Web Scrapping

In [1]:
```
# Web scrapping for xml sheet

import os
os.chdir(r"C:\Users\samru\Music\New folder\29th- webscrapping\xml_single art

import xml.etree.ElementTree as ET

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

root=ET.tostring(root, encoding='utf8').decode('utf8')

root

import re, string, unicodedata
import nltk

from bs4 import BeautifulSoup
#from nltk import word_tokenize, sent_tokenize
#from nltk.corpus import stopwords
#from nltk.stem import LancasterStemmer, WordNetLemmatizer

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

sample = denoise_text(root)
```

```
C:\Users\samru\anaconda3\Lib\site-packages\bs4\builder\__init__.py:545: XMLP
arsedAsHTMLWarning: It looks like you're parsing an XML document using an HT
ML parser. If this really is an HTML document (maybe it's XHTML?), you can i
gnore or filter this warning. If it's XML, you should know that using an XML
parser will be more reliable. To parse this document as XML, make sure you h
ave the lxml package installed, and pass the keyword argument `features="xm
l"` into the BeautifulSoup constructor.
  warnings.warn(
```

In [2]:
```
print(sample)
```

0901c7918047d0e2

Orphan Drug Approvals

WebMD, LLC

index

0901c79180555528

News Alert

FDA Grants Orphan Drug Status to Gevokizumab

The FDA has granted orphan drug designation to gevokizumab for the treatment of noninfectious intermediate uveitis, posterior uveitis, or panuveitis, or chronic noninfectious anterior uveitis.

Troy Brown

Journalist

Troy Brown is a freelance writer for Medscape.

Disclosure
Troy Brown has disclosed no relevant financial relationships.

Title

29
08
2012

choroiditis,cyclitis,intermediate uveitis,orphan drugs,pars planitis,posteri
or uveitis

29
08
2012

August 29, 2012 — The US Food and Drug Administration (FDA) has granted orph
an drug status to gevokizumab (Xoma 052, Xoma Corp), a monoclonal antibody t
hat binds strongly to interleukin 1β (IL-1β), for the treatment of noninfect
ious intermediate uveitis, posterior uveitis, or panuveitis, or chronic noni
nfectious anterior uveitis.

The Orphan Drug Act of 1983 was passed to encourage companies to develop tre
atments 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 u
nprofitable to develop. Companies that develop orphan drugs receive a 50% ta
x credit for the cost of conducting human clinical trials, 7-year marketing
exclusivity, and other incentives.

Behçet's disease is a rare multisystem disease that causes blood vessel infl
ammation throughout the body. Common symptoms are mouth sores, genital sore
s, and a type of panuveitis known as Behçet's uveitis, an inflammation of th
e uvea, retina, and vitreous humor that can lead to retinal detachment, vitr
eous hemorrhage, glaucoma, and blindness.

"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 diseas
e pathogenesis," Christine Kay, MD, the director of Retinal Clinical Researc
h and the director of the Electrophysiology Service in the Vitreoretinal Div
ision of the Department of Ophthalmology at the University of Florida in Gai
nesville, told Medscape Medical News. Dr. Kay is a clinical correspondent fo
r the American Academy of Ophthalmology.

"Gevokizumab regulates the activation of IL-1 receptors and can be intraveno
usly or subcutaneously administered," Dr. Kay added.

Patients with Behçet's uveitis have few treatment options. "There are curren
tly only 2 drugs FDA-approved for the treatment of chronic noninfectious int
ermediate, posterior, and panuveitis (Retisertand Ozurdex), and both are ex
tended-release corticosteroid ocular implants," Dr. Kay said.

Results 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.

"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.

Although 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.

"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.

Other Potential Indications

"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.

"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.

Dr. Kay has disclosed no relevant financial relationships.

References

Acknowledgements

In [3]:
```
# Sent_tokenize and word_tokenize
from nltk.tokenize import word_tokenize
tokens = word_tokenize(sample)

from nltk.tokenize import sent_tokenize
sentences = sent_tokenize(sample)

# Create BOW
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X = cv.fit_transform(sentences).toarray()
```

