Extracting or webscrapping from single XML file

In [1]:
```
import os
os.chdir(r"/Users/myyntiimac/Desktop/single xml file")
```

In [2]:
```
import xml.etree.ElementTree as ET

tree = ET.parse("769952.xml")
root = tree.getroot()
```

The encoding='utf8' argument specifies that the byte string should be encoded using the UTF-8 encoding.

.decode('utf8'): After converting the root element to a byte string, we use the decode() method to convert it back into a Unicode string.

In [3]:
```
root=ET.tostring(root, encoding='utf8').decode('utf8')

root
```

Out[3]:

'<?xml version=\'1.0\' encoding=\'utf8\'?>\n<article>\n <front>\n <journal-meta>\n <journal-id journal-id-type="publication">0901c7 918047d0e2</journal-id>\n <journal-id journal-id-type="publisher" />\n <journal-title>Orphan Drug Approvals</journal-title>\n <abbrev-journal-title abbrev-type="publication" />\n <issn />\n <publisher>\n <publisher-name>\n WebMD, LLC \n </publisher-name>\n </publisher>\n <notes notes-type="support-page">\n index \n </notes>\n </journal-meta>\n <article-meta>\n <arti cle-id>0901c79180555528</article-id>\n <article-categories>\n <subj-group>\n <subject>News Alert</subject>\n </subj-group>\n <series-title />\n </article-categ ories>\n <title-group>\n <article-above-title />\n <article-title>FDA Grants Orphan Drug Status to Gevokizumab</article-title> \n <subtitle />\n <alt-title>The FDA has grant ed orphan drug designation to gevokizumab for the treatment of noninfectious intermediate uveitis, posterior uveitis, or panuveitis, or chronic noninfect ious anterior uveitis.</alt-title>\n </title-group>\n <contrib-group>\n <contrib contrib-type="Journalist">\n <name>\n <surname>Troy Brown</surname>\n </name>\n <role>Journalist</role>\n <b io>\n Troy Brown is a freelance writer for Medscap e. \n </bio>\n <author-comment>\n <title>Disclosure</title>\n Troy Brown has disclos ed no relevant financial relationships. \n </author-co mment>\n <author-comment>\n <title >Title</title>\n \n </author- comment>\n </contrib>\n </contrib-group>\n <pub-date>\n <day>29</day>\n <month>08</month> \n <year>2012</year>\n </pub-date>\n <v olume />\n <issue />\n <fpage />\n <lpage / >\n <copyright-year />\n <copyright-statement />\n <kwd-group>\n <kwd>choroiditis,cyclitis,intermediate uveitis, orphan drugs,pars planitis,posterior uveitis</kwd>\n </kwd-group> \n <history>\n <date date-type="posting">\n <day>29</day>\n <month>08</month>\n <y ear>2012</year>\n </date>\n </history>\n </ article-meta>\n </front>\n <body>\n <sec sec-type="page">\n <title />\n <sec sec-type="Default">\n <title />\n <sec sec-type="section">\n <title />\n August 29, 2012 — The US Food and Drug Administration (FDA) has granted o rphan drug status to gevokizumab (<italic>Xoma 052</italic>, Xoma Corp), a m onoclonal antibody that binds strongly to interleukin 1β (IL-1β), for the tr eatment of noninfectious intermediate uveitis, posterior uveitis, or panuvei tis, or chronic noninfectious anterior uveitis. \n The Orphan Drug Act of 1983 was passed to encourage companies to develop treatments for rare dis eases (diseases that affect fewer than 200,000 people in the United States). Because the market is so small, such treatments can be unprofitable to devel op. Companies that develop orphan drugs receive a 50% tax credit for the cos t of conducting human clinical trials, 7-year marketing exclusivity, and oth er incentives. \n Behçet\'s disease is a rare multisystem disease that causes blood vessel inflammation throughout the body. Common symptoms are mo uth sores, genital sores, and a type of panuveitis known as Behçet\'s uveiti s, an inflammation of the uvea, retina, and vitreous humor that can lead to retinal detachment, vitreous hemorrhage, glaucoma, and blindness. \n "A genetic association has been shown between Behçet\'s disease and the IL-1 ge ne cluster, and IL-1β has been implicated as a mediator in Behçet\'s disease pathogenesis," Christine Kay, MD, the director of Retinal Clinical Research and the director of the Electrophysiology Service in the Vitreoretinal Divis ion of the Department of Ophthalmology at the University of Florida in Gaine sville, told <italic>Medscape Medical News</italic>. Dr. Kay is a clinical c orrespondent for the American Academy of Ophthalmology. \n "Gevokizumab regulates the activation of IL-1 receptors and can be intravenously or subcu taneously administered," Dr. Kay added. \n Patients with Behçet\'s uvei

```
tis have few treatment options. "There are currently only 2 drugs FDA-approv
ed for the treatment of chronic noninfectious intermediate, posterior, and p
anuveitis (<italic>Retisert</italic> [Bausch &amp;amp; Lomb] and <italic>Ozurdex
</italic> [Allergan]), and both are extended-release corticosteroid ocular i
mplants," Dr. Kay said.</p>\n<p>Results of a proof-of-concept phase 2 trial
of intravenous gevokizumab in 7 patients with Behçet\'s uveitis were publish
ed in the April issue of the <italic>Annals of Rheumatic Diseases</italic>.
In that trial patients were given a single infusion of gevokizumab (0.3 mg/k
g), and all patients experienced complete reduction of intraocular inflammat
ion in between 4 and 21 days (median, 14 days). There were no treatment-rela
ted adverse events.</p>\n<p>"In clinical trials, so far, gevokizumab has bee
n studied in nearly 500 patients. The studies have shown that gevokizumab is
well-tolerated, and no drug-related adverse events have been reported," Fred
Kurland, chief financial officer of Xoma, said in an email interview with <i
talic>Medscape Medical News</italic>.</p>\n<p>Although it appears that gevok
izumab "may offer a viable treatment option in Behçet\'s disease, it remains
to be seen if an IL-1 antibody will have an effect in other forms of noninfe
ctious uveitis. A phase 3 clinical trial to evaluate the efficacy of [gevoki
zumab] in the treatment of noninfectious uveitis is in the recruitment proce
ss," Dr. Kay said.</p>\n<p>"Gevokizumab does offer the possibility of a path
ophysiology-driven targeted therapy for IL-1 related uveitis, and if proven
safe and effective in a phase 3 trial, this could provide a valuable option
in the treatment of noninfectious intermediate uveitis, posterior uveitis, a
nd panuveitis. Even if this drug is only shown to be effective in Behçet\'s
disease, this could provide a useful and targeted treatment for an extremely
aggressive condition, perhaps limiting broader and more toxic immunosuppress
ion," Dr. Kay said.</p>\n<p>\n                                <bold>Other Potential
Indications</bold>\n                    </p>\n<p>"As an IL-1β inhibitor, gev
okizumab has potential in a very large number of indications that are driven
by inflammation, such as noninfectious uveitis.... [W]e are also engaged in
2 proof-of-concept phase 2 trials using gevokizumab in patients with moderat
e to severe acne vulgaris and in erosive osteoarthritis of the hand, and we
will initiate a third proof-of-concept trial in another indication later thi
s year," Kurland explained.</p>\n<p>"With respect to the [noninfectious uvei
tis] market specifically, we estimate that there are approximately 150,000 p
atients in the [United States who have noninfectious uveitis]," Kurland adde
d, noting they are not discussing the drug\'s pricing yet.</p>\n<p>\n
<italic>Dr. Kay has disclosed no relevant financial relationships.</italic>
\n                    </p>\n                </sec>\n            </sec>\n
</sec>\n    </body>\n    <back>\n        <ref-list>\n            <title>Refe
rences</title>\n            <list>\n                <list-item>\n
<p />\n                </list-item>\n            </list>\n        </ref-list
>\n        <ack>\n            <title>Acknowledgements</title>\n            <
p />\n        </ack>\n        <fn-group>\n            <fn fn-type="bkmtr_fro
nt">\n                <p />\n            </fn>\n        </fn-group>\n    </
back>\n</article>'
```

In [5]:
```
import re, string, unicodedata
import nltk
```

In [6]:
```
from bs4 import BeautifulSoup
from nltk import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import LancasterStemmer, WordNetLemmatizer
```

In [7]:
```
#Define all function for Root as input
def strip_html(text):
    soup = BeautifulSoup(text, "html.parser")
    return soup.get_text()

def remove_between_square_brackets(text):
    return re.sub('\[[^]]*\]', '', text)
```

```
def denoise_text(text):
    text = strip_html(text)
    text = remove_between_square_brackets(text)
    text=re.sub('  ','',text)
    return text
```

In [8]:
```
sample = denoise_text(root)
```

```
/Users/myyntiimac/anaconda3/lib/python3.10/site-packages/bs4/builder/__init_
_.py:545: XMLParsedAsHTMLWarning: It looks like you're parsing an XML docume
nt using an HTML parser. If this really is an HTML document (maybe it's XHTM
L?), you can ignore or filter this warning. If it's XML, you should know tha
t using an XML parser will be more reliable. To parse this document as XML,
make sure you have the lxml package installed, and pass the keyword argument
`features="xml"` into the BeautifulSoup constructor.
  warnings.warn(
```

In [9]:
```
sample
```

Out[9]: '\n\n\n\n0901c7918047d0e2\n\nOrphan Drug Approvals\n\n\n\n\nWebMD, LLC\n\n\n\nindex\n\n\n\n0901c79180555528\n\n\nNews Alert\n\n\n\n\n\nFDA Grants Orphan Drug Status to Gevokizumab\n\nThe FDA has granted orphan drug designation to gevokizumab for the treatment of noninfectious intermediate uveitis, posterior uveitis, or panuveitis, or chronic noninfectious anterior uveitis.\n\n\n\n\nTroy Brown\n\nJournalist\n\nTroy Brown is a freelance writer for Medscape.\n\n\nDisclosure\nTroy Brown has disclosed no relevant financial relationships.\n\n\nTitle\n\n\n\n\n\n29\n08\n2012\n\n\n\n\n\n\n\nchoroiditis,cyclitis,intermediate uveitis,orphan drugs,pars planitis,posterior uveitis\n\n\n\n29\n08\n2012\n\n\n\n\n\n\n\n\n\n\nAugust 29, 2012 — The US Food and Drug Administration (FDA) has granted orphan drug status to gevokizumab (Xoma 052, Xoma Corp), a monoclonal antibody that binds strongly to interleukin 1β (IL-1β), for the treatment of noninfectious intermediate uveitis, posterior uveitis, or panuveitis, or chronic noninfectious anterior uveitis.\nThe Orphan Drug Act of 1983 was passed to encourage companies to develop treatments for rare diseases (diseases that affect fewer than 200,000 people in the United States). Because the market is so small, such treatments can be unprofitable to develop. Companies that develop orphan drugs receive a 50% tax credit for the cost of conducting human clinical trials, 7-year marketing exclusivity, and other incentives.\nBehçet\'s disease is a rare multisystem disease that causes blood vessel inflammation throughout the body. Common symptoms are mouth sores, genital sores, and a type of panuveitis known as Behçet\'s uveitis, an inflammation of the uvea, retina, and vitreous humor that can lead to retinal detachment, vitreous hemorrhage, glaucoma, and blindness.\n"A genetic association has been shown between Behçet\'s disease and the IL-1 gene cluster, and IL-1β has been implicated as a mediator in Behçet\'s disease pathogenesis," Christine Kay, MD, the director of Retinal Clinical Research and the director of the Electrophysiology Service in the Vitreoretinal Division of the Department of Ophthalmology at the University of Florida in Gainesville, told Medscape Medical News. Dr. Kay is a clinical correspondent for the American Academy of Ophthalmology.\n"Gevokizumab regulates the activation of IL-1 receptors and can be intravenously or subcutaneously administered," Dr. Kay added.\nPatients with Behçet\'s uveitis have few treatment options. "There are currently only 2 drugs FDA-approved for the treatment of chronic noninfectious intermediate, posterior, and panuveitis (Retisertand Ozurdex), and both are extended-release corticosteroid ocular implants," Dr. Kay said.\nResults of a proof-of-concept phase 2 trial of intravenous gevokizumab in 7 patients with Behçet\'s uveitis were published in the April issue of the Annals of Rheumatic Diseases. In that trial patients were given a single infusion of gevokizumab (0.3 mg/kg), and all patients experienced complete reduction of intraocular inflammation in between 4 and 21 days (median, 14 days). There were no treatment-related adverse events.\n"In clinical trials, so far, gevokizumab has been studied in nearly 500 patients. The studies have shown that gevokizumab is well-tolerated, and no drug-related adverse events have been reported," Fred Kurland, chief financial officer of Xoma, said in an email interview with Medscape Medical News.\nAlthough it appears that gevokizumab "may offer a viable treatment option in Behçet\'s disease, it remains to be seen if an IL-1 antibody will have an effect in other forms of noninfectious uveitis. A phase 3 clinical trial to evaluate the efficacy ofin the treatment of noninfectious uveitis is in the recruitment process," Dr. Kay said.\n"Gevokizumab does offer the possibility of a pathophysiology-driven targeted therapy for IL-1 related uveitis, and if proven safe and effective in a phase 3 trial, this could provide a valuable option in the treatment of noninfectious intermediate uveitis, posterior uveitis, and panuveitis. Even if this drug is only shown to be effective in Behçet\'s disease, this could provide a useful and targeted treatment for an extremely aggressive condition, perhaps limiting broader and more toxic immunosuppression," Dr. Kay said.\n\nOther Potential Indications\n\n"As an IL-1β inhibitor, gevokizumab has potential in a very large number of indications that are driven by inflammation, such as noninfectious uveitis.... e are also engaged in 2 proof-of-concept phase 2 trials using gevokizumab in patients with moderate to severe acne vulgaris and in erosive osteoarthritis of the hand, and we will initiate a third proof-of-concept trial in another indication later this year," Kurland explained.\n"With respect to themarket specifically, we estimate that there are

```
approximately 150,000 patients in the ," Kurland added, noting they are not
discussing the drug\'s pricing yet.\n\nDr. Kay has disclosed no relevant fin
ancial relationships.\n\n\n\n\n\n\n\nReferences\n\n\n\n\n\n\n\nAcknowledgeme
nts\n\n\n\n\n\n\n\n\n'
```

In [ ]: