DSBA-6100-U91:ITCS-6100-U91

Extract Problem Solved By Patent

Heath Styles hstyles@uncc.edu, **Kushal Venkateshgupta** kguduruv@uncc.edu, **Raj Shah** rshah62@uncc.edu, **Gordon Willingham** gwillin3@uncc.edu, **and Anoosh Guddehithlu Prathap Kumar** aguddehi@uncc.edu

Project Report October:26 th 2021

1 Abstract

For this project the approach is to perform data analysis on the provided patent xml data files accessed from the USPTO dataset. This is done to assess the xml data structure in terms of identifying the location of the target “description” tag. The description tag has been identified as the target tag, because it contains data pertaining to the problems solved by the various patents. The xml file will then be parsed using xml.etree.ElementTree (an xml parser) that parses, explores, modify and populate XML files, which would allow for the separation of each individual patent based on the location of starting and ending xml file tags. Once each patent has been separated, we then target the description tag of each patent to extract the needed information and write each description to a txt file. The purpose of this txt file is to hold all the information that will be used in the topic modelling process to identify keywords that indicate a problem statement within the patent.

2 Introduction

This project is mainly focusing on extracting problems solved by the patents using USPTO dataset. It is achieved using Beautiful soup, xml.etree.ElementTree parser, topic modelling, regex etc methods.

There are three primary steps involved in this process which are:
1. Extraction from the patent dataset: In this step we extract data from the patent xml document using beautiful soup.
2. Cleaning the data extracted: Here we try to remove as many stop words as possible which makes comparison easier and remove all the unnecessary xml tags.
3. Analyze each of them using Entity/Topic extraction: After Cleaning and Tokenizing the data, we check for the correlation and the means.

3 Problem Statement

Our aim is to extract the actual problem solved by a patent from the patent XML document.

4 Data Description

Input dataset Size :
Input dataset consists of **7363** Patents with various tags and Information in XML format.

Input Data:
The training data provided are XML files with many patent documents. Which consist of problem information, solutions to it and how the problem is being solved.

The XML document has the details of patent information, detailed description of a problem, experiments performed, published information, category, background, brief descriptions with images, solutions to the actual problem.
Detailed description paragraphs with little cleanup will give us all the required information which would help us to find the actual problem solved by the patent.

Output Data:
The expected output in this project is to find the actual problem solved by the patent in the patent document i.e extracting description from patent file to find the keyword like "However", "Although" etc.

4.1 Subsection motivating your approach

Our approach is divided into phases:

1) Analyse pattern of problem description:

   By conducting analysis of the problem description layout with the xml file, we are able to identify tag structures necessary for proper extraction of the problem description tag element from the xml

2) Description extraction:

   Once the description tag patterns have been identified in step 1, an xml parser (xml.etree.ElementTree) can be used to separate the specific description sections from the xml file. By doing so this allows for further analysis which is more focused on identifying the problem statements contained within each description section.

3) Xml tag removal & stop word cleanup:

Once the descriptions have been collectively written to a txt file, the next approach would be cleaning the data to allow for quicker, more efficient data processing

4) Running NLTK concordance:
   With all description tag content extract from xml and written to txt file, nltk concordance tool can be used on the descriptions.txt file identify and sentences containing words which identify the start of a problem statement

5) Identify Problems Solved:

   Once sentences can be identified using the nltk concordance tool, they must then be extracted into a separate file.

4.2 Subsection describing the details of your methods

Analyze patent document:
Skim through the patent xml files manually and find some patterns in problem statements.
Extraction of description tags:
Extract all the descriptions from the huge xml file with 8000 patent xmls and copy them to a text file
XML tag removal:
Remove all the internal xml tags present in the extracted description document as part of cleanup
Stop words cleanup:
Remove all the unnecessary stop words from the extracted description like then, is, are, them, you etc

Nltk Concordance:
Tokenize descriptions data in txt file and and locate occurrences of words such as “however” and “advantages”, which indicate the start of a problem solution statement

Problem extraction:
Extract all the sentences with the important keywords like however, although from the extracted description text. Using NLTK and similarity functions we need to analyze the actual problem solved from the cleaned up text.

4.3 Subsection details of your experiments

To analyze the xml documents everyone in the team has skimmed through 125 patent documents manually and found some interesting patterns in which the problem statements are structured.
Processed a huge XML file with 8000 patent documents and extracted all the description tags using beautiful soup library
Once extracted descriptions were written to a text file.

Tested Topic Modelling on extracted descriptions.txt file

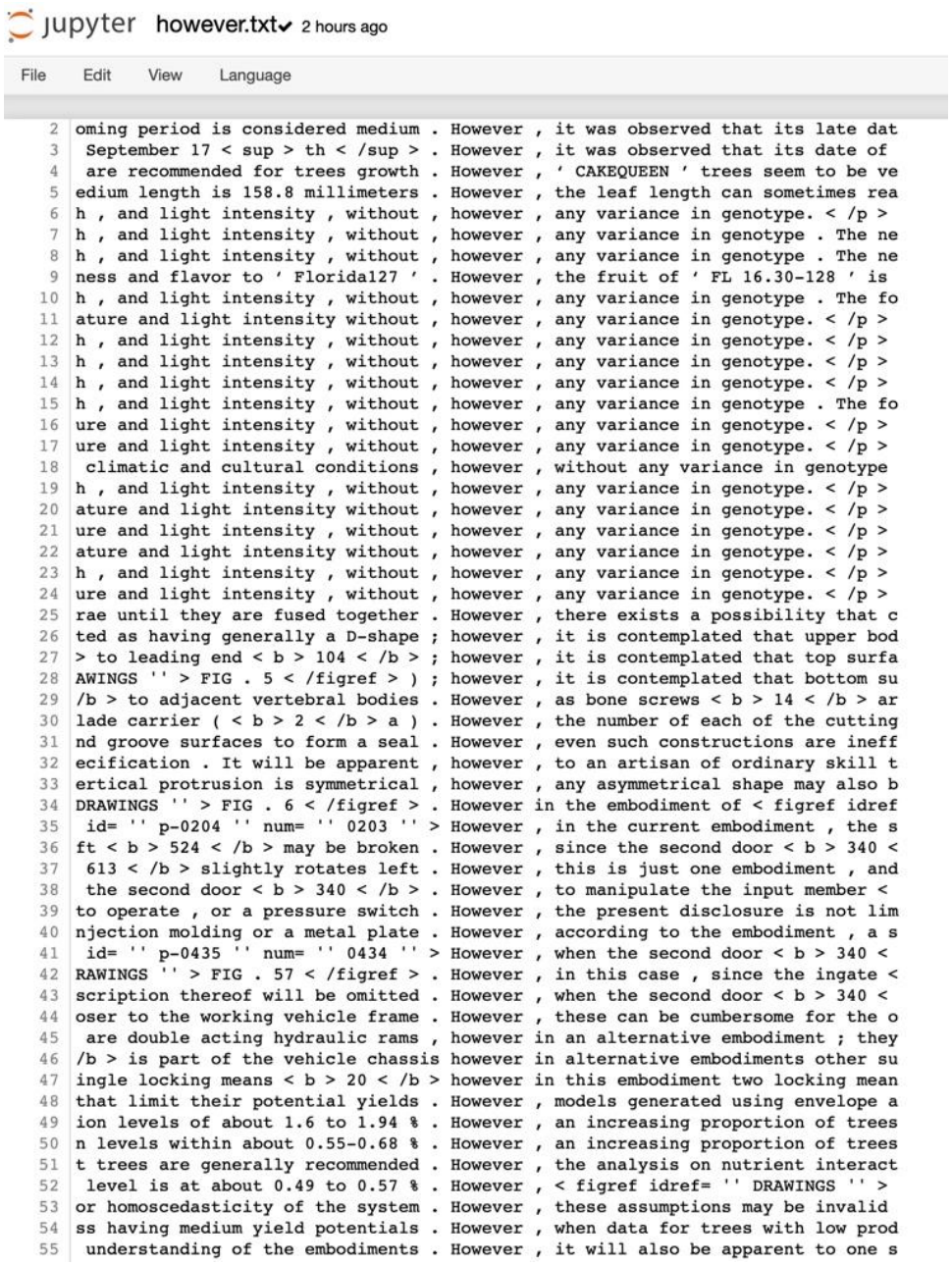
```
0       5       device network system data memory storage devices communication computer processor
1       5       unit signal control circuit power sensor time output operation state
2       5       cell group sequence acid cells target sample amino weight comprising
3       5       layer light num material structure region element formed substrate plurality
4       5       data user information image based system content display object location
5       5       figref drawings idref fig num embodiment shown view block figs
6       5       portion side direction position surface end member body assembly frame
7       5       embodiments num present invention method disclosure include components component elements
8       5       num flow temperature pressure system high fluid gas valve liquid
9       5       num level end application entry step description heading pt reference
```

```
<150> LL/token: -8.10832
<160> LL/token: -8.10753
<170> LL/token: -8.10671
<180> LL/token: -8.10548
<190> LL/token: -8.10498

0       5       device network system data memory storage devices communication computer processor
1       5       unit signal control circuit power sensor time output operation state
2       5       cell group sequence acid cells target sample id amino weight
3       5       layer light material num structure region element formed substrate surface
4       5       data user information image based system content display object location
5       5       figref drawings idref fig num embodiment shown view block figs
6       5       portion side direction position surface end member body assembly frame
7       5       embodiments num present invention method disclosure include component components apparatus
8       5       flow num temperature system pressure vehicle high fluid gas valve
9       5       num level end application entry step description heading reference pt

<200> LL/token: -8.10492

Total time: 11 minutes 21 seconds
```

Why We Choose Topic Modeling :
Topic modelling was run in an attempt to identify and further select words which indicate or could be used to identify the problem statement that is being solved by any particular patent.

After further discussion, it was decided that a better approach would be to use nltk concordance to identify the occurrences of words which identify the beginning of a problem statement. Example of concordance usage can be seen below:

```
import nltk.corpus
import sys
from nltk.text import Text

from nltk import word_tokenize

nltk.download('punkt')

contents = open('descriptions.txt','r').read()

#nltk.download('gutenberg')
#textList = Text(nltk.corpus.gutenberg.words('descriptions.txt'))

tokens = word_tokenize(contents)
textList = Text(tokens)
#collect = textList.concordance('however', lines=24197)


saveout = sys.stdout
file2 = open('however.txt', 'w')
sys.stdout = file2
collect = textList.concordance('however', lines=24197)

sys.stdout = saveout
file2.close()
```


jupyter however.txt 2 hours ago

File Edit View Language

```
oming period is considered medium . However , it was observed that its late dat
 September 17 < sup > th < /sup > . However , it was observed that its date of
 are recommended for trees growth . However , ' CAKEQUEEN ' trees seem to be ve
edium length is 158.8 millimeters . However , the leaf length can sometimes rea
h , and light intensity , without , however , any variance in genotype. < /p >
h , and light intensity , without , however , any variance in genotype . The ne
h , and light intensity , without , however , any variance in genotype . The ne
ness and flavor to ' Florida127 ' . However , the fruit of ' FL 16.30-128 ' is
h , and light intensity , without , however , any variance in genotype . The fo
ature and light intensity without , however , any variance in genotype. < /p >
h , and light intensity , without , however , any variance in genotype. < /p >
h , and light intensity , without , however , any variance in genotype. < /p >
h , and light intensity , without , however , any variance in genotype. < /p >
h , and light intensity , without , however , any variance in genotype . The fo
ure and light intensity , without , however , any variance in genotype. < /p >
ure and light intensity , without , however , any variance in genotype. < /p >
 climatic and cultural conditions , however , without any variance in genotype
h , and light intensity , without , however , any variance in genotype. < /p >
ature and light intensity without , however , any variance in genotype. < /p >
ure and light intensity , without , however , any variance in genotype. < /p >
ature and light intensity without , however , any variance in genotype. < /p >
h , and light intensity , without , however , any variance in genotype. < /p >
ure and light intensity , without , however , any variance in genotype. < /p >
rae until they are fused together . However , there exists a possibility that c
ted as having generally a D-shape ; however , it is contemplated that upper bod
> to leading end < b > 104 < /b > ; however , it is contemplated that top surfa
AWINGS '' > FIG . 5 < /figref > ) ; however , it is contemplated that bottom su
/b > to adjacent vertebral bodies . However , as bone screws < b > 14 < /b > ar
lade carrier ( < b > 2 < /b > a ) . However , the number of each of the cutting
nd groove surfaces to form a seal . However , even such constructions are ineff
ecification . It will be apparent , however , to an artisan of ordinary skill t
ertical protrusion is symmetrical , however , any asymmetrical shape may also b
DRAWINGS '' > FIG . 6 < /figref > . However in the embodiment of < figref idref
 id= '' p-0204 '' num= '' 0203 '' > However , in the current embodiment , the s
ft < b > 524 < /b > may be broken . However , since the second door < b > 340 <
 613 < /b > slightly rotates left . However , this is just one embodiment , and
 the second door < b > 340 < /b > . However , to manipulate the input member <
to operate , or a pressure switch . However , the present disclosure is not lim
njection molding or a metal plate . However , according to the embodiment , a s
 id= '' p-0435 '' num= '' 0434 '' > However , when the second door < b > 340 <
RAWINGS '' > FIG . 57 < /figref > . However , in this case , since the ingate <
scription thereof will be omitted . However , when the second door < b > 340 <
oser to the working vehicle frame . However , these can be cumbersome for the o
 are double acting hydraulic rams , however in an alternative embodiment ; they
/b > is part of the vehicle chassis however in alternative embodiments other su
ingle locking means < b > 20 < /b > however in this embodiment two locking mean
that limit their potential yields . However , models generated using envelope a
ion levels of about 1.6 to 1.94 % . However , an increasing proportion of trees
n levels within about 0.55-0.68 % . However , an increasing proportion of trees
t trees are generally recommended . However , the analysis on nutrient interact
 level is at about 0.49 to 0.57 % . However , < figref idref= '' DRAWINGS '' >
or homoscedasticity of the system . However , these assumptions may be invalid
ss having medium yield potentials . However , when data for trees with low prod
 understanding of the embodiments . However , it will also be apparent to one s
```

5 Discussion and Related Work

The first phase of this project was identifying the target tag/area of the provided xml file. The target tag was identified as seen in the photo below:

```
<description id="description">
<?BRFSUM description="Brief Summary" end="lead"?>
<heading id="h-0001" level="1">FIELD OF THE INVENTION</heading>
<p id="p-0002" num="0001">The present invention relates to a pneumatic control system and, more specifically, to such a system used in agricultural implements.</p>
<heading id="h-0002" level="1">BACKGROUND OF THE INVENTION</heading>
<p id="p-0003" num="0002">Modern seeding agricultural implements provide the function of distributing bulk seeds from an onboard hopper past a metering system to a series
<p id="p-0004" num="0003">In such an implement, it is important to have uniform flow throughout the pneumatic system. In many cases the conduits are split into branches and
<p id="p-0005" num="0004">Accordingly, what is needed in the art is a system for providing uniform flow through the pneumatic passages in an agricultural implement and othe
<heading id="h-0003" level="1">SUMMARY OF THE INVENTION</heading>
<p id="p-0006" num="0005">The invention provides control valves within a system having a plurality of pneumatic flow passages to provide uniform and preselected flow.</p>
<p id="p-0007" num="0006">In one form, the disclosure is a pneumatic flow system for an agricultural implement with a source of pressurized air and a system of hoses connec
<p id="p-0008" num="0007">In another form, the disclosure is a pneumatic flow control system with a source of pressurized air and a tube receiving pressurized air from the
<p id="p-0009" num="0008">In still another form, the disclosure an agricultural implement for planting seeds and movable in a given direction for planting seeds in a field.
<p id="p-0010" num="0009">One advantage of the disclosure is to provide a simplified yet effective control of air through the plurality of hoses in a pneumatic conveying sy
<p id="p-0011" num="0010">Another advantage is to ensure uniform distribution of seeds and improved accuracy in an air seeder planting unit.</p>
<?BRFSUM description="Brief Summary" end="tail"?>
<?brief-description-of-drawings description="Brief Description of Drawings" end="lead"?>
<description-of-drawings>
<heading id="h-0004" level="1">BRIEF DESCRIPTION OF THE DRAWINGS</heading>
<p id="p-0012" num="0011">For the purpose of illustration, there are shown in the drawings certain embodiments of the present invention. It should be understood, however, t
<p id="p-0013" num="0012"><figref idref="DRAWINGS">FIG. 1</figref> shows an agricultural implement with which the present disclosure may be used;</p>
<p id="p-0014" num="0013"><figref idref="DRAWINGS">FIG. 2</figref> is a control valve incorporated in the agricultural implement of <figref idref="DRAWINGS">FIG. 1</figref>
<p id="p-0015" num="0014"><figref idref="DRAWINGS">FIG. 3</figref> is an alternative embodiment of the control valve incorporated in <figref idref="DRAWINGS">FIG. 1</figref
<p id="p-0016" num="0015"><figref idref="DRAWINGS">FIG. 4</figref> is still another alternative of a control valve for use with the agricultural planter of <figref idref="D
</description-of-drawings>
<?brief-description-of-drawings description="Brief Description of Drawings" end="tail"?>
<?DETDESC description="Detailed Description" end="lead"?>
<heading id="h-0005" level="1">DETAILED DESCRIPTION OF THE INVENTION</heading>
<p id="p-0017" num="0016">Referring first to <figref idref="DRAWINGS">FIG. 1</figref>, there is shown rear view of an agricultural implement <b>100</b>. The agricultural im
<p id="p-0018" num="0017">A bulk seed hopper <b>112</b> is mounted on frame <b>102</b> and provides seeds in a bulk fashion. A seed metering assembly <b>114</b> receives se
<p id="p-0019" num="0018">In prior art air seeding units, the rate of flow through the various supply and distribution hoses is essential for accurate metering of the seeds
<p id="p-0020" num="0019">Valve element <b>214</b> is connected to a piston <b>218</b> displaceable in a housing <b>220</b> to provide a pressure chamber <b>222</b> on the
<p id="p-0021" num="0020">In practice, air under pressure along with seeds flowing from supply hose <b>116</b> enters inlet passage <b>208</b> and the total pressure at the
<p id="p-0022" num="0021">The control valve <b>120</b> shown in <figref idref="DRAWINGS">FIG. 2</figref> has a pair of outlet passages. It should be apparent to those skil
<p id="p-0023" num="0022">The illustrations in <figref idref="DRAWINGS">FIGS. 2 and 3</figref> show the manipulation of a control valve in purely pneumatic, mechanical fash
<p id="p-0024" num="0023">These and other advantages of the present invention will be apparent to those skilled in the art from the foregoing specification. Accordingly, it
<?DETDESC description="Detailed Description" end="tail"?>
</description>
```