In [4]:
```
tokens
```

Out[4]:
```
['0901c7918047d0e2',
 'Orphan',
 'Drug',
 'Approvals',
 'WebMD',
 ',',
 'LLC',
 'index',
 '0901c79180555528',
 'News',
 'Alert',
 'FDA',
 'Grants',
 'Orphan',
 'Drug',
 'Status',
 'to',
 'Gevokizumab',
 'The',
 'FDA',
 'has',
 'granted',
 'orphan',
 'drug',
 'designation',
 'to',
 'gevokizumab',
 'for',
 'the',
 'treatment',
 'of',
 'noninfectious',
 'intermediate',
 'uveitis',
 ',',
 'posterior',
 'uveitis',
 ',',
 'or',
 'panuveitis',
 ',',
 'or',
 'chronic',
 'noninfectious',
 'anterior',
 'uveitis',
 '.',
 'Troy',
 'Brown',
 'Journalist',
 'Troy',
 'Brown',
 'is',
 'a',
 'freelance',
```

```
'for',
'Medscape',
'.',
'Disclosure',
'Troy',
'Brown',
'has',
'disclosed',
'no',
'relevant',
'financial',
'relationships',
'.',
'Title',
'29',
'08',
'2012',
'choroiditis',
',',
'cyclitis',
',',
'intermediate',
'uveitis',
',',
'orphan',
'drugs',
',',
'pars',
'planitis',
',',
'posterior',
'uveitis',
'29',
'08',
'2012',
'August',
'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',
'.',
'The',
'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',
'.',
'Behçet',
"'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',
'.',
'``',
'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',
'.',
'``',
'Gevokizumab',
'regulates',
'the',
'activation',
'of',
'IL-1',
'receptors',
'and',
'can',
'be',
'intravenously',
'or',
'subcutaneously',
'administered',
',',
"''"
```

```
'Dr.',
'Kay',
'added',
'.',
'Patients',
'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',
'.',
'Results',
'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',
'.',
'``',
'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',
'.',
'Although',
'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',
'.',
'``',
'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',
'.',
'Other'
```

```
'Potential',
'Indications',
"''",
'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',
'.',
'``',
'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',
'.',
'Dr.',
'Kay',
'has'
```

```
 'disclosed',
 'no',
 'relevant',
 'financial',
 'relationships',
 '.',
 'References',
 'Acknowledgements']
```

In [6]:
```
sentences
```

Out[6]: ['\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.',
 'Troy Brown\n\nJournalist\n\nTroy Brown is a freelance writer for Medscape.',
 'Disclosure\nTroy Brown has disclosed no relevant financial relationships.',
 'Title\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\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.',
 'The 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.',
 "Behçet'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.",
 '"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.',
 '"Gevokizumab regulates the activation of IL-1 receptors and can be intravenously or subcutaneously administered," Dr. Kay added.',
 "Patients 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.',
 "Results 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.',
 '"In clinical trials, so far, gevokizumab has been studied in nearly 500 patients.',
 'dies have shown that gevokizumab is well-tolerated, and no drug-re

```
lated adverse events have been reported," Fred Kurland, chief financial off
icer of Xoma, said in an email interview with Medscape Medical News.',
 'Although 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 a
n effect in other forms of noninfectious uveitis.',
 'A phase 3 clinical trial to evaluate the efficacy ofin the treatment of n
oninfectious uveitis is in the recruitment process," Dr. Kay said.',
 '"Gevokizumab does offer the possibility of a pathophysiology-driven targe
ted 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 non
infectious intermediate uveitis, posterior uveitis, and panuveitis.',
 'Even if this drug is only shown to be effective in Behçet\'s disease, thi
s could provide a useful and targeted treatment for an extremely aggressive
condition, perhaps limiting broader and more toxic immunosuppression," Dr.
Kay said.',
 'Other Potential Indications\n\n"As an IL-1β inhibitor, gevokizumab has po
tential in a very large number of indications that are driven by inflammati
on, such as noninfectious uveitis.... e are also engaged in 2 proof-of-conc
ept phase 2 trials using gevokizumab in patients with moderate to severe ac
ne vulgaris and in erosive osteoarthritis of the hand, and we will initiate
a third proof-of-concept trial in another indication later this year," Kurl
and explained.',
 '"With respect to themarket specifically, we estimate that there are appro
ximately 150,000 patients in the ," Kurland added, noting they are not disc
ussing the drug\'s pricing yet.',
 'Dr. Kay has disclosed no relevant financial relationships.',
 'References\n\n\n\n\n\n\n\nAcknowledgements']
```

In [8]:
```
# BOW CounterVectorizer convert sentences to vector

X
```

Out[8]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [1, 0, 0, ..., 0, 0, 1],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

Streamlit to deploy

In [9]:
```
# Streamli app
import streamlit as st

custom_css = """
<style>
body {
    background-color: #f0f2f6; /* Change this color to your desired backgrou
}
</style>
"""

# Streamlit app title
st.title('XML to Para')
# custom CSS
```

```
st.markdown(custom_css, unsafe_allow_html=True)

def clean_text(text):
    # Define your cleaning logic here
    # For demonstration, let's just return the input text as is
    return text



uploaded_file = st.file_uploader("Upload XML file", type=["xml"])

if uploaded_file is not None:
    try:
        # Load XML from the uploaded file
        tree = ET.parse(uploaded_file)
        root = tree.getroot()
        # Extract text from XML
        xml_text = ET.tostring(root, encoding='unicode', method='text')
        # Clean the extracted text
        cleaned_text = clean_text(xml_text)
        # Display cleaned text
        st.text_area("Cleaned Text", value=cleaned_text, height=400)
    except Exception as e:
        st.error(f"Error: {e}")
```

```
2024-06-06 09:43:02.655
  Warning: to view this Streamlit app on a browser, run it with the followin
g
  command:

    streamlit run C:\Users\samru\AppData\Roaming\Python\Python311\site-packa
ges\ipykernel_launcher.py [ARGUMENTS]
```

In [ ]: