Patent Recommendation System

CMPE -256 Group 6

Amrita Kasaundhan
Reg. No: 013854204
Computer Engineering Department
San Jose State University
amrita.kasaundhan@sjsu.edu

Shivangi Nagpal
Reg. No: 013852696
Computer Engineering Department
San Jose State University
shivangi.nagpal@sjsu.edu

Nidhi Tattur Aravinda Kumar
Reg. No: 013845494
Computer Engineering Department
San Jose State University
nidhi.tatturaravindakumar@sjsu.edu

Mahesh Reddy Konatham
Reg. No: 013823095
Computer Engineering Department
San Jose State University
maheshreddy.konatham@sjsu.edu

Arpit Sharma
Reg. No: 013826722
Computer Engineering Department
San Jose State University
arpit.sharma@sjsu.edu

***Abstract*—"A patent is a form of intellectual property that gives its owner the legal right to exclude others from making, using, selling, and importing an invention for a limited period of years, in exchange for publishing an enabling public disclosure of the invention." During this era of innovation with total more than 50,000 patent file by only one country, it would be helpful to know if we get a system that gives related already filed patents to avoid duplicate while filing for patents. This will help organization to deal with such huge increase in patents counts. Several Machine learning models have been applied to the system to get the best results. The system is based on extracting the keywords from title and abstract of patents and gives the top five similar patents.**

Keywords—Content based filtering, TF-IDF, Web scraping, LDA, Patent recommendation System

I. Introduction

Impalpable assets such as Intellectual Property Rights (IPR) has a very significant role in the net worth of industries. In particular, the intellectual property needs to be registered internationally as patents and is used to legally protect the proprietary technology and ensures that the market advantage of the firm, promote further commercialization, royalties, licensing, and sales. When new technology is developed and a patent is issued, official claims ensure that the assignee(s) maintain their competitiveness by preventing others from using the patented technology without prior permission. As declared by US Patent Law (US Patent Act), patents are to be used to encourage and promote commercial development by providing legal protection. Whoever without authority makes, uses, imports, or sells any patented invention during the term of the patent stands in violation of the law and infringes upon the patent assignee. Certainly, companies with high quality patents hold a competitive and sustainable market position. This paper is organized into several sections. In Section 2, ease of use briefs about how system identifies and recommends related patents to the user. Section 3 describes the literature review, Section 4, is reviewed with a focus on data collection and data cleaning. Section 5, proposed methodology and algorithms used to formulate the user search. Section 6 Finally, summarizes the conclusion of the research

II. Ease of Use

The World Intellectual Property Organization reports the sharp increase in number of patents filed per year. A company or individual gets overwhelmed by the difficulty of reviewing, and analyzing illustrations, information, claims, and technical knowledge whenever they attempt to search, interpret, compare, and classify patent documents. Thus, to facilitate the manual processing of information about the patents, a computer assisted patent information and knowledge management systems are needed.

While the numbers of patents application filed by US based firms continues to grow, the patent recommendation system must be useful for finding related patents in rapid and effective manner. Thus, this research gives in insight of an intelligent recommendation methodology based on the users' search. The research implements a content based filtering approach to construct the intelligent patent recommendation system. When a user searches patents in the system, the system automatically identifies and recommends related patents to the user.

A company or individual inventor may gain or lose tangential profits and other market advantage if their innovation unknowingly conflicts with any already existing technology. Hence Its mandate to search, review, and interpret patents in different patent databases to understand prior inventions before filing new patent or any new product and to maintain its legal authority. Hence, such recommendation methodologies plays pivotal function of a patent recommendation system. When a user lacks the domain knowledge of their patent research, efforts are often impeded which leads to relevant searches.

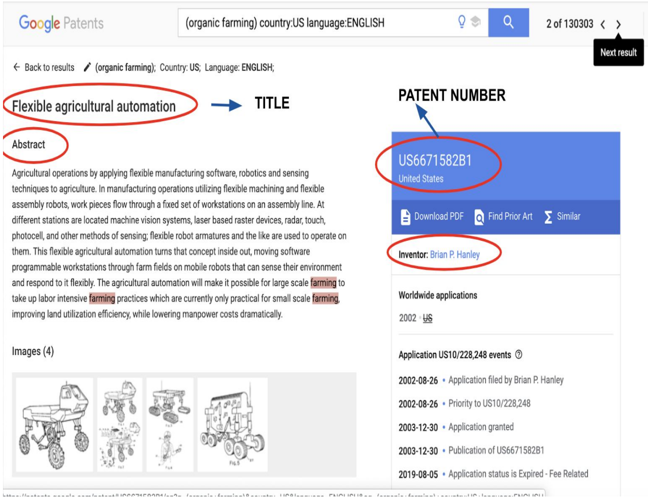
III. Literature Review

The purpose of the patents is to protect the invention and intellectual property rights (IPR) which is claimed by the inventor. Inventors is to write their research and development findings or any form of innovation as patent documents by following a required patent document format with a specification guideline. the patent evaluations committee then evaluates for patent's qualification. This qualifies the patent owner(also called the patent assignee) to take legal action to prevent unauthorized manufacturing, selling or usage of the invention. In this section, related literature covering patent search, patent analysis, and patent recommendation systems is described.

Different patent databases are used by the organizations or individuals to search the patents, e.g. the United States Patent and Trademark Office (USPTO), the Indian Patent Office (IPO), and the global Intellectual Property Organization (WIPO). Additionally, They also use national and international available open databases as well as the databases which requires subscription fees or database purchases.. These databases have various functions for the patent search, i.e., a simple keyword search, assignee, inventors, and year etc., and also a patent number search. When users search patents without using much of defined specific keywords or search options, they usually end up getting search results with errors where they either miss patents or retrieve the wrong patents from the database. Therefore, to ease the work many automatic approaches have been proposed for patent search. The time and cost of search can be reduced if patents are classified before searching,. The Indian Patent Office (IPO) uses a clustering approach based on the k-nearest neighbor algorithm. Researchers have outlined that the patent classification can be more accurate by considering both the patent's metadata and the full abstract of the patent (Richter and MacFarlane, 2005). Trappey et al. (2010) propose a clustering method to group the patent documents into the form of overlapping clusters. This approach determines whether a given patent can possibly be categorized into multiple clusters, which is consistent with the principle of a patent allowing multiple claims. Further, Chiang et al. (2011) implemented a system which takes automated binary knowledge document classification and does content analysis on the document. The system is actually designed using a back-propagation artificial neural network(also called RNN), and normalized term frequency methods with a purpose of improving the patent classification in a binary and hierarchical manner. Therefore, the system keeps on identifying patents iteratively until a sufficient reduced number of highly related patents are collected. After this process, the collected patents are then analyzed to extract the exact detailed information. Identification of technology trends as well as the performance of legal due diligence is included in the analysis. There are four major analytical approaches used which includes time series analysis, patent citation analysis, international patent classification (IPC) analysis, and the construction of patent maps. The purpose of time series analysis is to analyze the change in the number of patents applied for overtime. Its patent citations that determine the difference between two different patents by using their citations and other references. IPC analysis helps to determine which technology is being developed. If some IPC regions are restricted to a small number of patents, it may turn up for R&D bottleneck and for business potential. Thus, the output of IPC analysis gives the information to support government or enterprise R&D strategy. As a result, its patent maps that are used to draw the potential relationships between two patent group or clusters.Different patent groups described by patent map that belong to technology groups or assignees Huang and Li (2010). Here, the feedback given by user is a type of rating that users express their satisfaction with a search. And this behavior is not necessarily the explicit one. (Nichols, 1997). The explicit feedback means of representing the information that is directly provided by the user in the form of personal information, common replies, survey results, and work experience. And the Implicit feedback is something that is collected indirectly from user and it is usually extracted from user's actions or browsing records or query logs. Oard and Kim (1998) says that the implicit information given by user can be extracted and can be classified as three behavioral types which in turn can serve as to infer data information for the recommendation system. To get the feedback loop the researchers have also considered various techniques and technologies to define user behaviors and methods to trace their queries and feedback. For example, a matchmaker service plays an important role to ensure that the proper connection between the user and the service provider. However, the lack of relevant service domain knowledge and incorrect service queries may end up preventing matchmakers from identifying the service concepts that correctly represent the service requests. Wang et al. (2007) explains that the information coverage and update problems are also common obstacle to give accurate result given current web search engines. These problems limit the services offered by enterprise and make the resolution of complaints difficult to achieve.Hence, a new search algorithm based on DNS is proposed in their research to solve the problems. In this approach, the system follows a layered distributed architecture, which is different from current commercial search engines. Semantic similarity model for describing the service ontology environment whereas Bouras and Poulopoulos (2012) propose a web personalization mechanism based on dynamic creation and automatic updates of user profiles to better match users preferences by Dong et al. (2011). This approach assumes that a user's profile is affected by other user's grouping details which are constructed with similar profiles. As a result, a real-time user-centric document grouping mechanism is implemented to support the web personalization system and provide data for experimental evaluation of the system.

IV. Data collection and Data analysis

Data Collection: we collected the data from (https://patents.google.com) using Selenium web driver and stored the results in .csv file. The function of the web scraper is to extract “Abstract”, “Title”, “Patent Number”, “classification”, “inventor”, “current Assignee” from the page store it in .csv file and save it, then run loop to iterate through pages for predefined many times. The use of Selenium web driver reduced and eased the efforts of collecting data manually from the site thousands of times. Roughly we have collected thousands of rows to train and test the built model.



← Back to results (organic farming); Country: US; Language: ENGLISH;

Flexible agricultural automation

Abstract

Agricultural operations by applying flexible manufacturing software, robotics and sensing techniques to agriculture. In manufacturing operations utilizing flexible machining and flexible assembly robots, work pieces flow through a fixed set of workstations on an assembly line. At different stations are located machine vision systems, laser based raster devices, radar, touch, photocell, and other methods of sensing; flexible robot armatures and the like are used to operate on them. This flexible agricultural automation turns that concept inside out, moving software programmable workstations through farm fields on mobile robots that can sense their environment and respond to it flexibly. The agricultural automation will make it possible for large scale farming to take up labor intensive farming practices which are currently only practical for small scale farming, improving land utilization efficiency, while lowering manpower costs dramatically.

US6671582B1
United States

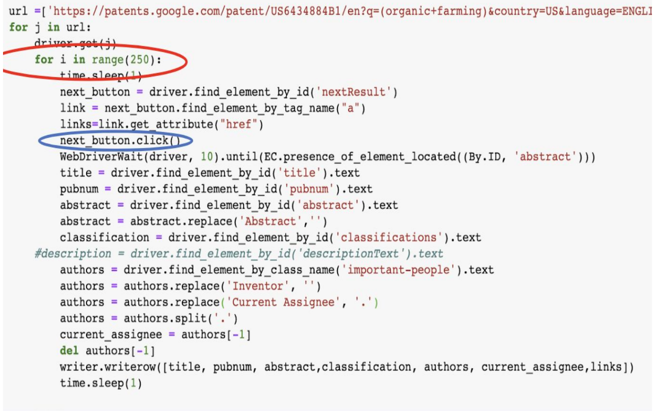
Inventor: Brian P. Hanley

Worldwide applications
2002 · US

Application US10/228,248 events

2002-08-26 • Application filed by Brian P. Hanley
2002-08-26 • Priority to US10/228,248
2003-12-30 • Application granted
2003-12-30 • Publication of US6671582B1
2019-08-05 • Application status is Expired - Fee Related

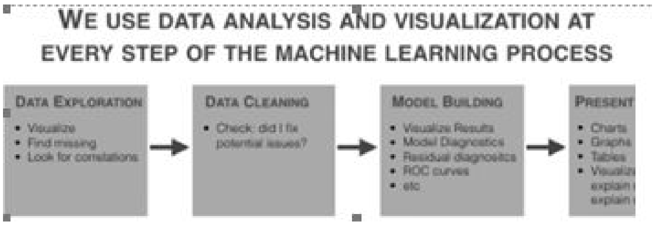
Web Scraper: Selenium web driver is used to collect data from site. The function of Selenium web scraper is to open url and scrape the data from specified html tags and write to the “output.csv” file .We collected multiple links each one regarding a specific domain from google patents website.. These multiple links are then stored in a list and iterated to runs a loop for more than 250 times for each link which extracts “title”, “pubnub”, “abstract”, ”classification”, ”important people”, ”inventor”, ”current assignee”.



```
url =['https://patents.google.com/patent/US6434884B1/en?q=(organic+farming)&country=US&language=ENGLI
for j in url:
    driver.get(j)
    for i in range(250):
        time.sleep(1)
        next_button = driver.find_element_by_id('nextResult')
        link = next_button.find_element_by_tag_name("a")
        links=link.get_attribute("href")
        next_button.click()
        WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.ID, 'abstract')))
        title = driver.find_element_by_id('title').text
        pubnum = driver.find_element_by_id('pubnum').text
        abstract = driver.find_element_by_id('abstract').text
        abstract = abstract.replace('Abstract','')
        classification = driver.find_element_by_id('classifications').text
    #description = driver.find_element_by_id('descriptionText').text
        authors = driver.find_element_by_class_name('important-people').text
        authors = authors.replace('Inventor', '')
        authors = authors.replace('Current Assignee', '.')
        authors = authors.split('.')
        current_assignee = authors[-1]
        del authors[-1]
        writer.writerow([title, pubnum, abstract,classification, authors, current_assignee,links])
        time.sleep(1)
```

Data Analysis: The data scientist spends most of their time doing analysis of data and prepare the data for modeling. The traditional statement is that data scientists *“spend 80% of their time on data preparation.”* While I think the statement is true, a more precise statement is that you spend more than 80% of the time on gathering the data, aggregating the data, cleaning the data and preprocessing the data. And ultimately, the importance of data analysis applies not only to data science generally, but machine learning specifically.

If you wish to build a machine learning model, you’ll spend huge amounts of time doing data analysis as a *precursor* to that process. Moreover, the data analysis will be used to explore *the results* of your model after you’ve applied a machine learning algorithm.



WE USE DATA ANALYSIS AND VISUALIZATION AT EVERY STEP OF THE MACHINE LEARNING PROCESS

DATA EXPLORATION
- Visualize
- Find missing
- Look for correlations

DATA CLEANING
- Check: did I fix potential issues?

MODEL BUILDING
- Visualize Results
- Model Diagnostics
- Residual diagnositcs
- ROC curves
- etc

PRESENT
- Charts
- Graphs
- Tables
- Visualiz
explain
explain

data was cleaned using Pandas and cleared the rows having missed or null values. Stop words like “is”, “am”, “in” were removed from title to the purpose of having only keywords with some meaning. Python library NLTK is used to remove the stop words. Duplicate patents were removed from the dataset based on patent id and kept the first row appeared in the dataset.

V. Methodology

The figure below (Fig. 1) describes the methodology that summarizes the users search records based on specific search conditions and when the target user searches patents, the methodology filters the results for the patent recommendation system. The following section describes the patent content based filtering process as the core of the dynamic patent recommendation methodology.

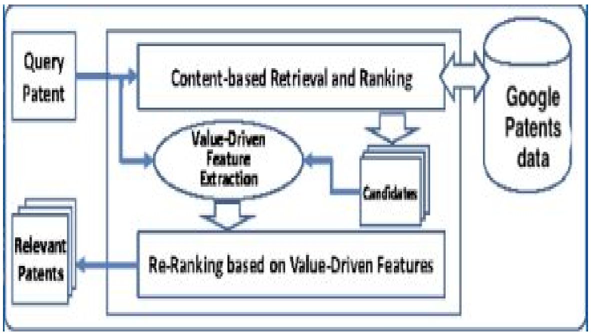
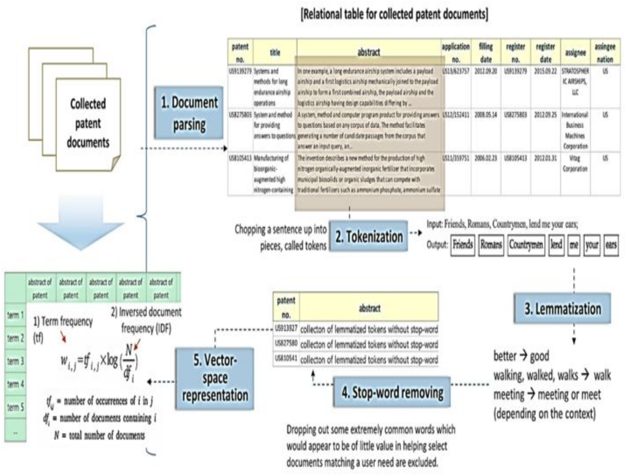

Query Patent
Content-based Retrieval and Ranking
Google Patents data
Value-Driven Feature Extraction
Candidates
Relevant Patents
Re-Ranking based on Value-Driven Features

Fig. 1. Research procedure

The Content based filtering which is based on keywords of the items matching with user keywords, which indicates the users interest. The system gives the list of the patents that are similar to the particular given patent filter. More specifically, the pairwise similarity score will be computed for all the patents based on the filter criteria and recommend top five patent list based on the similarity score.


[Relational table for collected patent documents]
Collected patent documents
1. Document parsing
patent no.
title
abstract
application no.
filling date
register no.
register date
assignee
assingee nation
2. Tokenization
Chopping a sentence up into pieces, called tokens
Input: Friends, Romans, Countrymen, lend me your ears;
Output: Friends Romans Countrymen lend me your ears
3. Lemmatization
better → good
walking, walked, walks → walk
meeting → meeting or meet (depending on the context)
4. Stop-word removing
Dropping out some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded.
collecton of lemmatized tokens without stop-word
5. Vector-space representation
abstract of patent
term 1
term 2
term 3
term 4
term 5
1) Term frequency (tf)
2) Inversed document frequency (IDF)
$w_{i,j}=tf_{i,j}\times\log(\frac{N}{df_i})$
tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Cosine similarity, Euclidean distance and Pearson correlation were initially used to calculate the similarity between different patents. However, all three similarity measures gave us the same result. Eventually, we proceeded with cosine similarity because even if the two similar items are far apart by the Euclidean distance chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity.

