep("wiki", 8)
```

Allow for case insensitive matches

We can allow for case insensitive matches by converting both the target and the file contents to lowercase before they are matched.

In [73]:
```
def map_grep_insensitive(file_names):
    results = {}
    for fn in file_names:
        with open(fn) as f:
            lines = [line.lower() for line in f.readlines()]
        for line_index, line in enumerate(lines):
            if target.lower() in line:
                if fn not in results:
                    results[fn] = []
                results[fn].append(line_index)
    return results

def mapreduce_grep_insensitive(path, num_processes):
    file_names = [os.path.join(path, fn) for fn in os.listdir(path)]
    return map_reduce(file_names, num_processes,  map_grep_insensitive, reduce_grep)

target = "data"
new_data_occurrences = mapreduce_grep_insensitive("wiki", 8)
```

Checking for additional matches

We already stored the results into variables data_occurences and new_data_occurrences. To check for additional matches with the second version of the algorithm, we can loop over the file names and print the length difference between the results.

Let's verify that the new implementation works by seeing if it finds more matches than the previous implementation.

In [74]:
```
for fn in new_data_occurrences:
    if fn not in data_occurrences:
        print("Found {} new matches on file {}".format(len(new_data_occurrences[fn]), fn))
    elif len(new_data_occurrences[fn]) > len(data_occurrences[fn]):
        print("Found {} new matches on file {}".format(len(new_data_occurrences[fn]) - len
```

```
Found 1 new matches on file wiki/Table_Point_Formation.html
Found 1 new matches on file wiki/Ingrid_GuimarC3A3es.html
Found 2 new matches on file wiki/Jules_Verne_ATV.html
Found 1 new matches on file wiki/Pictogram.html
Found 2 new matches on file wiki/Claire_Danes.html
Found 1 new matches on file wiki/PTPRS.html
Found 1 new matches on file wiki/A_Beautiful_Valley.html
Found 1 new matches on file wiki/Mudramothiram.html
Found 2 new matches on file wiki/Gordon_Bau.html
Found 1 new matches on file wiki/Embraer_Unidade_GaviC3A3o_Peixoto_Airport.html
Found 3 new matches on file wiki/Code_page_1023.html
Found 1 new matches on file wiki/Cryptographic_primitive.html
Found 1 new matches on file wiki/Alex_Kurtzman.html
Found 1 new matches on file wiki/Filip_Pyrochta.html
Found 1 new matches on file wiki/Morgana_King.html
Found 1 new matches on file wiki/Don_Parsons_(ice_hockey).html
Found 1 new matches on file wiki/Bias.html
Found 2 new matches on file wiki/Tomohiko_ItC58D_(director).html
Found 1 new matches on file wiki/Imperial_Venus_(film).html
Found 1 new matches on file wiki/Camp_Nelson_Confederate_Cemetery.html
Found 1 new matches on file wiki/Benny_Lee.html
Found 1 new matches on file wiki/Kul_Gul.html
Found 1 new matches on file wiki/Medicago_murex.html
Found 1 new matches on file wiki/Oldfield_Baby_Great_Lakes.html
Found 1 new matches on file wiki/Wilson_Global_Explorer.html
Found 1 new matches on file wiki/Craig_Chester.html
Found 1 new matches on file wiki/Derek_Acorah.html
Found 1 new matches on file wiki/Jack_Goes_Home.html
Found 1 new matches on file wiki/Morning_Glory_(2010_film).html
Found 1 new matches on file wiki/Tim_Spencer_(singer).html
```

```
Found 1 new matches on file wiki/Lower_Blackburn_Grade_Bridge.html
Found 1 new matches on file wiki/1953E2809354_FA_Cup_qualifying_rounds.html
Found 1 new matches on file wiki/Sol_Eclipse.html
Found 1 new matches on file wiki/Jonathan_A._Goldstein.html
Found 1 new matches on file wiki/83_(number).html
Found 1 new matches on file wiki/Devil_on_Horseback.html
Found 1 new matches on file wiki/Harry_Hill_Bandholtz.html
Found 2 new matches on file wiki/Shpolskii_matrix.html
Found 6 new matches on file wiki/Dragnet_(franchise).html
Found 1 new matches on file wiki/Qalat_Kat.html
Found 3 new matches on file wiki/Maniitsoq_structure.html
Found 1 new matches on file wiki/Ordinary_Virginia.html
Found 1 new matches on file wiki/Dewoitine_D.21.html
Found 1 new matches on file wiki/Furto_di_sera_bel_colpo_si_spera.html
Found 1 new matches on file wiki/Rudy_The_Rudy_Giuliani_Story.html
Found 1 new matches on file wiki/Exploratorium_(film).html
Found 1 new matches on file wiki/Foulonia.html
Found 1 new matches on file wiki/Amborella.html
Found 1 new matches on file wiki/Rally_for_Democracy_and_Progress_(Benin).html
Found 1 new matches on file wiki/Swathi_Chinukulu.html
Found 2 new matches on file wiki/Precorrin6A_reductase.html
Found 1 new matches on file wiki/The_Gentleman_Without_a_Residence_(1915_film).html
Found 1 new matches on file wiki/Manhattan_Murder_Mystery.html
Found 2 new matches on file wiki/Viva_Villa.html
Found 1 new matches on file wiki/Companys_procC3A9s_a_Catalunya.html
Found 1 new matches on file wiki/Avengers_Academy.html
Found 1 new matches on file wiki/Antibiotic_use_in_livestock.html
Found 1 new matches on file wiki/Syngenor.html
Found 1 new matches on file wiki/Cobble_Hill_Brooklyn.html
Found 1 new matches on file wiki/Typhoon_Hester_(1952).html
Found 1 new matches on file wiki/WintersWimberley_House.html
Found 1 new matches on file wiki/Kokan_Colony.html
Found 1 new matches on file wiki/Wilhelm_Wagenfeld_House.html
Found 2 new matches on file wiki/Taipa_HousesE28093Museum.html
Found 1 new matches on file wiki/WLSR.html
Found 1 new matches on file wiki/Lake_County_Examiner.html
Found 1 new matches on file wiki/Copamyntis_infusella.html
Found 1 new matches on file wiki/C11orf30.html
Found 1 new matches on file wiki/Old_Mill_Creek_Illinois.html
Found 1 new matches on file wiki/Bahmanabade_Olya.html
Found 1 new matches on file wiki/Ek_Dil_Sau_Afsane.html
Found 1 new matches on file wiki/Daniel_Cerone.html
Found 1 new matches on file wiki/Shoreyjehye_Do.html
Found 1 new matches on file wiki/Failing_Office_Building.html
Found 1 new matches on file wiki/Pushkar.html
Found 1 new matches on file wiki/List_of_Uzbek_films_of_2014.html
Found 1 new matches on file wiki/KMTZ.html
Found 1 new matches on file wiki/Golabkhvaran.html
Found 1 new matches on file wiki/CurtissWright_Hangar_(Columbia_South_Carolina).html
Found 1 new matches on file wiki/Blue_SWAT.html
Found 1 new matches on file wiki/Danish_Maritime_Safety_Administration.html
Found 1 new matches on file wiki/Don_Raye.html
Found 1 new matches on file wiki/Lis_LC3B8wert.html
Found 1 new matches on file wiki/Doumanaba.html
Found 1 new matches on file wiki/Sahanpur.html
Found 1 new matches on file wiki/Meleh_Kabude_Sofla.html
Found 1 new matches on file wiki/Panchamrutham.html
Found 1 new matches on file wiki/Bibiana_Beglau.html
Found 1 new matches on file wiki/Kattukukke.html
Found 1 new matches on file wiki/Acceptance_(Heroes).html
Found 1 new matches on file wiki/Westchester_Los_Angeles.html
Found 1 new matches on file wiki/Appa_(film).html
Found 1 new matches on file wiki/HD_90156.html
Found 2 new matches on file wiki/The_Audacity_to_Podcast.html
Found 1 new matches on file wiki/Brownfield_(software_development).html
Found 1 new matches on file wiki/Boardman_Township_Mahoning_County_Ohio.html
Found 1 new matches on file wiki/King_Parker_House.html
Found 2 new matches on file wiki/List_of_Spaghetti_Western_films.html
Found 1 new matches on file wiki/The_Future_(film).html
Found 1 new matches on file wiki/Weiser_River.html
```

