IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

Patent Landscape Analysis Using NLP

K Uttham[1], Mrs. Darshana A Naik[2], Dr. Seema[3]

Dept of Computer Science, Ramaiah Institute of Technology[1,2,3]

Abstract: The text in a patent is so rich that it has the capacity to contribute towards creativity and ingenuity for the engineers, policy makers and scientist of those countries which are straggling in the technology [5]. As per the reports of World Intellectual Property Organizations, patent documents contain around 95% of the inventions hence proving to be a vital source for technological fruition and development with the passage of time [6]. The patent text data is broadly classified into two domains as the patent text and metadata. Where, the patent text data encompasses title, abstract, claims, narrative and explanation and the background, while the metadata is associated with information like inventor, issue date, examiner and the one who applied for patent. A patent landscape is an analysis of patent data that reveals business, scientific and technological trends. Landscape reports typically focus on a single industry, technology or geographic region.

I. INTRODUCTION

Both text and metadata associated with patent can instigate revolutionary inclinations in the rapidly transforming technological industry. Since there could be an enormous amount of data associated with patents in a specific industry, this is where tools like NLP (Natural Language Processing) comes in to extract the required information from the patent texts. NLP is acknowledged as a powerful tool to discern, construe and comprehend the contemporary technological landscapes [7]. NLP has the capacity and trend of being incorporated in the parsing of sentiment, subject segregation and identification, rapport mining and domain of machine translations. NLP in amalgamation with machine learning has been doing wonders in patent comprehension along diverse fields like sustainability, design process and the health sector [5, 6].

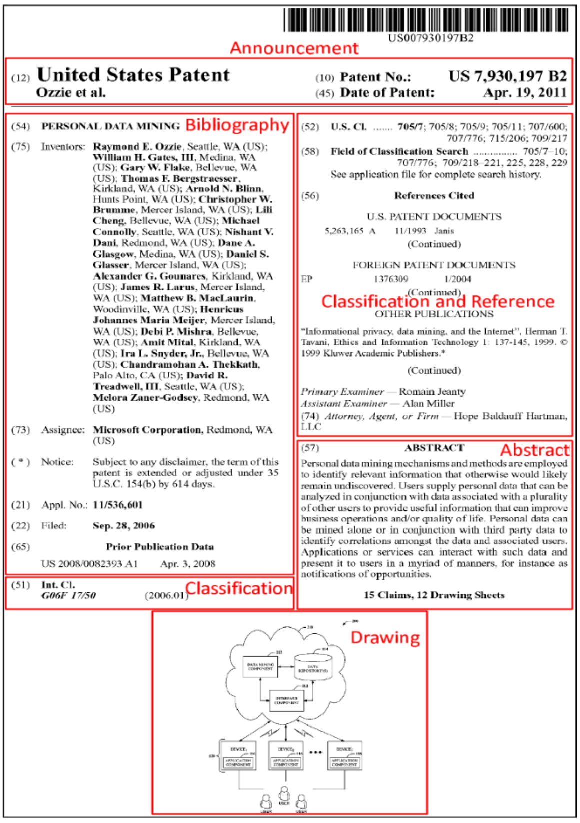
Recently NLP has been employed in a range of data mining and information extraction complexities in the patent landscape domain. A review of the literature indicates that NLP has been incorporated for patent data of USPTO (United States Patent & Trademark Office) to analyze, derive, and disambiguate it and consequently develop an SQL database. Furthermore, in this study an advanced novelty evaluation system was developed built specifically around the term "novelty" [7]. Similarly, NLP has been employed in the telecom industry to identify and reveal the patent registration trends on the basis of company names and assigning years by using topic modelling and applying it to more than 150,000 full text patents. Apart from the technology association and forecast, NLP has been employed in the domain of technology classification. This has been achieved by focusing at a specific technology stream and target the complexities, excessive time consumption and higher costs associated with manual classification. The solution of NLP based automatic classification was achieved through consolidation of active learning with the fusion of multi-classifier [8].

