Matching IP Buyers with Sellers: An Intellectual Property Recommendation System

Gautam Harinarayanan, Mayank Gulati, Akshay Kurapaty, Vaibhav Diyora, Matthew A. Lanham
Purdue University Krannert School of Management
gharinar@purdue.edu; mgulati@purdue.edu; akurapat@purdue.edu; vdiyora@purdue.edu; lanhamm@purdue.edu

Abstract

This study creates an intellectual property (IP) recommendation system that provides a list of firms, who are most likely to buy patents, to those that are trying to patent their IP. Our recommendation algorithm uses natural language processing techniques, which involves both syntactic and semantic analysis. The algorithm also utilizes other quantitative data such as the history of patents purchased by the target firm, firm's financial health, and other propriety features to increase match accuracy. To find the relevant patents, feature and key word extraction, stemming, POS tagging, lemmatization and chunking methods were used followed by TF-IDF vectorization and concept extraction to find the similarities between the patents. Each patent is given a similarity score on a scale of 0 to 1, and all the patents are ranked based on this score. Quantitative measures such as patent buying frequency, relevancy, etc. are used to re-rank the list. The ranked list acts as a recommendation list of firms with the greatest likelihood of buying at the top. We cross-validated our recommendation system using previous buyers and sellers, which gave highly promising results. Our partnering firm is considering using our recommendation as a new feature within their existing software.

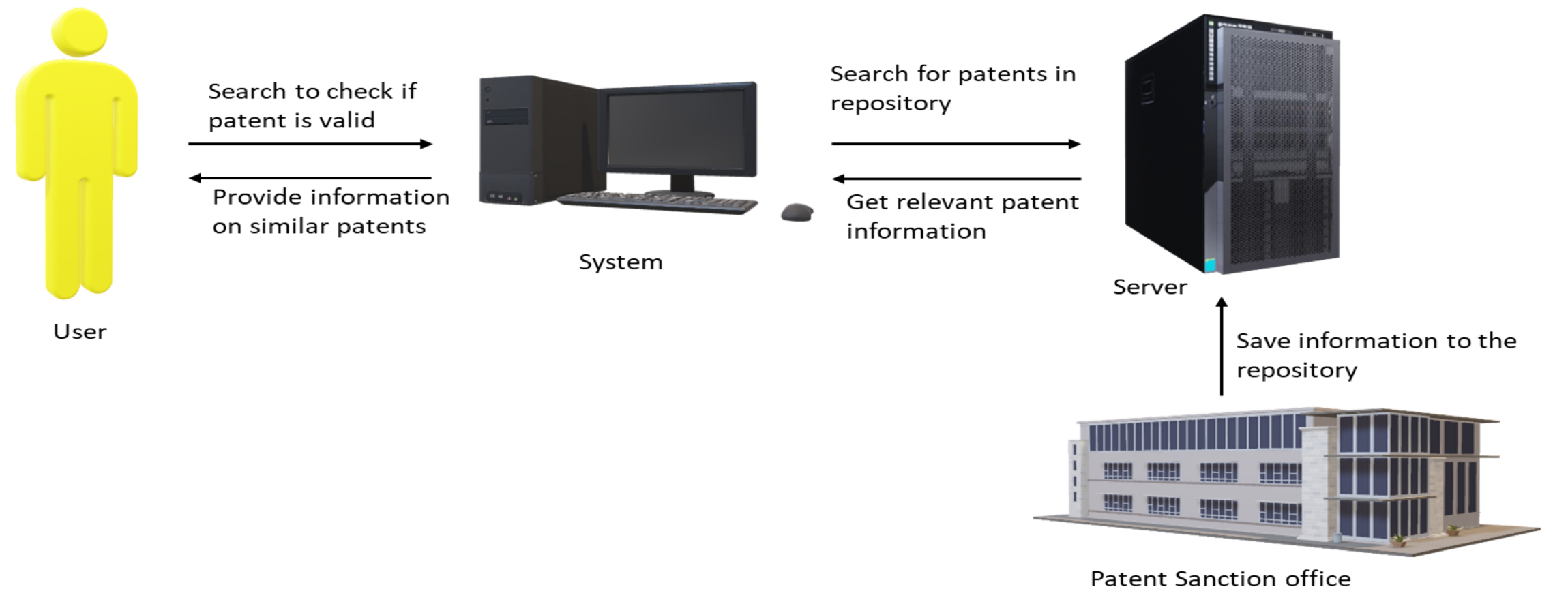
Introduction

The motivation for this study is that matching IP buyers with sellers is not a trivial task. For starters, the property can be a physical as well as a virtual asset. The government plays a major role in protecting both people's physical and virtual assets. While physical assets are protected by legal documents and trespassing rules, virtual assets such as an idea, is protected by issuing a patent to the person who owns the idea. Such patents are called intellectual property (IP). Exclusive access to develop and reap the benefits of patents create a competitive advantage for firms, allowing them to capture market share. Although IPs guarantee the right of business for that particular idea to the IP holder, there may arise situations where the IP holder does not want to hold the IP. In many cases, the IP holder would want to sell their IP to someone or some company for a reasonable price, which is usually the amount the idea protected by the IP is worth. These reasons could be:

- The company holding the IP is under financial stress and requires cash urgently.
- The tests to be conducted to validate the idea protected by the IP is expensive and not something an individual or a startup could afford.
- The company or individual holding the IP decided to pursue their business in another direction and would like to sell off their IP to some other individual or business working in that industry.

