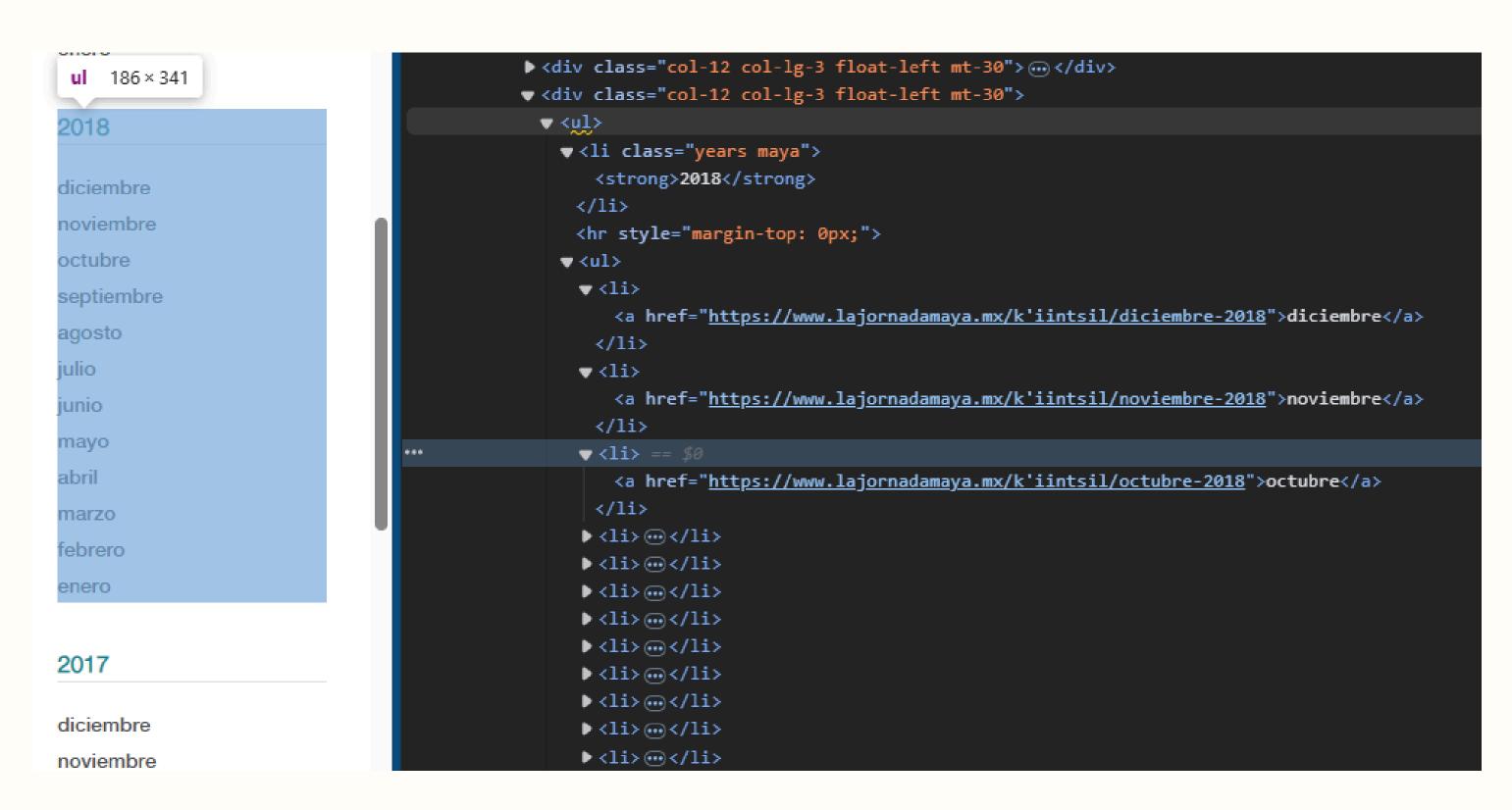
Realizado por:
José Luis Puc Moo
Jean Buenfil



Un con clase "years maya" contiene los años de las noticias, en este caso buscamos el año 2018. En el mismo nivel, se tiene otro el cual cada contiene los enlaces de las noticias por cada mes del año.



Con esta función obtenemos el listado de los enlaces de cada mes del 2018.

```
#Función para obtener la sección con los url de los meses con noticias del 2018

def get_2018_months_tag(url):
    url_jornada_maya = url
    data = requests.get(url_jornada_maya).text
    soup = BeautifulSoup(data, "html5lib")
    #La clase years maya contiene la lista de links a los meses según el año, por lo que se usa el string 2018
    #para especificar el año
    year_2018_li = soup.find("li", class_="years maya", string="2018")
    year_2018_ul = year_2018_li.find_next("ul")
    return year_2018_ul
```

Como ya tenemos un con los enlaces para las noticias por cada mes, es cuestión de obtener las noticias de por cada mes. Por ello, en cada bloque de mes, se enlistan las noticias y cada una de ella tiene su enlace colocado con una clase llamada "post-headline".



Función que obtiene el listado de noticas y su contenido de todos los meses

```
def get_month_blogs(ul_tag):
       year_2018_ul = ul_tag
       month_url = ""
       blogs_list = []
       for link in year_2018_ul.find_all("a", href=True):
           month_url = link["href"]
           data = requests.get(month_url).text
           soup = BeautifulSoup(data, "html5lib")
           blogs = soup.find_all("div", class_="single-blog-area blog-style-2 mb-15 wow fadeInUp")
           for blog in blogs:
               headline_tag = blog.find("a", class_="post-headline")
               headline = headline_tag.text.strip()
               link_ref = headline_tag["href"]
               content = get_post_content(link_ref)
               blogs_list.append((headline,link_ref, content))
       return blogs_list
```

El contenido de cada blog se encuentra en un div con clase "single-blog-content".

```
aiv.single-biog-content 509 × 597.47 min all amount of the service ser
   'IINTSIL
La Jornada Maya
27/12/2018
Ángel Vargas y Carlos Paul
La Jornada Maya
 Ciudad de México
 Jueves 27 de diciembre, 2018
Ti' le k'iino'oba',chi'ichnakile' táan u jach yantal ichil u ya'abil u aj
meyajilo'ob miatsil ku bo'otalo'ob ichil ba'ax k'ajóolta'an beey capítulo
3000 (wa yéetel honorarios) tumen ma' náach k'iin yanik le ken
ts'o'okspajak u contrato'ob yéetel tumen túumben jala'ach oka'an
meyaje', tu beetaj jayp'éel k'eexilo'ob ti' le taak'in yaan ti'al xupbilo'.
Ti'al ya'abach ti' le aj meyajo'obe', u ts'ook k'iinil u meyajo'ob ti' u
mola'ayilo'ob Secretaría de Cultura ti' u Jala'achil u lu'umil Méxicoe' leti'e'
31 ti' diciembre ku taala', ts'o'okole' ma'u yojelo'ob wa yaan u yantal ti'ob
meyaj tu yáax k'iinil enero, beey úuchik u tsilbaltiko'ob ti' La Jornada
ba'ale' tu k'áatajo'ob ma'u ya'alal máax y k'aaba'ob tumen saajko'ob ka
beeta'ak wa ba'ax ti'ob.
Aj meyajo'obe' tu ya'alajo'ob chi'ichnak yaniko'ob tumen mina'an mix u
jo'olpóopilo'ob le mola'ay tu'ux ku meyajo'obo' ti'al ka a'alak ti'ob ba'ax
bîin úuchuk wa ba'ax je'el u páajtal u beetiko'obe'.
```

```
▼ <div class="container">
 ▼ <div class="row"> (flex)
     <!--Bloque de notas por minuto -->
   ▼ <div class="col-12 col-lg-8" id="vertical-2">
     ▼ <div class="single-blog-area blog-style-2 wow fadeInUp" data-wow-delay="0.</p>
       style="visibility: visible; animation-duration: 1000ms; animation-delay: 0
       ▼ <div class="row"> (flex)
         ▶ <div class="col-12"> ••• </div>
         ▼ <div class="col-12">
            <!-- Blog Content -->
           ▼ <div class="single-blog-content"> == $0
              <div class="line top-line-Política"></div>
              <a href="#" class="post-tag Política">K'iintsil</a>
            ▶ <div class="post-meta"> ... </div>
                "Ángel Vargas y Carlos Paul"
                <br>
                " La Jornada Maya"
                <br>
                <br>
                " Ciudad de México"
                <br>
                " Jueves 27 de diciembre, 2018"
                <br>
                <br>
                " Ti' le k'iino'oba',chi'ichnakile' táan u jach yantal ichil u ya
                ku bo'otalo'ob ichil ba'ax k'ajóolta'an beey capítulo 3000 (wa yée
                k'iin yanik le ken ts'o'okspajak u contrato'ob yéetel tumen túumbe
                beetaj jayp'éel k'eexilo'ob ti' le taak'in yaan ti'al xupbilo'."
                <br>
                " Ti'al ya'abach ti' le aj meyajo'obe', u ts'ook k'iinil u meyajo
```

Función que nos permite obtener el contenido de cada noticia. Un problema para obtener unicamente el contenido sin el autor, fecha y demás es que dichos datos junto con el contenido se encuentran dentro de un lo que lo hace complicado verificar para todos pues no siguen un patrón.

```
#Función que utilize get_month_blogs para obtener el contenido de una noticia
def get_post_content(post_href):
    blog_url = post_href
    data = requests.get(blog_url).text
    soup = BeautifulSoup(data, "html5lib")
    blog content div = soup.find all("div", class = "single-blog-content")
    #Existen dos div con la clase single-blog-content en cada noticia, por lo que se usa [-1]
    #porque el útlimo div con esa clase es el que contiene el texto de la noticia
    blog content div = blog content div[-1]
    #Se eliminan secciones no deseadas de la noticia
    for meta in blog_content_div.find_all(["div", "a", "h1", "h6", "figcaption"]):
        meta.extract()
    noticia = blog_content_div.get_text(separator="\n").strip()
    return noticia
```

```
url_jornada_maya = "https://www.lajornadamaya.mx/k'iintsil/archivo"
year_2018_ul = get_2018_months_tag(url_jornada_maya)
blogs_list = get_month_blogs(year_2018_ul)
df = pd.DataFrame(blogs_list)
f.to_csv("noticias.csv", index=False, header=False)
```