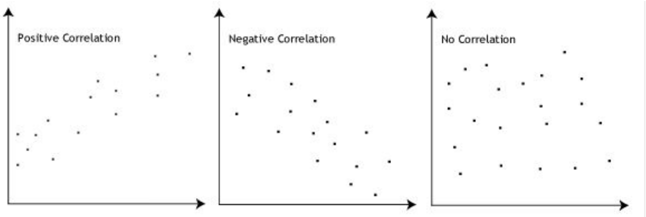
Euclidean distance: This similarity matrix calculates the linear distance between the two documents in vector space model. In other words, euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. Note that the formula treats the values of X and Y seriously: no adjustment is made for differences in scale. Euclidean distance is only appropriate for data measured on the same scale.

$$d(p_1, p_2) = \sqrt{\sum_{i \in \text{item}} (s_{p_1} - s_{p_2})^2}$$

Pearson Correlation: Pearson correlation coefficient, is a measure of the strength of a linear association between two variables. Basically, this draws correlation to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, indicates how far away all these data points are to this line of best fit.

$$\rho_{xy} = \frac{\frac{1}{n}\sum((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} * \sqrt{\frac{1}{n}\sum(y_i - \bar{y})^2}}$$

The Pearson correlation coefficient, ranges from the values from +1 to -1. Where the value 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Cosine Similarity: It determines the similarity between two vectors in vector space model. Mathematically it the the angle between two vectors in vector space. This method have been found to be the most time efficient method for similarity calculation.

$$cos(\theta) = \frac{\sum x_i * y_i}{\sqrt{\sum x_i^2} * \sqrt{\sum y_i^2}}$$

The cosine similarity matrix disregards the magnitude of the vectors, hence it eliminates the possibility of having higher similarity scores because of unequal length of documents.

A. Term Frequency-Inverse Document Frequency

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

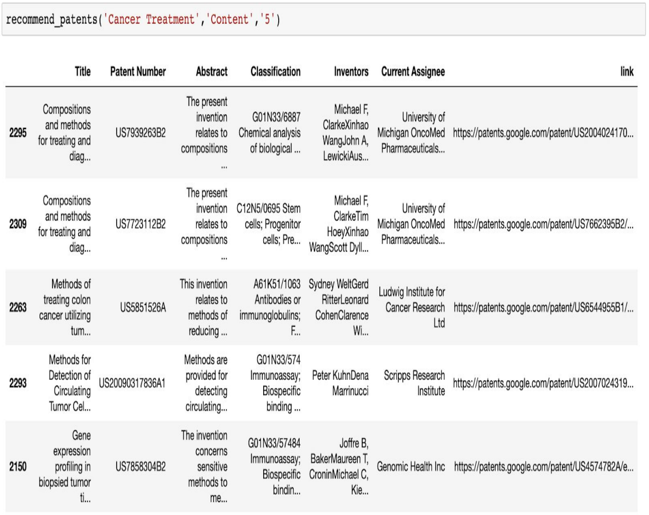
Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).



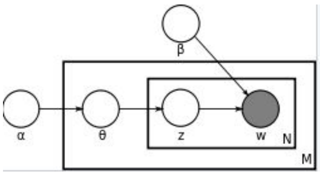
```
recommend_patents('Cancer Treatment','Content','5')
```

	Title	Patent Number	Abstract	Classification	Inventors	Current Assignee	link
2295	Compositions and methods for treating and diag...	US7939263B2	The present invention relates to compositions ...	G01N33/6887 Chemical analysis of biological ...	Michael F, ClarkeXinhao WangJohn A, LewickiAus...	University of Michigan OncoMed Pharmaceuticals...	https://patents.google.com/patent/US2004024170...
2309	Compositions and methods for treating and diag...	US7723112B2	The present invention relates to compositions ...	C12N5/0695 Stem cells; Progenitor cells; Pre...	Michael F, ClarkeTim HoeyXinhao WangScott Dyll...	University of Michigan OncoMed Pharmaceuticals...	https://patents.google.com/patent/US7662395B2/...
2263	Methods of treating colon cancer utilizing tum...	US5851526A	This invention relates to methods of reducing ...	A61K51/1063 Antibodies or immunoglobulins; F...	Sydney WeltGerd RitterLeonard CohenClarence Wi...	Ludwig Institute for Cancer Research Ltd	https://patents.google.com/patent/US6544955B1/...
2293	Methods for Detection of Circulating Tumor Cel...	US20090317836A1	Methods are provided for detecting circulating...	G01N33/574 Immunoassay; Biospecific binding ...	Peter KuhnDena Marrinucci	Scripps Research Institute	https://patents.google.com/patent/US2007024319...
2150	Gene expression profiling in biopsied tumor ti...	US7858304B2	The invention concerns sensitive methods to me...	G01N33/57484 Immunoassay; Biospecific bindin...	Joffre B, BakerMaureen T, CroninMichael C, Kie...	Genomic Health Inc	https://patents.google.com/patent/US4574782A/e...