The second phase consisted of using an xml parser (xml.etree.ElementTree) to parse the xml file data and separate then separate each patent contained in the provided file.

```
import re
import pandas as pd
import os
```

```
import xml.etree.ElementTree as ET
import xml.etree.ElementTree as x
import matplotlib.pyplot as plt
%matplotlib inline


d = []
s = ""
filesize= os.path.getsize("output.txt")
print(filesize)
if(filesize<=0):
    f = open("./ipg210907.xml")

    for l in f:
        if l == "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n":
            if len(s)>0:
                d.append(s)
            s = ""
        s += l
    d.append(s)
    #print(d[0])
else:
    f = open("./ipg210907.xml")
    for l in f:
        if l == "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n":
            if len(s)>0:
                d.append(s)
            s = ""
        s += l
    d.append(s)
```

Once each patent is separated into an array, the next step is extracting the description section of each patent by targeting the “description” tag in the xml file and then writing each description to a txt file, as seen below:

```
In [8]: pip install lxml

Requirement already satisfied: lxml in /Users/heathst/opt/anaconda3/lib/python3.8/site-packages (4.6.3)
Note: you may need to restart the kernel to use updated packages.

In [9]: from bs4 import BeautifulSoup as bs

In [10]: size = len(d)
#print(size)
for i in range(size):
    content = d[i]
    bs_content = bs(content, "lxml")
    description_list.append(bs_content.find_all("description"))

print(len(description_list))

7363

In [11]: print(description_list[7000])

[<description id="description">
<?RELAPP description="Other Patent Relations" end="lead"?>
<heading id="h-0001" level="1">PRIORITY</heading>
<p id="p-0002" num="0001">This patent application claims priority to U.S. Provisional Application Ser. No. 62/728
1, which was filed 6 Sep. 2018, which is titled OPTICAL MANDREL, AND FIBER-OPTIC-SENSING SYSTEM THAT INCLUDES THE
DREL, and which is incorporated by reference.</p>
<?RELAPP description="Other Patent Relations" end="tail"?>
<?BRFSUM description="Brief Summary" end="lead"?>
<heading id="h-0002" level="1">SUMMARY</heading>
<p id="p-0003" num="0002">This disclosure applies to a class of optical-fiber sensors that launch light into one
of an optical-fiber assembly and use the light reflected or scattered back from different locations or zones in t
iber to detect a disturbance and to determine where along the fiber the disturbance occurs. A system is configure
r sensing an acoustic signal incident on an optical-fiber assembly. For example, if the system is ground based, t
the acoustic signal may be generated by a vibration caused by a walking human or animal or by a moving vehicle.</
<p id="p-0004" num="0003">Applications for such a system include providing perimeter security for a ground-based
tion such as a nuclear power plant, monitoring oil, natural gas, and other types of wells, and detecting and loca
ng unauthorized crossings of a land border between two or more jurisdictions.</p>
<p id="p-0005" num="0004">To improve the ability of such a system to sense an acoustic signal, an optical-fiber a
bly may include, in addition to an optical fiber, optical mandrels spaced apart along the optical fiber. The opti

In [15]: descriptionsHolder = 'descriptions.txt'

file = open(descriptionsHolder, "w");

for i in description_list:
    file.write(str(i) + "\n")
```

jupyter descriptions.txt 17 hours ago

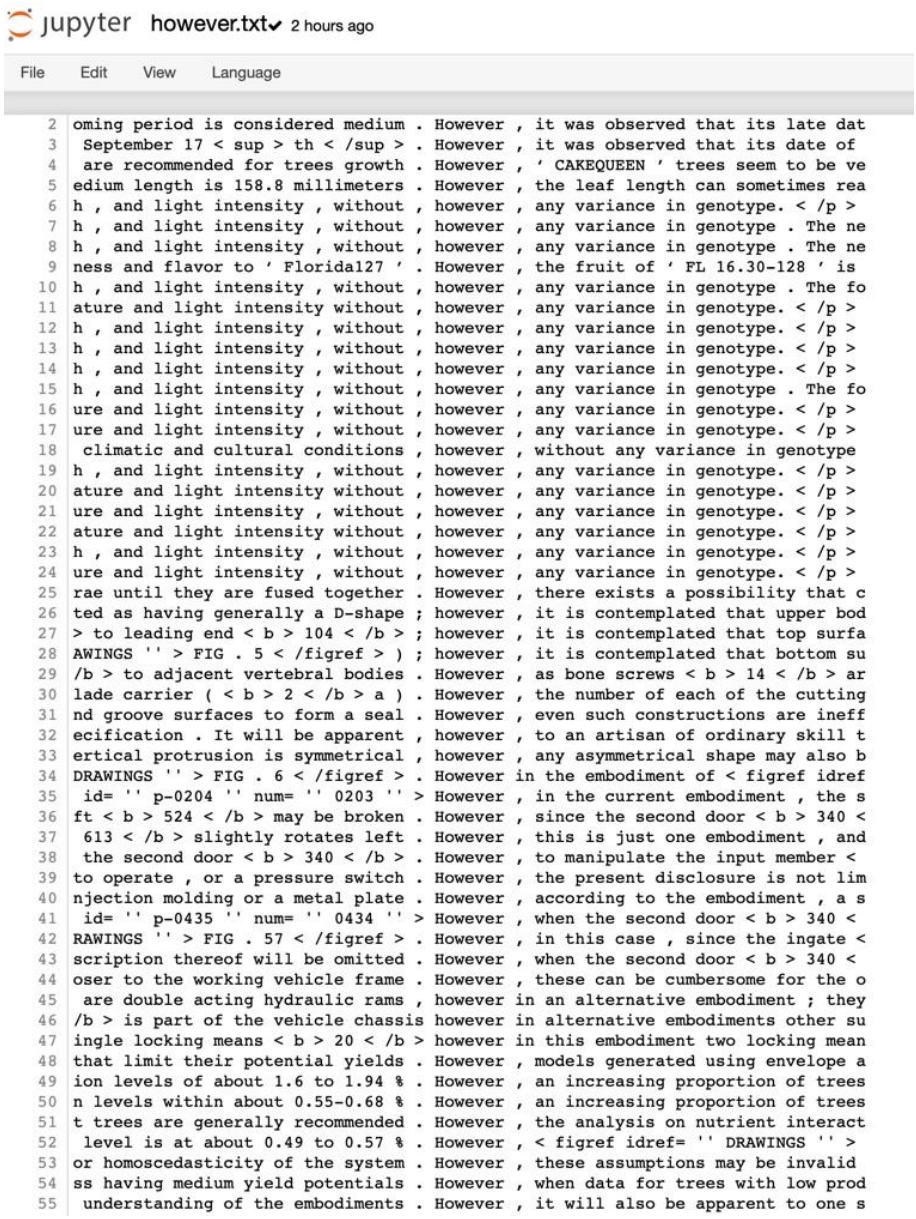
Logout

File Edit View Language Plain Text

```
[<description id="description">
<?brief-description-of-drawings description="Brief Description of Drawings" end="lead"?>
<description-of-drawings>
<p id="p-0001" num="0001"><figref idref="DRAWINGS">FIG. 1</figref> is a top, front, left side perspective view of a butter stick;</p>
<p id="p-0002" num="0002"><figref idref="DRAWINGS">FIG. 2</figref> is a top plan view thereof, the bottom plan view being a mirror image thereof;</p>
<p id="p-0003" num="0003"><figref idref="DRAWINGS">FIG. 3</figref> is a front elevation view thereof, the back elevation view being a mirror image thereof;</p>
<p id="p-0004" num="0004"><figref idref="DRAWINGS">FIG. 4</figref> is a left side elevation view thereof; and,</p>
<p id="p-0005" num="0005"><figref idref="DRAWINGS">FIG. 5</figref> is a right side elevation view thereof.</p>
<p id="p-0006" num="0006">The dot-dash broken lines illustrated in the butter stick surfaces define the bounds of the claim, forming no part thereof. The evenly-spaced broken lines at the breaks in the drawings illustrate portions of the butter stick that form no part of the claim. Any portion of the article between those breaks forms no part of the claimed design.</p>
</description-of-drawings>
<?brief-description-of-drawings description="Brief Description of Drawings" end="tail"?>
</description>]
[<description id="description">
<?brief-description-of-drawings description="Brief Description of Drawings" end="lead"?>
<description-of-drawings>
<p id="p-0001" num="0001"><figref idref="DRAWINGS">FIG. 1</figref> is an upper perspective view of a glove, showing our new design;</p>
<p id="p-0002" num="0002"><figref idref="DRAWINGS">FIG. 2</figref> is a front elevation view thereof;</p>
<p id="p-0003" num="0003"><figref idref="DRAWINGS">FIG. 3</figref> is a rear elevation view thereof;</p>
<p id="p-0004" num="0004"><figref idref="DRAWINGS">FIG. 4</figref> is a top view thereof;</p>