Similarly, to provide meaningful and instant insights to researchers and associated professionals, NLP has been employed for information extraction from the patent documents. NLP has been employed to develop problem graphs from the corpus text of patents through reorganization and streamlining. This NLP based tool presented the contemporary state of art in the specified fileds [9]. Likewise, NLP based genre eccentricities evaluation of the patents was conducted to summarize the patents and hence acquire the necessary information of invention [10]. In this study, NLP was aimed at the analysis of lexical chains and text fragments present in the claims and summaries for relevancy evaluation. Moreover, NLP has been found fruitful in determining the preposition based association between the key words and the congruent technologies through semantic analysis [11]. In addition to these, NLP because of its efficacy and fruitfulness has been used in combination with machine learning and open source implementations of the web oriented visualizations to augment the text and data mining of patent text corpus, its semantic analysis, shortcomings and insights retrieval [12].

Figure 1.1: Front Page of a Patent Document

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022
DOI: 10.17148/IJARCCE.2022.11817

Structure of Patents:



Patents are legal documents issued by a government that grant the owner of an idea or an invention a set of exclusive rights and protection. The right of exclusivity allows the patent owner to prevent others from making, using, selling, offering for sale, or importing the patented invention during the patent term, which typically lasts from the date of filing to the date of expiration, and in the country or countries where patent protection exists.

Patent documents are one of the most important components in protecting patent owners' intellectual property rights. Patents and inventions are two distinct but related concepts: patents are legal documents, whereas inventions are the main subject of patents. Different countries or regions might have their own patent laws and regulations; nevertheless, there are two sorts of patents that are commonly used: utility patents and design patents.

Utility patents describe technological solutions relating to a product, method, or useful improvement, whereas design patents frequently represent original designs related to product specifications. In practice, due to the distinct properties of these two types of patents, the structure of the patent document may differ slightly; nevertheless, a typical patent document often contains several requisite sections, such as a front page, detailed specifications, claims, declaration, and/or a list of drawings to illustrate the solution's idea.

II.LITERATURE SURVEY

Longhui Zhang, et.al [1], mentioned about presenting a description of existing patent mining research activities, patent analysts and interested readers get a bigger picture of patent mining. The paper first gives an overview of the structure of a general patent. Before publication of patents, each patent is assigned one or more classification codes based on their textual contents for efficient management and retrieval.

The sophisticated structure of patent documentation frequently hinders analysts from quickly comprehending the core concept of patents.

Patent documentation are at the heart of many technological organizations and businesses. It is critical to analyze the quality of patent documents for additional actions in order to support decision making. In practice, patent valuation refers

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁞ Impact Factor 7.39 ⁞ Vol. 11, Issue 8, August 2022
DOI: 10.17148/IJARCCE.2022.11817

to a standard technique of evaluating the importance/quality of patent documents, with the goal of assisting internal decision making for patent protection tactics.

Patent documents are highly sensitive to region, which means that patents from various regions may be documented in various languages. A patent may be described in several languages, however localized patent information is preferred by analysts of patents. For instance, a patent document may be prepared in English, but a Korean analyst anticipates that it will also be translated into Korean for ease of understanding. In principle, a cross-language patent retrieval system can be created using machine translation and semantic correspondence.

Yucesoy, Selen, et al [2], This study discusses several trials to improve patent categorization accuracy.

10 trials and 46 sub-trials are organized in this study, with the general framework preserved for each trial. The presence of extra stop words, phrases, or distinct elements of patent texture distinguishes the trials.

The classification procedure is divided into four steps: (a) preparing the framework for patent classification; (b) preprocessing; (c) keyword extraction analysis; and (d) patent classification, which comprises k-nearest neighbor extraction and SVM analysis.

Following questions are answered:

• In the analysis, which part of the patent should be chosen?

• In each trial, which classification algorithm should be used?

To generate phrases in this work, the RapidMiner 7.1 software's n-gram generation function is used.

The purpose of this paper is to determine which combination of patent sections is most suited and offers the most accurate results in patent classification. The results of the trials have demonstrated that the SVM method produces better outcomes.

When compared to using all textual information, the combination of abstract, title, and description yields similar outcomes.

Mattas, Nisha, et al [3], This paper compares the key features of different techniques used for mining hidden patterns from patents.

Main Text Mining approaches used: Natural Language Processing, Rule-Based Techniques, Sematic Analysis Based and Neural Networks based Techniques.

The typical steps in Visualization Approach include: Ahead Citation, Niche Patent and Reverse Citation.

Patent Analysis has proved to play an important role in defining business strategy in various organizations and strategic planning with its help is also useful in competitive intelligence.

This paper concludes an overview of widely adopted techniques along with their advantages and limitations.

Ivanov, Alexander, and Zeljko Tekic,[4] This paper talks about problems in Patent search, namely

1. Patent language. Often contain specific words or new words related to the patent

2. Number of patents and their scope

Overall aim is to create new analytical methods for producing intelligence related to technological innovation and technology business based on combining patent data with contextual data.

BFS search is used with blogs about input as patents and list of raw articles which mentions patents as output. This raw data is used to obtain links to specific patents. These links need to be cleaned from any accidental links and other data which is not needed. Referred to as pruning. The cleaned data is used to extract valuable information about patents.

The main result of this paper is not only the algorithms and software, but the data found.

This demonstrated that contextual data can provide useful information and improvements to patent databases. Many intellectual property management activities will be simplified by adding keywords that characterise product features and

ISO 3297:2007 Certified ⁒ Impact Factor 7.39 ⁒ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

goods. In the case of the Jeep patent, the authors demonstrated that current patent databases do not allow for a thorough search, whereas their approach and findings do.

F. Madani et.al [5], Stated that the text in a patent is so rich that it has the capacity to contribute towards creativity and ingenuity for the engineers, policy makers and scientist of those countries which are straggling in the technology. In this paper they offer a complete overview of the many options for privacy, authentication, and message propagation that solve these challenges in this study.

A. Souili et.al [6], in this paper they found that as per the reports of World Intellectual Property Organizations, patent documents contain around 95% of the inventions hence proving to be a vital source for technological fruition and development with the passage of time.

A. Suominen et.al [7], in this paper they suggested that the patent text data is broadly classified into two domains as the patent text and metadata. Where, the patent text data encompasses title, abstract, claims, narrative and explanation and the background, while the metadata is associated with information like inventor, issue date, examiner and the one who applied for patent.

Both text and metadata associated with patent can instigate revolutionary inclinations in the rapidly transforming technological industry. Since there could be an enormous amount of data associated with patents in a specific industry, this is where tools like NLP (Natural Language Processing) comes in to extract the required information from the patent texts. NLP is acknowledged as a powerful tool to discern, construe and comprehend the contemporary technological landscapes

R. McEntire et al [8], G. Cao, et al [9], in this paper NLP has the capacity and trend of being incorporated in the parsing of sentiment, subject segregation and identification, rapport mining and domain of machine translations. NLP in amalgamation with machine learning has been doing wonders in patent comprehension along diverse fields like sustainability, design process and the health sector.

B. Balsmeier et al [9], Recently NLP has been employed in a range of data mining and information extraction complexities in the patent landscape domain. A review of the literature indicates that NLP has been incorporated for patent data of USPTO (United States Patent & Trademark Office) to analyze, derive, and disambiguate it and consequently develop an SQL database. Furthermore, in this study an advanced novelty evaluation system was developed built specifically around the term “novelty”.

Different vulnerabilities exist as a result of VANET's opposing features, causing the network to crash. Numerous researchers have focused on various security vulnerabilities and solutions in VANET during the last decade. As a result, this article presents a thorough examination of the most recent security-related detection approaches in VANET. The research is based on the VANET attacks and their countermeasures, as revealed by the survey.

X. Zhang et al [10], has been employed in the telecom industry to identify and reveal the patent registration trends on the basis of company names and assigning years by using topic modelling and applying it to more than 150,000 full text patents. Apart from the technology association and forecast, NLP has been employed in the domain of technology classification. This has been achieved by focusing at a specific technology stream and target the complexities, excessive time consumption and higher costs associated with manual classification. The solution of NLP based automatic classification was achieved through consolidation of active learning with the fusion of multi-classifier

Similarly, to provide meaningful and instant insights to researchers and associated professionals, NLP has been employed for information extraction from the patent documents. NLP has been employed to develop problem graphs from the corpus text of patents through reorganization and streamlining. This NLP based tool presented the contemporary state of art in the specified fields [6].

Likewise, NLP based genre eccentricities evaluation of the patents was conducted to summarize the patents and hence acquire the necessary information of invention [13]. In this study, NLP was aimed at the analysis of lexical chains and text fragments present in the claims and summaries for relevancy evaluation. Moreover, NLP has been found fruitful in determining the preposition-based association between the key words and the congruent technologies through semantic analysis [14]. In addition to these, NLP because of its efficacy and fruitfulness has been used in combination with machine

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

learning and open-source implementations of the web oriented visualizations to augment the text and data mining of patent text corpus, its semantic analysis, shortcomings and insights retrieval [5].

III.IMPLEMENTATION

Algorithm: Patent Parsing

Patents that have been downloaded in bulk are in .xml format. This is problematic for Python pandas to use, and therefore it is converted to .csv format. The resulting .csv file is usable by pandas and is simple to read and modify to satisfy our requirements. Each patent in the xml file begins with the following line:

Each patent in the xml file begins with the following line:

<\?xml version\=\"1\.0\" encoding\=\"UTF\-8\"\?>

The script builds a Python dictionary with a set of keys and values, and thereafter scans the XML document for the set of keys and appends them to their respective key value pairs in the dictionary. Afterwards, the dictionary can be easily converted into a data frame and then into a .CSV file.

Bibliographic Analysis:

Bibliographic analysis is a technique for analyzing and interpreting huge quantities of scientific data. In this instance, it assists in obtaining relevant information from a large number of patents. The script displays four informative graphs from which we can draw various inferences. The function bibliographic_plots() outputs various graphs and pie charts for us to infer from.

• The plot function, which displays the number of patents assigned to each assignee or company, first turns the current pandas series into a list that can be separated and sorted into a dictionary from high to low number of patents. The outcome is represented on a bar graph as Assignees/Number of Patents.

Content Analysis: To examine content and its properties, content analysis is a qualitative research method or methodology that is frequently employed. It is a technique for quantifying qualitative information by classifying and contrasting numerous data points in order to condense it into information that is useful. When conducting content analysis, NLP is used.

• First the various stops words in English language are imported. Custom words can also be added. Stop words are common English words for example “a”, “has”, “the”, “is”, “are” and etc. These words need to be removed as they carry very little information.

• Next, key phrase extraction is done. This is extracting most common words, and a BoW (Bag of Words) corpus is created. BoW is a set of vectors containing the count of word occurrences in the document. BoW vectors are easy to interpret. This BoW is created with the help of Python module Gensim. Gensim is an open-source library for unsupervised topic modeling and NLP.

• Topic modelling is then used to identify the best themes. Finding the topics that best characterize a group of documents is known as topic modelling. These topics won't appear before the topic modelling procedure. The topics' coherence scores are calculated. The coherence score calculates how closely related words are to one another within a topic. The degree of semantic similarity between high scoring terms in a topic is used to calculate the coherence score for that topic. These assessments aid in differentiating issues that can be understood semantically from topics that are the result of statistical inference. This again is done with the help of the Gensim module. The result is a dictionary of LDA models and coherence scores. The LDA model with the maximum coherence score is taken. This is commonly referred to as the best LDA model. It is used to sort topics accordingly.

• The final result from this is the best LDA model, the BoW corpus and a dictionary.

pyLDAvis:

pyLDAvis is a python module that helps in visualizing LDA models and see the most frequent terms in a selected topic. The result is a set a topics and the top 30 frequent terms. It is up to the user to interpret what domain these topics belong to. The frequent terms will give a basic idea on what the topic they are related to.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

IV.RESULTS

Figure 1.2 shows the top 20 assignees and the number of patents filed by them. This gives us an insight into which company or individual is filing the most number of patents. This also gives us a small insight into which domain most patents are being filed.

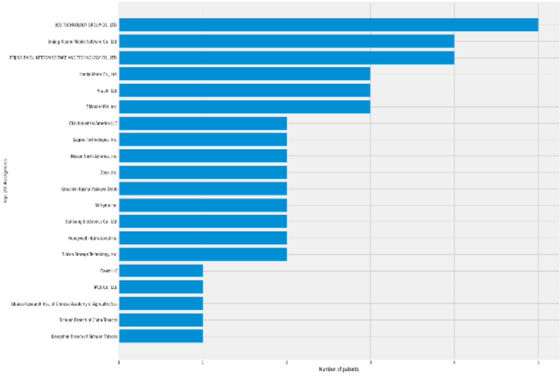


Figure 1.2: Number of Patents per company/assignee

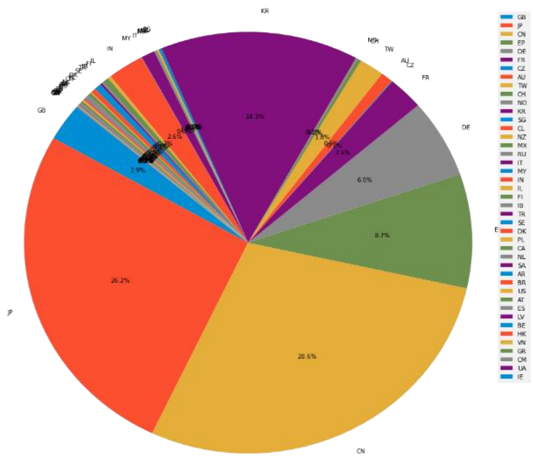


Figure 1.3: Prior Country

Figure 1.3 shows in which countries the patents were filed before being filed under USPTO. As we can see the majority of the patents are from China, Japan and Korea. The USPTO site has a list Country codes.

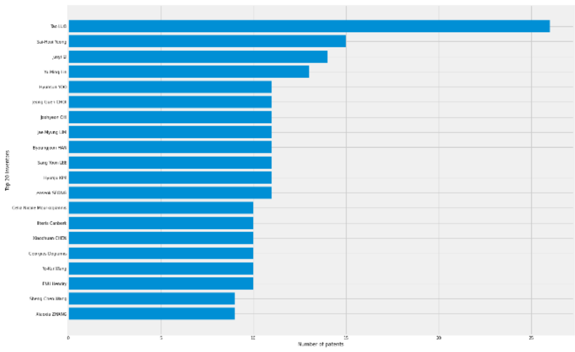


*Figure **Error! No text of specified style in document.Error! No text of specified style in document.**:1.4: Top inventors and Number of Patents*

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 8, August 2022
DOI: 10.17148/IJARCCE.2022.11817

The Figure 1.4 shows the top inventors and the number of patents they have filed. This gives us an insight into who is filing the most patents irrespective of the company.

	Number of Topics	Coherence Score
77	79	0.588726
72	74	0.588505
80	82	0.586475
81	83	0.584807
68	70	0.584238
...	...	...
10	12	0.487020
7	9	0.485729
6	8	0.484895
3	5	0.478913
5	7	0.468883

99 rows × 2 columns

Figure 1.5: Coherence Scores

Figure 1.5 shows the coherence scores for 100 topics. It is ordered from highest to lowest score.

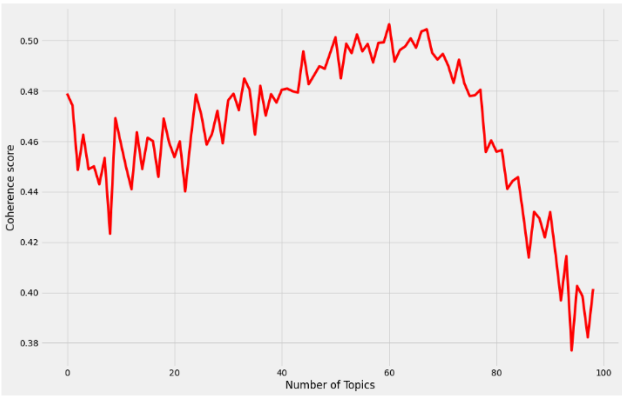


Figure 1.6: Coherence Score vs Number of Topics Graph

From figure 1.5 and 1.6 we can see that the coherence score increases as the number of topics increase, and from Fig 1.5 we can see that as the number of topics increase the coherence score either decreases or has very minute difference to average it around 0.513. The best LDA model having a score of 0.577.

The PyLDAvis module shows us a visual representation of the most prominent terms. The following results are for patents filed in the year 2021 up to the 30th of December. 9846 patents are present.

 ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⪤ Impact Factor 7.39 ⪤ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

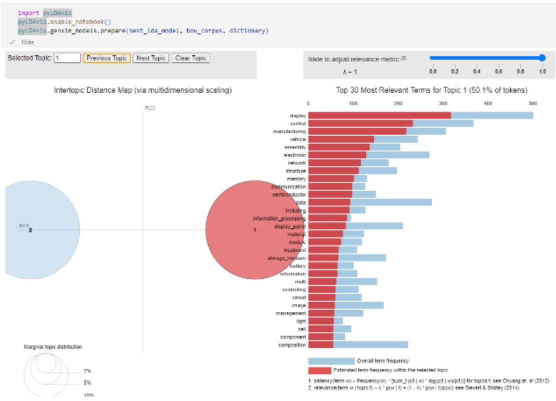


Figure1.7: PyLDAvis Results

Fig 1.7 shows the pyLDAvis module which shows the most relevant terms for a topic. Each bubble of the left represents a topic. The larger the bubble, the higher percentage of the number of relevant terms in the corpus is about that topic.

Blue bars represent the overall frequency of each word in the corpus. If no topic is selected, the blue bars of the most frequently used words will be displayed. Red bars give the estimated number of times a given term was generated by a given topic. As we can see from Fig 6.11, there are around 500 occurrences of the word ‘display’, and this term is used about 300 times within topic 1. The word with the longest red bar is the word that is present most in that respective topic.

The above results are for patents filed in the year 2021 up to the 30th of December. 9846 patents are present.

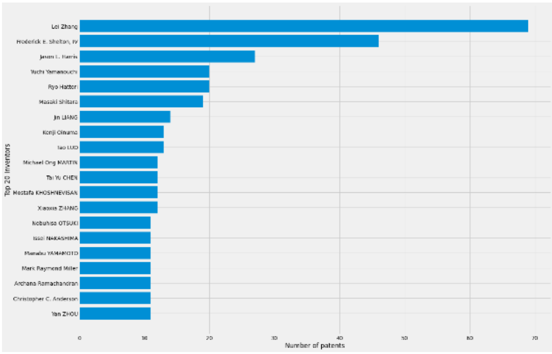


Figure 1.8: Top 20 Assignees

Figure 1.8 shows the top 20 assignees and the number of patents filed by them. This gives us an insight into which company or individual is filing the most number of patents. This also gives us a small insight into which domain most patents are being filed.