B. *Latent Dirichlet Allocation*

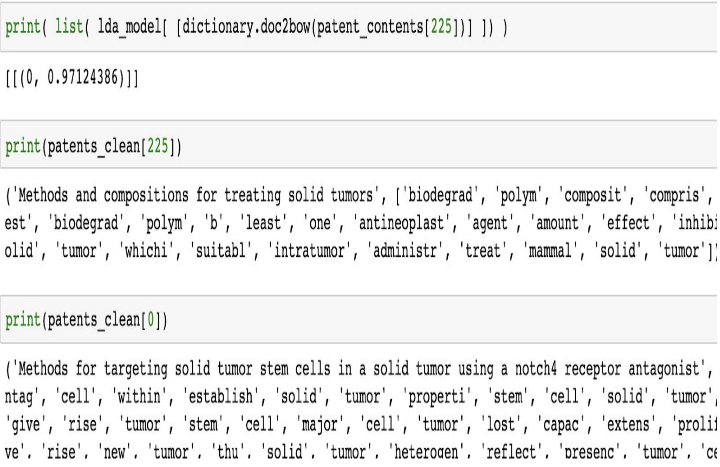
In LDA, each document is seen as a combination of other various keywords where each document is considered to have a set of keywords that are assigned to it by LDA. It is very much similar to the probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a sparse Dirichlet prior. The function of sparse Dirichlet priors is to encode the intuition that document cover only a small set of topics and that topics use only a small set of words frequently. In practice, this results in a better disambiguation of words and a more precise assignment of documents to topics. LDA is a generalization of the pLSA model, which is equivalent to LDA under a uniform Dirichlet prior distribution. With plate notation, which is often used to represent probabilistic graphical models (PGMs), the dependencies among the many variables can be captured concisely. The boxes are "plates" representing replicates, which are repeated entities. The outer plate represents documents, while the inner plate represents the repeated word positions in a given document, each of which position is associated with a choice of topic and word. M denotes the number of documents, N the number of words in a document. The variable names are defined as follows:

- α is the parameter of the Dirichlet prior on the per-document topic distribution.
- β is the parameter of the Dirichlet prior on the per-topic word distribution
- is the topic distribution for document
- is the word distribution for topic k
- is the topic for the j-th word in document is the specified word



Researchers merged the two list filtered based on “title” and “abstract”. Then the data was cleaned by removing stopped words and duplicates. The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation. The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables θd are document-level variables, sampled once per document. Finally, the variables zdn and wdn are word-level variables and are sampled once for each word in each document.