<p id="p-0005" num="0005"><figref idref="DRAWINGS">FIG. 5</figref> is a bottom view thereof;</p>
<p id="p-0006" num="0006"><figref idref="DRAWINGS">FIG. 6</figref> is a right side elevation view thereof; and,</p>
<p id="p-0007" num="0007"><figref idref="DRAWINGS">FIG. 7</figref> is a left side elevation view thereof.</p>
<p id="p-0008" num="0008">The dash-dash broken lines in the drawings show portions of the glove and form no part of the claimed design. The dash-dot broken lines in the drawings illustrate the boundaries of the claim and form no part of the claimed design.</p>
</description-of-drawings>
<?brief-description-of-drawings description="Brief Description of Drawings" end="tail"?>
</description>]
[<description id="description">
<?brief-description-of-drawings description="Brief Description of Drawings" end="lead"?>
<description-of-drawings>
<p id="p-0001" num="0001"><figref idref="DRAWINGS">FIG. 1</figref> is an upper perspective view of a glove, showing our new design;</p>
<p id="p-0002" num="0002"><figref idref="DRAWINGS">FIG. 2</figref> is a front elevation view thereof;</p>
<p id="p-0003" num="0003"><figref idref="DRAWINGS">FIG. 3</figref> is a rear elevation view thereof;</p>
<p id="p-0004" num="0004"><figref idref="DRAWINGS">FIG. 4</figref> is a top view thereof;</p>
<p id="p-0005" num="0005"><figref idref="DRAWINGS">FIG. 5</figref> is a bottom view thereof;</p>
<p id="p-0006" num="0006"><figref idref="DRAWINGS">FIG. 6</figref> is a right side elevation view thereof; and,</p>
<p id="p-0007" num="0007"><figref idref="DRAWINGS">FIG. 7</figref> is a left side elevation view thereof.</p>
<p id="p-0008" num="0008">The dash-dash broken lines in the drawings show portions of the glove and form no part of the claimed design. The dash-dot broken lines in the drawings illustrate the boundaries of the claim and form no part of the claimed design.</p>
</description-of-drawings>
<?brief-description-of-drawings description="Brief Description of Drawings" end="tail"?>
</description>]
```

