Patent Quest: An Unsupervised Learning Algorithm that Returns the Most Common Words Found in Recently Filed Patents and Abstracts

Ben Rogers
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
bsrodgers@ncsu.edu

Ryan Mott
Department of Computer Science
North Carolina State University
Bronxville, NY 10708
rmmott@ncsu.edu

Ralph Keyser
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
rkeyser@ncsu.edu

Abstract

The United States Patent and Trademark Office ("USPTO") maintains the records of millions of issued patents in an online database and categorizes each patent by a number of attributes, including inventor; location; applicant; and technological class. The exact nature of what each of these patents claims and describes is often a hotly contested legal issue, which requires, among other things, the testimony of expert witnesses; the advocacy of patent attorneys; and the holding of a federal judge. But there are many cases in which the exact meaning of one patent is not needed so much as the general subject matter of many patents. For example, a business may want to be apprised of the general R&D efforts of its competitor; or data scientists may want to investigate general invention trends of a geographic region. Thus, we have devised a method of finding, in response to a given query, the most common keywords used in the abstracts of each patent, and generating a WordCloud-style visualization that can be used to quickly surmise the general subject matter of a group of patents. Accuracy tests, in which the visualization will be compared to an actual opinion of a patent lawyer as to a group of patents' most significant terms, will also be conducted.

1 Background & Introduction

From the Singer sewing machine (see U.S. Patent No. 8,294) to the Edison lightbulb (see U.S. Patent No. 223,898) to the Wright Brothers' aircraft (see U.S. Patent No. 821,393) , much of American history can be told through the stories of invention and innovation that propelled it forward into new eras. A window into these stories exists through the U.S. patent system, which publishes these inventions, and the best way of making and using them, in documents called "patents."

The United States Patent and Trademark Office ("USPTO") is in charge of issuing these patents and making them accessible to the public, which it has done perpetually since its inception in 1790. As of 2023, there are now nearly 13 million patents that have been published through the U.S. patent system, which remain freely accessible by all.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The patents themselves are written by the inventor (and usually, a patent attorney), and include, among other things, a title, an abstract, a brief description of the invention and a set of patent claims. Patents are notoriously complex, and the exact nature of what they describe and claim is often subject to intense disagreement, leading to multi-year legal battles that can cost millions of dollars in legal fees.

But often, the exact nature of what a patent describes may be unimportant compared to the general subject matter that the patent represents. For example, a business may want to surveil the general progression of its competitor's research and development efforts, in order to ensure that it is pursuing similar research goals and market developments. Or perhaps, the business may want to see if the competitor is venturing into its own patents research efforts - and possibly infringing on patents of its own. As another example, a data scientist may want to collect information on general R&D trends in a given industry or region, and thus would care more about aggregated information about many patents than the exact meaning of a single patent.

Thus, there exists a need to quickly and accurately compile information about the general subject matter of publicly available patents and patent applications. Accordingly, this project seeks to use unsupervised learning methods for the purpose of gleaning information about the subject matter of not one, but many patents, by analyzing the text found in a patent's title and abstract and producing a visualization of the most common keywords found therein. The goal of the project would be to allow a user to perform a basic search – by competitor, inventor, geographic region or technological class – within a specified date range, and then produce a word cloud visualization, like the one in Figure 1, showing the most common keywords used in the patents that are included in the search.

conveyor
underneath
Defects
classifier
cooling
leaking
appearing
continues
liquid
spraying
rattling
slight
smoke
split
heard
unexpectedly
Extreme
capacitors
overheat
products
constructing
collected
cracked
fails
Strange
line
yield
touch
broken
freezes
blown
belt
Things
signs
assembler
time
floor
cuts
stuck
emitted
connect
agent
hot
fall
whine
inside
roller
shaking
occasionally
sorting
mixer
fuse
coming
sometimes
wear
get
up
supply
beginning
robot
sorter
scanner
keeps
plant
off
output
coolant
spools
showing
arm
bent
pipe
Loud
material
Burst
sound
items
programming
mixing
startup
blender
crashed
jammed
interface
controller
flick
software
high-pitched
break
Transport
tripped
electrical
power
reel
spilling
bending
overheating
engine
starting
pistons

Figure 1: Example WordCloud[2]

2 Method

2.1 Locate the Available Data Set

Although full text patents are available directly as bulk downloads from the United States Patent and Trademark Office, the sheer volume of the data – which includes every word written in every patent issued since 2002 – made it cumbersome to use. Each week, around 7,000 U.S. patents are issued, which translates to hundreds of thousands of pages of dual-column text. Since our project is focused primarily on a small subset of this text (the abstract and the title), collecting the data in bulk form would have been unwieldy and enormously inefficient.