```
Found 1 new matches on file wiki/Jon_Mullich.html
Found 1 new matches on file wiki/Saravan_Gilan.html
Found 2 new matches on file wiki/Agaritine_gammaglutamyltransferase.html
Found 1 new matches on file wiki/Nuno_Leal_Maia.html
Found 1 new matches on file wiki/Battle_of_Wattignies.html
Found 1 new matches on file wiki/Colchester_Village_Historic_District.html
Found 1 new matches on file wiki/Hayateumi_Hidehito.html
Found 7 new matches on file wiki/List_of_people_from_Bangor_Maine.html
Found 1 new matches on file wiki/Mirisah.html
Found 1 new matches on file wiki/Teiji_Ito.html
Found 1 new matches on file wiki/L._Fry.html
Found 1 new matches on file wiki/Tropical_sprue.html
Found 1 new matches on file wiki/Roxbury_Presbyterian_Church.html
Found 1 new matches on file wiki/Peter_Collingwood.html
Found 4 new matches on file wiki/List_of_molecular_graphics_systems.html
Found 1 new matches on file wiki/Functoid.html
Found 1 new matches on file wiki/Vojin_C486etkoviC487.html
Found 1 new matches on file wiki/Julien_Boisselier.html
Found 1 new matches on file wiki/Jazz_in_Turkey.html
Found 2 new matches on file wiki/Kim_Yonghwa.html
Found 1 new matches on file wiki/Holly_Golightly_(comics).html
Found 1 new matches on file wiki/SalemAuburn_Streets_Historic_District.html
Found 2 new matches on file wiki/Kate_Harwood.html
Found 1 new matches on file wiki/Gulliver_Mickey.html
Found 1 new matches on file wiki/Urs_Burkart.html
Found 1 new matches on file wiki/Smilax_laurifolia.html
Found 1 new matches on file wiki/Taylor_Williamson.html
Found 1 new matches on file wiki/Claudia_Neidig.html
Found 1 new matches on file wiki/Dean_Kukan.html
Found 1 new matches on file wiki/Demographics_of_American_Samoa.html
Found 1 new matches on file wiki/C389cole_des_Mines_de_Douai.html
Found 1 new matches on file wiki/Frost_Township_Michigan.html
Found 1 new matches on file wiki/Shabbir_Kumar.html
Found 1 new matches on file wiki/West_Park_Bridge.html
```

Finding match indexes on lines

Given a string and a target, find all occurrences of the target within the string.

In [75]:
```
def find_match_indexes(line, target):
    results = []
    i = line.find(target, 0)
    while i != -1:
        results.append(i)
        i = line.find(target, i + 1)
    return results

# Test implementation
s = "Data science is related to data mining, machine learning and big data.".lower()
print(find_match_indexes(s, "data"))
```

```
[0, 27, 65]
```

Finding all match locations

We can use any of the above functions to find all match locations. We will use the third function.

After finding all indexes in one line, we need to create pairs by adding the line index.

In [76]:
```
def map_grep_match_indexes(file_names):
    results = {}
    for fn in file_names:
        with open(fn) as f:
            lines = [line.lower() for line in f.readlines()]
```

```
        for line_index, line in enumerate(lines):
            match_indexes = find_match_indexes(line, target.lower())
            if fn not in results:
                results[fn] = []
            results[fn] += [(line_index, match_index) for match_index in match_indexes]
    return results

def mapreduce_grep_match_indexes(path, num_processes):
    file_names = [os.path.join(path, fn) for fn in os.listdir(path)]
    return map_reduce(file_names, num_processes,  map_grep_match_indexes, reduce_grep)

target = "science"
occurrences = mapreduce_grep_match_indexes("wiki", 8)
```

Displaying the results

Our grep algorithms can now find all the matches, however, with the dictionary it produces, it's not very easy to see those matches.

We will write the results into a CSV file to more easily see those matches. We will create a CSV file listing all occurrences, and will also show the text around each occurrence.

In [77]:
```
import csv

# How many character to show before and after the match
context_delta = 30

with open("results.csv", "w") as f:
    writer = csv.writer(f)
    rows = [["File", "Line", "Index", "Context"]]
    for fn in occurrences:
        with open(fn) as f:
            lines = [line.strip() for line in f.readlines()]
        for line, index in occurrences[fn]:
            start = max(index - context_delta, 0)
            end   = index + len(target) + context_delta
            rows.append([fn, line, index, lines[line][start:end]])
    writer.writerows(rows)
```

Here's an example of the table created by our solution. The target was the string "science:"

In [78]:
```
import pandas
df = pandas.read_csv("results.csv")
df.head(10)
```

Out[78]:

	File	Line	Index	Context
0	wiki/Valentin_Yanin.html	6	840	embers of the USSR Academy of Sciences","Full ...
1	wiki/Valentin_Yanin.html	6	890	ers of the Russian Academy of Sciences","Demid...
2	wiki/Valentin_Yanin.html	66	90	href="/wiki/Soviet_Academy_of_Sciences" class=...
3	wiki/Valentin_Yanin.html	66	145	ect" title="Soviet Academy of Sciences">Soviet...
4	wiki/Valentin_Yanin.html	66	173	f Sciences">Soviet Academy of Sciences; he...
5	wiki/Valentin_Yanin.html	144	1440	rs_of_the_USSR_Academy_of_Sciences" title="Cat...
6	wiki/Valentin_Yanin.html	144	1502	rs of the USSR Academy of Sciences">Full Membe...
7	wiki/Valentin_Yanin.html	144	1548	rs of the USSR Academy of Sciences <li...
8	wiki/Valentin_Yanin.html	144	1632	of_the_Russian_Academy_of_Sciences" title="Cat...

	File	Line	Index	Context
9	wiki/Valentin_Yanin.html	144	1697	of the Russian Academy of Sciences">Full Membe...

Conclusion

Locating data from text files is a very common and time-consuming operation when many files are involved. By using MapReduce, we can significantly reduce the time required to locate that data.

In this project, we've implemented a MapReduce grep algorithm that locates all matches of a given string within all files in a given folder.