K'IINTSIL

La Jornada Maya

28/12/2018 |

Merry MacMasters

Oochel ch'a'aban ti' tomo IV, Francisco Toledo: obra 1957-2017 K'iintsil

Viernes 28 ti' diciembre, 2018

Tu ja'abil 2017e' Fomento Cultural Banamexe' (FCB), yéetel u yáantajil Citibanamex, tu ts'o'oksaj yáax editorial meyajil táan u beeta'al yóok'ol u meyajil Francisco Toledo (Juchitán, 1940), takmuk'ta'ab tumen j boon yéetel j póol meyaj, máax jaxh táakpaj tu súutukil táan u beeta'al.

Tu chowakil jo'op'éel ja'abe', u múuch'kabil FCB tu beetaj jump'éel noj xaak'al ichil u wakp'éel décadas ts'o'ok u meyaj j its'at máak, ts'o'okole'

```
<!-- Blog Content -->

▼ <div class="single-blog-content">
   <div class="line top-line-Política"></div>
   <a href="#" class="post-tag Política">K'iintsil</a>
  ▶ <div class="post-meta"> ••• </div>
  ▼  == $0
     "Merry MacMasters "
     <br>
     " Oochel ch'a'aban ti' tomo IV, Francisco Toledo: obra 1957-2017"
     " K'iintsil"
     <br>
     <br>
     " Viernes 28 ti' diciembre, 2018"
     <br>
     <br>
     " Tu ja'abil 2017e' Fomento Cultural Banamexe' (FCB), yéetel u yáantajil Citibanamex, tu ts'o'oksaj yáax
     editorial meyajil táan u beeta'al yóok'ol u meyajil Francisco Toledo (Juchitán, 1940), takmuk'ta'ab tumen j boon
     yéetel j póol meyaj, máax jaxh táakpaj tu súutukil táan u beeta'al."
     <br>
```

Katia Rejón Ilustración Saúl Cagnone K'iintsil

Jo', Yucatán Viernes 28 ti' diciembre, 2018

"Jump'éel kúuchil e'esjail ts'a'aban u k'aaba', jump'éel u yoochel beeta'an yéetel bronce, jump'éel kúuchil balts'am beeta'al u ti'al, u kis buuts'il turibús ku xîimbaltik kúuchilo'ob tu'ux lîik'ij yéetel u páajtalil u k'aay ti' úuchben kúuchilo'ob kaláanta'ane', ma'atáan u chukik. T beetaj jump'éel u woojil Armando Manzanero je'el bix ts'o'ok u beeta'al yéetel uláak' máako'ob, je'el bix Selena yéetel Michael Jackson", beey úuchik u ya'alik u secretarioil Turismo Federal, Miguel Torruco, tu súutukil úuchik u ts'áak k'ajóoltbil túumben k'aay yaan u beeta'al tumen yukatekoil ti'al u káajbal le túumben ja'abo'.

Woojile' yaan u beeta'al yéetel plástico je'el bix yanik Ken, ts'o'okole'

```
<a href="#" class="post-tag Política">K'iintsil</a>
▶ <div class="post-meta"> ··· </div>
> √p>
   "Katia Rejón" == $0
   " Ilustración Saúl Cagnone"
   <br>
   " K'iintsil"
   <br>
   " Jo', Yucatán"
   " Viernes 28 ti' diciembre, 2018"
   <br>
   " "Jump'éel kúuchil e'esjail ts'a'aban u k'aaba', jump'éel u yoochel beeta'an yéetel bronce, jump'éel kúuchil
   balts'am beeta'al u ti'al, u kis buuts'il turibús ku xíimbaltik kúuchilo'ob tu'ux líik'ij yéetel u páajtalil u
   k'aay ti' úuchben kúuchilo'ob kaláanta'ane', ma'atáan u chukik. T beetaj jump'éel u woojil Armando Manzanero
   je'el bix ts'o'ok u beeta'al yéetel uláak' máako'ob, je'el bix Selena yéetel Michael Jackson", beey úuchik u
   ya'alik u secretarioil Turismo Federal, Miguel Torruco, tu súutukil úuchik u ts'áak k'ajóoltbil túumben k'aay
   yaan u beeta'al tumen yukatekoil ti'al u káajbal le túumben ja'abo'. "
   <br>
```





www.unsitiogenial.es