Fortunately, data from patents (since 1976) and patent applications (since 2001) are also available as part of a mySQL relational database hosted at PatentsView.org. Using the available API query tools, we were able to pull only abstract and title data, as well as other identifying information, using a fraction of the space and bandwidth that would have been required from a bulk download. Thus, using this approach, we were able to obtain a dataset that contained the previous ten years of patent titles and abstracts - from 2013 to 2023 - and store that data in a manageable file size on our team's local machines and Google Drive. A representative sample of this data appears at `https://github.ncsu.edu/rmmott/-engr-ALDA-Fall2023-P4`.

Field	Sub-fields	Description
Applicants	First Name Last Name Organization	Provides information about the entity of record that applied for the patent.
Application	ID Filing Date	Provides information about the date the patent application was filed; and information about its unique, nominal application ID.
Assignees	First Name Last Name Organization	Provides information about the entity, if any, that has been assigned the legal rights of ownership of the patent.
CPC (Current)	CPC Class CPC Subclass CPC Group	Patents are categorized by their technological class according to the Cooperative Patent Classification ("CPC") scheme, a collection of 70,000+ classifiers that classify everything from "manure loaders" (A01C 3/04) to "thin magnetic films, e.g. of one-domain structure, characterized by the coupling or physical contact with connecting or interacting conductors" (H01F 10/06). In order of increasing specificity, the CPC classifiers are class, subclass and group, respectively[4].
Inventor	Inventor City Inventor State Inventor Country First Name Last Name	Provides information describing the identity and location of the inventor(s) of the patent.
Patent Abstract	n.a.	Paragraph describing, in high-level terms, the problem the inventor sought to overcome and how he did so through the patented invention.
Patent Date	n.a.	Date that the patent was granted.
Patent ID	n.a.	Unique, nominal identified assigned to each patent at time of grant.
Patent Title	n.a	Short, one-line descriptor of the patent provided by the inventor.

Table 1: API response fields

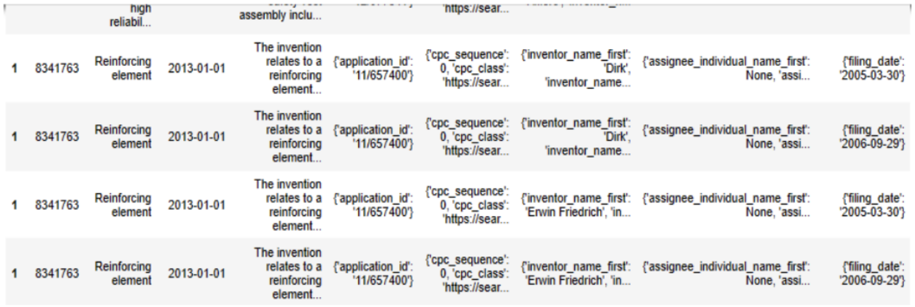
2.2 Parse the Data and Storing It for Processing

Data from the PatentsView mySQL database was collected using an API (PatentViewAPI.py), which pulled the fields shown in Table 1.

Collecting these fields enables the ability to search patents filed by an applicant organization or person; or an inventor; or an entity that was assigned rights to the patent (assignee); or from a given inventor's city/state/country; within a given date; or within a given CPC technological class; and within that it will let us assign each patent a nominal, unique identifier (patent ID or application ID) and it will let us search the words (abstract and title) for the purposes of creating a visualization or other analysis.