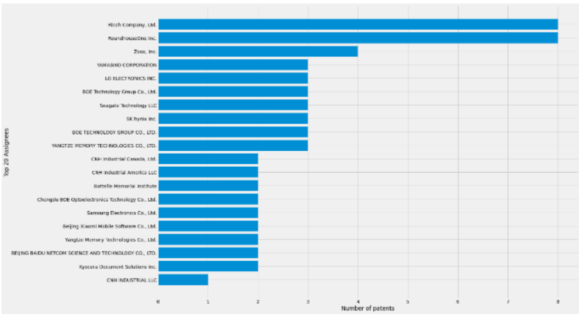


Figure 1.9: Top Inventors

The Figure 1.9 shows the top inventors and the number of patents they have filed. This gives us an insight into who is filing the most patents irrespective of the company.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

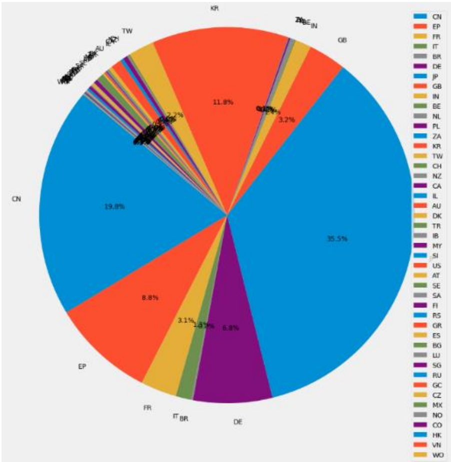


*Figure **Error! No text of specified style in document.**1.10: Prior Country*

Figure 1.10 shows in which countries the patents were filed before being filed under USPTO. As we can see the majority of the patents are from China, Japan and Korea.

	Number of Topics	Coherence Score
27	29	0.509367
24	26	0.508259
28	30	0.504863
38	40	0.502186
46	48	0.501426
...	...	...
81	83	0.412469
69	71	0.412139
75	77	0.409691
77	79	0.404341
80	82	0.402093

99 rows × 2 columns

Figure1.11: Coherence Scores

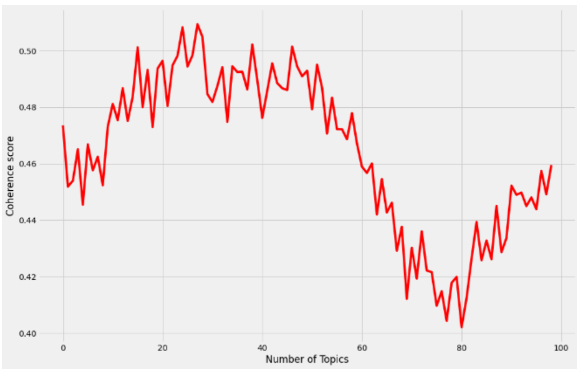


Figure 1.12: Coherence Scores vs Number of Topics

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11817

From figure 1.11 and 1.12 we can see that the coherence score increases as the number of topics increase. And from Fig 1.12 we can see that as the number of topics increase the coherence score either decreases or has very minute difference to average it around 0.464. The best LDA model having a score of 0.509.

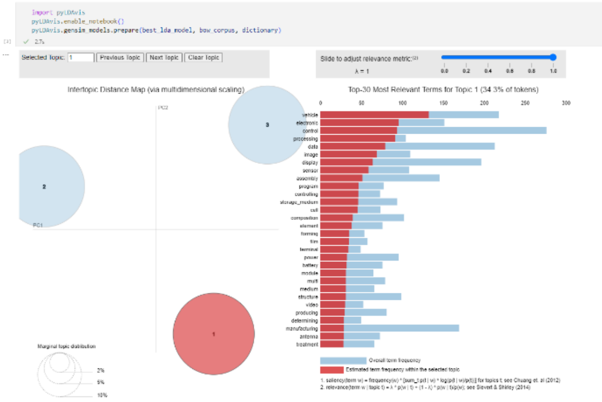


Figure 1.13: PyLDAvis Result

Fig 1.13 shows the pyLDAvis module which shows the most relevant terms for a topic. Each bubble of the left represents a topic. As we can see from Fig 1.13, there are a little more than 200 occurrences of the word 'vehicle', and this term is used about 100 – 150 times within topic 1.