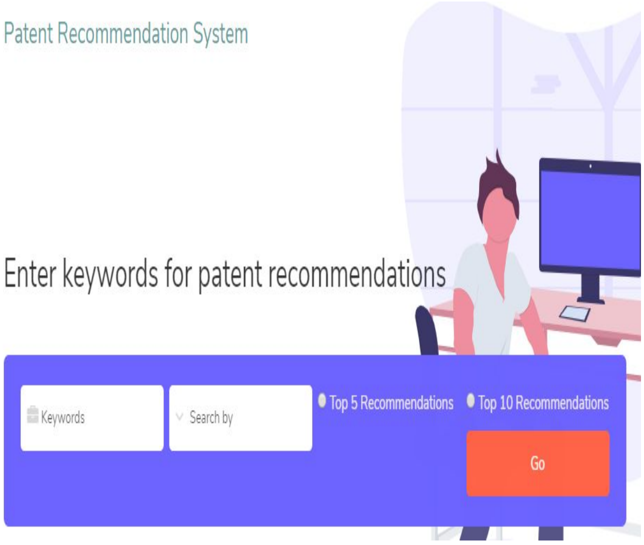
LDA will lead us to a list of ‘topics’, each consisting of multiple words. These words are what define the ‘topic’ - there is no explicit name or label for each topic. In the picture below, after training the LDA model we are trying to find the similarity of document at index 225 with the trained dataset. So, the LDA gives us an output recommending document with index zero and similarity value of 0.97. Further, the document title for document zero and document 225 are printed for easy human interpretation.



```
print( list( lda_model[ [dictionary.doc2bow(patent_contents[225])] ]) )

[[(0, 0.97124386)]]

print(patents_clean[225])

('Methods and compositions for treating solid tumors', ['biodegrad', 'polym', 'composit', 'compris',
est', 'biodegrad', 'polym', 'b', 'least', 'one', 'antineoplast', 'agent', 'amount', 'effect', 'inhibi
olid', 'tumor', 'whichi', 'suitabl', 'intratumor', 'administr', 'treat', 'mammal', 'solid', 'tumor'])

print(patents_clean[0])

('Methods for targeting solid tumor stem cells in a solid tumor using a notch4 receptor antagonist',
ntag', 'cell', 'within', 'establish', 'solid', 'tumor', 'properti', 'stem', 'cell', 'solid', 'tumor',
'give', 'rise', 'tumor', 'stem', 'cell', 'major', 'cell', 'tumor', 'lost', 'capac', 'extens', 'proli
ve'. 'rise'. 'new'. 'tumor'. 'thu'. 'solid'. 'tumor'. 'heterogen'. 'reflect'. 'presenc'. 'tumor'. 'ce
```

C. GUI:

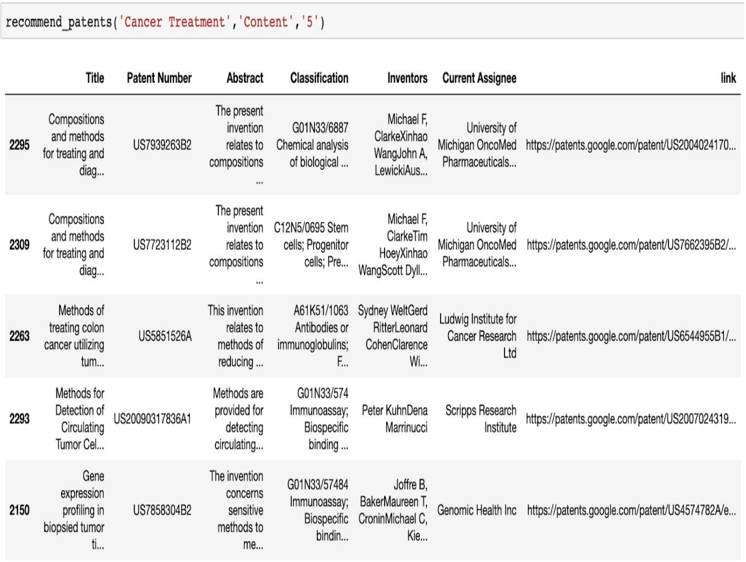
The GUI for our project lets user's search for a patent by entering keywords. The user can further refine his search by specifying the area of his search as content, author or assignee. The result gives related patent search information based on keywords searched by the user. One can visit the front-end UI of recommendation system by entering URL (http://www.arpsharma.com/)in world wide web. User must enter a keyword and select the search by option from list menu which has option of "Content", "Author", "Assignee" to filter the results. Along with that, user has option whether to get top five or top ten matching patents recommendation. We have used "Flask" library of python to connect the system with front end GUI, HTML5 and CSS3 templates are used to style the front end. There are four files named "main.py", "results.html", "profile.html", "recommendor.py" to design the front end. "main.py" is the main driver file to navigate to the URL. "output.csv " is used to feed the data into the system. Further, several functions are being called from" recommendor.py" to perform the actions and get the result. "profile.html" is used to display and style the home page of recommendation system in tags form and "result.html" is used to display the results styling. To give a bit of insight the picture below shows how the website for the project looks.


Patent Recommendation System
Enter keywords for patent recommendations
Keywords
Search by
○ Top 5 Recommendations
○ Top 10 Recommendations
Go

VI. Conclusion

This project for patent search is based on the analysis of user's behavior record which we get from user's keyword in patent search platform. The recommendation system extracts the most appropriate patents based on the TF-IDF and cosine similarity measures and helps users obtain related patents efficiently. The system requires little time to infer recommendations and the system works independently without using human brain or any mathematical formula. The system recommends the most relevant patents from its available database, which provides users more options to explore. Since the users' behavior records in form of input keyword is an input source, personal and other factors which may influence search the output are avoided here. As an endnote, it was very interesting to apply and compare different machine learning models and techniques like cosine similarity, Euclidean distance measure, Pearson correlation on the data set to calculate the similarity matrix and the filtering techniques like TF-IDF and LDA in our system.