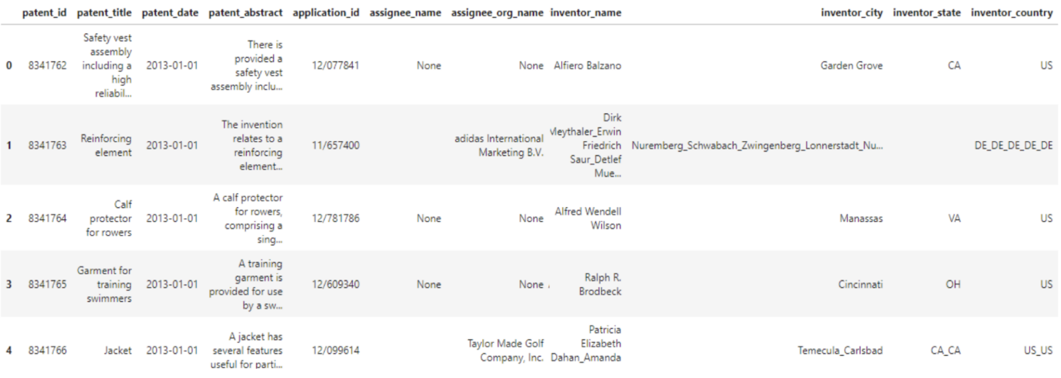
Although the data was retrieved as expected and appeared to be organized consistent with its accompanying technical documentation, there were some unexpected difficulties associated with collecting and parsing the data that required some novel approaches to concatenation. One example of an unexpected difficulty is shown below, with respect to U.S. Patent No. 8,341,769. Like many patents, U.S. Patent No. 8,341,769 has more than one inventor, but instead of listing all of the inventors as part of an "Inventor" attribute for a single patent entry, the database simply repeats the entry for each patent, with each entry corresponding to a different inventor. Thus, as shown in Figure 2, U.S. Patent No. 8,341,769 is duplicated at least five times across five different rows, with each row corresponding to a different inventor:

Duplicate patent entries, like those above, need to be managed for the purpose of our project. If they remain in the dataset without being culled, it would have the effect of duplicating abstract fields, which would have the effect of weighing more heavily the text from those patents with repeated entries. To solve this issue, we “flattened” the dataset and compressed each attribute into a single, patent-specific row, like the DataFrame shown in Figure 3.



		high reliabil...		assembly inclu...		'https://sear...			
1	8341763	Reinforcing element	2013-01-01	The invention relates to a reinforcing element...	{'application_id': '11/657400'}	{'cpc_sequence': 0, 'cpc_class': 'https://sear...	{'inventor_name_first': 'Dirk', 'inventor_name...	{'assignee_individual_name_first': None, 'assi...	{'filing_date': '2005-03-30'}
1	8341763	Reinforcing element	2013-01-01	The invention relates to a reinforcing element...	{'application_id': '11/657400'}	{'cpc_sequence': 0, 'cpc_class': 'https://sear...	{'inventor_name_first': 'Dirk', 'inventor_name...	{'assignee_individual_name_first': None, 'assi...	{'filing_date': '2006-09-29'}
1	8341763	Reinforcing element	2013-01-01	The invention relates to a reinforcing element...	{'application_id': '11/657400'}	{'cpc_sequence': 0, 'cpc_class': 'https://sear...	{'inventor_name_first': 'Erwin Friedrich', 'in...	{'assignee_individual_name_first': None, 'assi...	{'filing_date': '2005-03-30'}
1	8341763	Reinforcing element	2013-01-01	The invention relates to a reinforcing element...	{'application_id': '11/657400'}	{'cpc_sequence': 0, 'cpc_class': 'https://sear...	{'inventor_name_first': 'Erwin Friedrich', 'in...	{'assignee_individual_name_first': None, 'assi...	{'filing_date': '2006-09-29'}

Figure 2: DataFrame with repeated entries



	patent_id	patent_title	patent_date	patent_abstract	application_id	assignee_name	assignee_org_name	inventor_name	inventor_city	inventor_state	inventor_country
0	8341762	Safety vest assembly including a high reliabil...	2013-01-01	There is provided a safety vest assembly inclu...	12/077841	None	None	Alfiero Balzano	Garden Grove	CA	US
1	8341763	Reinforcing element	2013-01-01	The invention relates to a reinforcing element...	11/657400		adidas International Marketing B.V.	Dirk Vleythaler_Erwin Friedrich Saur_Detlef Mue...	Nuremberg_Schwabach_Zwingenberg_Lonnerstadt_Nu...		DE_DE_DE_DE_DE
2	8341764	Calf protector for rowers	2013-01-01	A calf protector for rowers, comprising a sing...	12/781786	None	None	Alfred Wendell Wilson	Manassas	VA	US
3	8341765	Garment for training swimmers	2013-01-01	A training garment is provided for use by a sw...	12/609340	None	None	Ralph R. Brodbeck	Cincinnati	OH	US
4	8341766	Jacket	2013-01-01	A jacket has several features useful for parti...	12/099614		Taylor Made Golf Company, Inc.	Patricia Elizabeth Dahan_Amanda	Temecula_Carlsbad	CA_CA	US_US

Figure 3: Flattened DataFrame

Each patent now has a single row, with a single abstract, corresponding to all of its accompanying attributes, including its inventors, organizations and countries of origin.