The above results are from patents filed in the year of 2020 up to 31st of December. 8759 patents are present.

V.CONCLUSION

In the year 2021 most the patents were field in the fields of display and electronics.

Overall, at the end we get a fairly good insight into which domain patents are being filed and from where majority of the patents are being filed from. However, the end result of the pyLDAvis module is left up to the user to interpret which domains the topics belong to with respect on their views.

REFERENCES

1. Zhang, Longhui, et al. "Patent Mining: A Survey: ACM SIGKDD Explorations Newsletter: Vol 16, No 2." ACM SIGKDD Explorations Newsletter, dl.acm.org, 1 Dec. 2014, https://dl.acm.org/doi/abs/10.1145/2783702.2783704.
2. Yucesoy, Selen, et al. "Patent Classification via Textual Analysis Which Sections to Be Included?" Patent Classification via Textual Analysis Which Sections to Be Included? ieeexplore.ieee.org, 28 Sept. 2018, https://ieeexplore.ieee.org/document/8620929/authors#authors.
3. Mattas, Nisha, et al. "Comparing Data Mining Techniques for Mining Patents." Comparing Data Mining Techniques for Mining Patents, ieeexplore.ieee.org, 21 Feb. 2015, https://ieeexplore.ieee.org/document/707908
4. Ivanov, Alexander, and Zeljko Tekic. "Using Contextual Data for Smart Patent Analysis." Using Contextual Data for Smart Patent Analysis, ieeexplore.ieee.org, 12 Dec. 2016, https://ieeexplore.ieee.org/document/783071
5. F. Madani and C. Weber, "The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis," *World Patent Information*, vol. 46, pp. 32-48, 2016/09/01/ 2016.
6. A. Souili, D. Cavallucci, and F. Rousselot, "Natural Language Processing (NLP) – A Solution for Knowledge Extraction from Patent Unstructured Data," *Procedia Engineering*, vol. 131, pp. 635-643, 2015/01/01/ 2015.
7. A. Suominen, H. Toivanen, and M. Seppänen, "Firms' knowledge profiles: Mapping patent data with unsupervised learning," *Technological Forecasting and Social Change*, vol. 115, pp. 131-142, 2017/02/01/ 2017.
8. R. McEntire *et al.*, "Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development," *Drug Discovery Today*, vol. 21, no. 5, pp. 826-835, 2016/05/01/ 2016.
9. G. Cao, P. Luo, L. Wang, and X. Yang, "Key Technologies for Sustainable Design Based on Patent Knowledge Mining," *Procedia CIRP*, vol. 39, pp. 97-102, 2016/01/01/ 2016.
10. B. Balsmeier *et al.*, "Machine learning and natural language processing on the patent corpus: Data, tools, and new measures," *Journal of Economics & Management Strategy*, https://doi.org/10.1111/jems.12259 vol. 27, no. 3, pp. 535-553, 2018/09/01 2018.

11. X. Zhang, "Interactive patent classification based on multi-classifier fusion and active learning," *Neurocomputing*, vol. 127, pp. 200-205, 2014/03/15/ 2014.
12. A. Souili, D. Cavallucci, and F. Rousselot, "A lexico-syntactic Pattern Matching Method to Extract Idm- Triz Knowledge from On-line Patent Databases," *Procedia Engineering*, vol. 131, pp. 418-425, 2015/01/01/ 2015.
13. J. Codina-Filbà *et al.*, "Using genre-specific features for patent summaries," *Information Processing & Management*, vol. 53, no. 1, pp. 151-174, 2017/01/01/ 2017.
14. J. An, K. Kim, L. Mortara, and S. Lee, "Deriving technology intelligence from patents: Preposition-based semantic analysis," *Journal of Informetrics*, vol. 12, no. 1, pp. 217-236, 2018/02/01/ 2018.