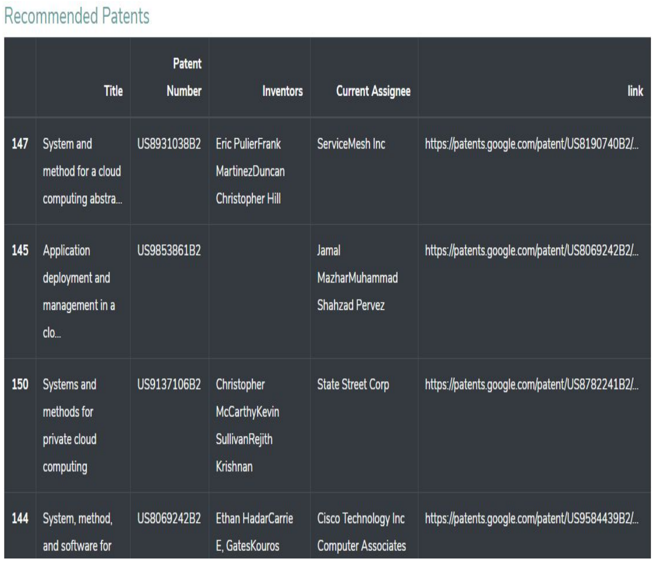
Research further can be extended beyond TF-IDF and LDA models. Models such as Rachhio's method, Patent Tree Model and Vector space model can be applied and results can be compared to discover the best model and to get better results.



```
recommend_patents('Cancer Treatment','Content','5')
```

	Title	Patent Number	Abstract	Classification	Inventors	Current Assignee	link
2295	Compositions and methods for treating and diag...	US7939263B2	The present invention relates to compositions ...	G01N33/6887 Chemical analysis of biological ...	Michael F, ClarkeXinhao WangJohn A, LewickiAus...	University of Michigan OncoMed Pharmaceuticals...	https://patents.google.com/patent/US2004024170...
2309	Compositions and methods for treating and diag...	US7723112B2	The present invention relates to compositions ...	C12N5/0695 Stem cells; Progenitor cells; Pre...	Michael F, ClarkeTim HoeyXinhao WangScott Dyll...	University of Michigan OncoMed Pharmaceuticals...	https://patents.google.com/patent/US7662395B2/...
2263	Methods of treating colon cancer utilizing tum...	US5851526A	This invention relates to methods of reducing ...	A61K51/1063 Antibodies or immunoglobulins; F...	Sydney WeltGerd RitterLeonard CohenClarence Wi...	Ludwig Institute for Cancer Research Ltd	https://patents.google.com/patent/US6544955B1/...
2293	Methods for Detection of Circulating Tumor Cel...	US20090317836A1	Methods are provided for detecting circulating...	G01N33/574 Immunoassay; Biospecific binding ...	Peter KuhnDena Marrinucci	Scripps Research Institute	https://patents.google.com/patent/US2007024319...
2150	Gene expression profiling in biopsied tumor ti...	US7858304B2	The invention concerns sensitive methods to me...	G01N33/57484 Immunoassay; Biospecific bindin...	Joffre B, BakerMaureen T, CroninMichael C, Kie...	Genomic Health Inc	https://patents.google.com/patent/US4574782A/e...

Figure.a Display of results at backend system



Recommended Patents

	Title	Patent Number	Inventors	Current Assignee	link
147	System and method for a cloud computing abstra...	US8931038B2	Eric PulierFrank MartinezDuncan Christopher Hill	ServiceMesh Inc	https://patents.google.com/patent/US8190740B2/...
145	Application deployment and management in a clo...	US9853861B2		Jamal MazharMuhammad Shahzad Pervez	https://patents.google.com/patent/US8069242B2/...
150	Systems and methods for private cloud computing	US9137106B2	Christopher McCarthyKevin SullivanRejith Krishnan	State Street Corp	https://patents.google.com/patent/US8782241B2/...
144	System, method, and software for	US8069242B2	Ethan HadarCarrie E, GatesKouros	Cisco Technology Inc Computer Associates	https://patents.google.com/patent/US9584439B2/...

Figure b Display of result at front end

VII. References

1. http://www.tfidf.com/

2. https://www.mdpi.com/2071-1050/10/8/2810

3.https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

4. https://kleiber.me/blog/2017/07/22/tutorial-lda-wikipedia/

5. https://radimrehurek.com/gensim/corpora/dictionary.html