2.3 Establish Means of Searching Parsed Data According to Business/Inventor/Class/Region and Date and Returning Words in Title and Abstract

At this stage of the project, we have not fully developed the algorithm that will return a DataFrame comprising the abstracts and titles of patents responsive to a given set of inputs. Ultimately, the goal of this code will be to return a body of text that will be treated as a “bag of words” from the responsive patents from which we can perform further analysis.

2.4 Establish Algorithm for Locating 100 Most Common Words in the Titles and Abstracts

Many techniques exist for locating the 100 most common words in the dataset[1]. In general, these techniques require a process of cleaning the text, by, among other things, (a) changing all letters to lowercase; (b) removing punctuation; and (3) eliminating noise, or “stop” words, which are English words that are well-known to be of little probative value in data analysis.

One popular dictionary of stop words is the set of NLTK stop words, which includes prepositions such as “in,” “from,” or “on”; pronouns such as “you” and “I”, and commonly used verbs such as “could” or “should.” In addition, there are several commonly used words in patents – such as “invention,” “claimed,” “present,” “method,” “product,” or “describe” – will similarly be of little probative value for our particular use case. As we test the code, we intend to use the NLTK stop words and add patent-specific noise words to the dictionary until we are satisfied that we are returning a meaningful set of words.

2.5 Return Visualization of Most Common Words

Once we obtain a set of the 100 most common words in the titles and abstracts that are responses to the user request, we intend to use a WordCloud visualization tool to return these responses, in graphic form, to the user.

One example of a WordCloud tool is the one written by Andres Mueller in 2012[3]. The WordCloud tool will enable presenting the information in various shapes, colors and sizes. For example, it could return something neutral, such as a rectangle or circle, or something commonly associated with inventions, such as the lightbulb.

3 Experiment Setup

3.1 Experiment 1: Functionality Testing

Searches by Business, Inventor, and Geography.

Once we create a program that can complete the tasks above, we intend to conduct two experiments - functionality testing, wherein we test to ensure that the program performs as intended; and accuracy testing, where we compare the programs output to the claims of a small number of individual patents responsive to the same set of inputs and date ranges.

3.2 Experiment 2: Accuracy Test

Compare Visualized Words to Claims of Actual Patents

In this experiment, the 100 most common words returned in the algorithm will be compared to the words found in the actual claims of a random selection of that company's patents that were issued in the same time period. For example, if the algorithm is used to return the most common words in the abstracts of Microsoft's patents from January to March of 2017, then this experiment will take a random subset of those patents and compare the most common words to the text of the actual patent claims.

Reverse – Analyze Patent for "Important Words" and Compare to Visualization

In this experiment, we randomly pull a set of 10 patents from a single company within a given company, and have a patent attorney identify the 10 most important words used in those patents. They will then be compared to a WordCloud visualization generated using the code from the same time period. For example, if ten patents are chosen from the patents that issued to IBM between April and August of 2019, the patent attorney will first analyze the patents for the most important terms; and following that analysis, the terms will be compared to the results of our project's code and visualization.

4 Results

Based on our progress so far, we are optimistic that the program will function to create the visualizations described above. The results of the accuracy testing are currently unknown, and depending on the results, we may propose strategies to improve it.

5 Conclusion

In summary, we hope to create a searchable database of the words used in published patent abstracts and titles; devise an algorithm to search for the most common words used in a subset of this database; and create a program that returns a useful visualization that highlights these words in common. Using this code, we will run some experiments to show proof-of-concept and provide some measure of accuracy.

References

[1] Mast. Finding the most frequent words in pandas dataframe, Apr 2021.

[2] MathWorks. Visualize text data using word clouds, 2023.

[3] Andres Mueller. A little word cloud generator in python, Nov 2012.

[4] The United States Patent and Trademark Office. 905 cooperative patent classification [r-07.2015], 2023.