Following the extraction of the descriptions and writing them to a txt file, topic modelling was done on the data contained in the txt file. The purpose of this is to identify keywords that signal the start of a problem statement or what problem is being solved by each patent. Results of this topic modelling process can be seen below:

```
import nltk.corpus
import sys
from nltk.text import Text

from nltk import word_tokenize

nltk.download('punkt')

contents = open('descriptions.txt','r').read()

#nltk.download('gutenberg')
#textList = Text(nltk.corpus.gutenberg.words('descriptions.txt'))

tokens = word_tokenize(contents)
textList = Text(tokens)
#collect = textList.concordance('however', lines=24197)


saveout = sys.stdout
file2 = open('however.txt', 'w')
sys.stdout = file2
collect = textList.concordance('however', lines=24197)

sys.stdout = saveout
file2.close()
```



After further discussion, it was decided that a better approach would be to use nltk concordance to identify the occurrences of words which identify the beginning of a problem statement.

6 Previous Problems

Problem: While Running The concordance Program on a Data set We found some of the description tags Data do not have words like ‘However’ Or ‘Advantages’.

Problem: Experienced errors when attempting to find occurrences of word “advantages” within the descriptions.txt file.

Problem: When extraction was attempted using Beautiful Soup there was a memory and time complexity, when using google collabs to import data files.

Solution: To solve this problem we migrated our experiments from the google collab environment to jupyter notebooks and ran the below command in the local terminal to allow for space requirements in jupyter notebook:
jupyter notebook --NotebookApp.iopub_data_rate_limit=1e10

6.1 Project Weekly Summary

After successfully extracting the Description Tags data into List . We modeled the list Data into Input Format for Concordance . We ran a concordance program on an input dataset of descriptions to do topic modeling on words like ‘However’ and ‘Advantages’**.**

7 The objective for the next week is:

Next week’s objective is to work on the ways to format the bad data (‘No However’) Need to Discuss with the professor on these issues.

You can use as many pages as you need/want for this summary

	Gordon Willingham	Raj Shah	Kushal	Anoosh G P	Heath vonn Styles
WEEK 1	Created Jupyter Notebook to document for our Patent Project	Created Detail notebook to clean xml data	Understood the actual problem we are going to solve	Analyzed problem statement and dataset	Analyze what is the problem being solved
WEEK 2	Examined XML Data of Patent website.	Presented Group Progress in Class	Gone through 20 plus patent xml files to understand the structure	Tried to extract some tags/entities from the xml file and also convert it to text file using Beautiful soup.	Examine xml
WEEK 3	Delivered In-class presentation of our week 3 progress	Extracted key information like kind ,description and tags.	Tried multiple libraries to extract description tag and found beautiful soup	Analyzing description tag in xml file to search/extract the keywords like “However”	Locate description tag in xml, for extraction

WEEK 4	Examined key-word modifiers & net number of patents.	Performing Experiments on Extracting Specific keywords.	Figuring out how to extract problem being solved by patent	Still trying to figure out how to extract description tag in particular.	Run xml text extraction tests using python libraries in jupyter notebook and visual studio
WEEK 5	Research concordance program implementation	Cleaning up the xml tags in the description.txt file	Cleaning the stop words from the description.txt file	Research about the NLTK library and how to use it to find the frequency of words.	Run more topic modelling to identify the keywords in the patents
WEEK 6	Research on Inputs for Concordance	Extraction of However word through concordance	Cleaning up the stop words from output of Concordance	Research on NLTK Library to Extract the proper words.	Debugging error when using concordance on word “advantage”

8 Conclusion:

Our data is available at:
https://developer.uspto.gov/product/patent-grant-full-text-dataxml

Our Project is available at:
https://colab.research.google.com/drive/1B5d9mMnHy9TF0GryN8B2-aqp_Z5-WzAi?usp=sharing

9 References:

1) https://developer.uspto.gov/product/patent-grant-full-text-dataxml
2) https://towardsdatascience.com/processing-xml-in-python-elementtree-c8992941efd2
3) https://www.nltk.org/
4) https://www.crummy.com/software/BeautifulSoup/bs4/doc/
5) https://jupyter.org/
6) https://evidencen.com/4-ways-to-load-data-in-jupyter-notebook-and-visual-studio-code/
7) https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
8) https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/