Selling the IP to make immediate money would surely help the companies who are in the above situations however, there is still the uncertainty of which company would want to buy this IP. In order to get a prospective buyer, the company would have to advertise their idea and search far and wide. Sometimes, the company who would be willing to buy this product may not even reside in the same country as the seller! This search process is not just time-consuming but also expensive, making it unaffordable to the IP holder.

An innovative patent search company, Loci.io, has found a solution to this problem. Loci.io provides their customers services which include finding if their client's idea is unique and if so, assists their clients in securing a patent for them. As an additional step, Loci.io wants to provide their clients with a set of companies that would be willing to buy this patent, should the need to sell arise. With this service, Loci.io would be able to increase its revenue by charging a percentage of the total sale amount if the sale of the patent goes through. It is this part of the project, finding potential buyers for a patent, that Loci.io has entrusted with us, and which will be explained in the later parts of this paper.


Search to check if patent is valid
Provide information on similar patents
User
System
Search for patents in repository
Get relevant patent information
Server
Save information to the repository
Patent Sanction office

Literature Review

The motive for the literature review was to check techniques which are tested and proven to find keywords, and for document matching. Deep learning and Feed Forward Neural Techniques have been proven as the best techniques, however as our problem statement is to find documents similar to the patent abstract being searched, we need a technique which is less complex and finds similar documents, not exact searches.

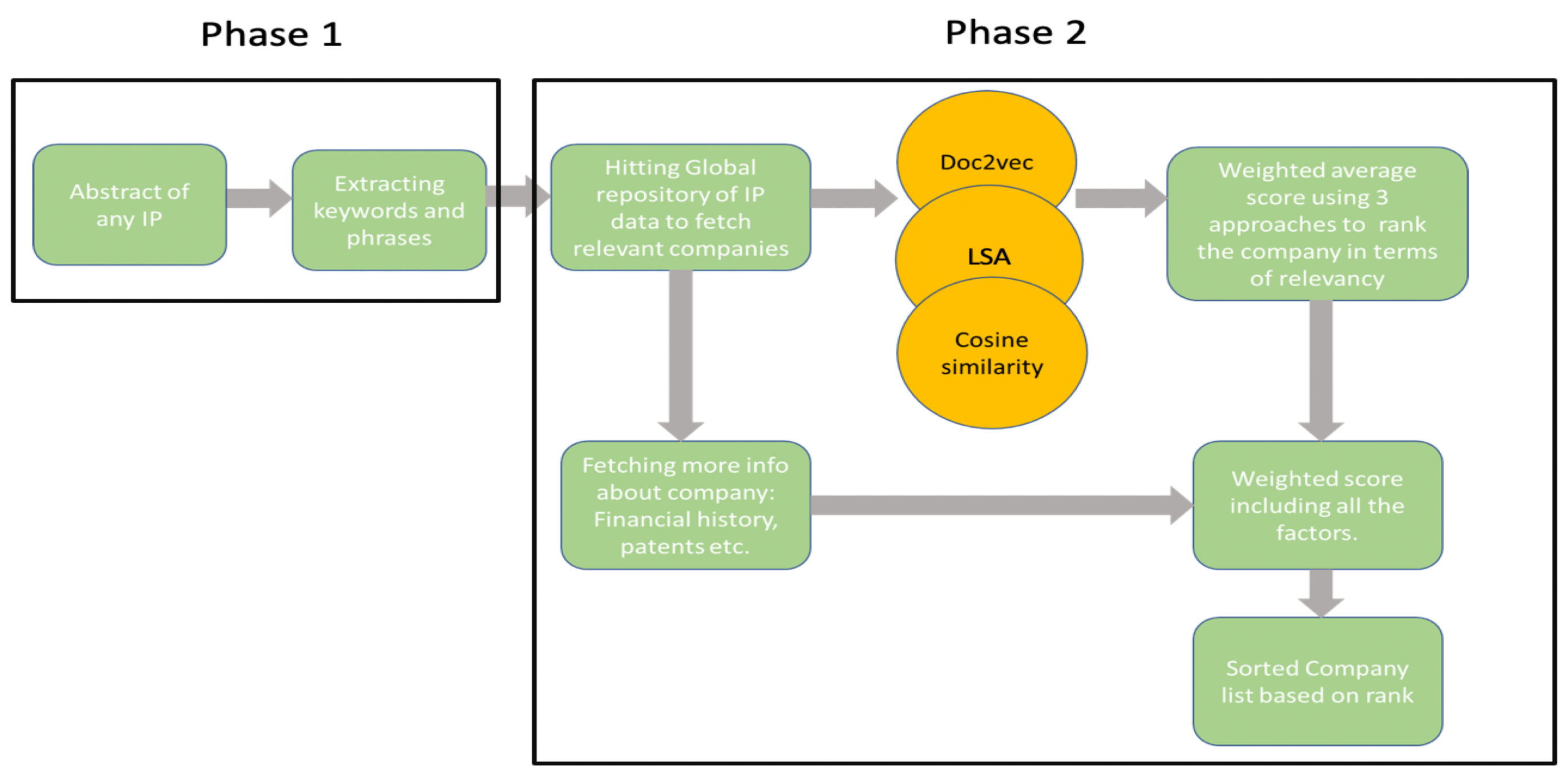
Document Similarity using Feed Forward Neural Networks:
Jackson and Leonard, 2015, used deep learning to compare documents to decide whether they were related to one another or not. They describe how a traditional feed forward network is the best option as neural networks are very good at distinguishing similar documents, and are able to learn the difference between related and unrelated ones [1].

Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality:
Reagan, Luke, Aaron and Jason, 2018, characterize a framework for matching text documents that decomposes existing methods into: (1) the choice of text representation, and (2) the choice of distance metric. The experimental results are then enhanced by developing a predictive model to estimate the match quality of pairs of text documents as a function of our various distance scores [2].

Methodology

This study was divided into two phases for faster completion. The first phase involved generating the keywords from a given patent abstract (instigating patent) to find all target patents which have those keywords. The second phase was to compare the instigating patent abstract with those of the target patent abstracts to find which one is most similar, proving that the target companies with most similar abstracts as that of the instigating patent would be working in the same field/area as the instigating company, and hence more likely to buy the patent from them.

Figure below shows the flowchart for both phases.


Phase 1
Phase 2
Abstract of any IP
Extracting keywords and phrases
Hitting Global repository of IP data to fetch relevant companies
Doc2vec
LSA
Cosine similarity
Weighted average score using 3 approaches to rank the company in terms of relevancy
Fetching more info about company: Financial history, patents etc.
Weighted score including all the factors.
Sorted Company list based on rank

Phase 1: Extracting Keywords

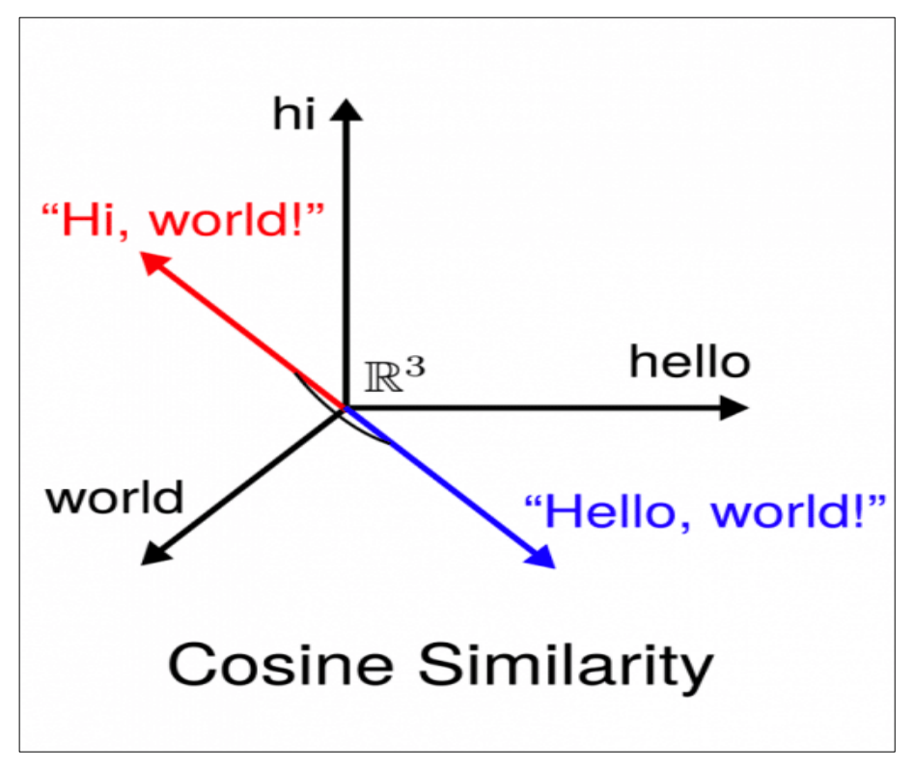
This process involves extracting keywords from the given abstract. These keywords are used to search for related patents in the patent repository. The following techniques are applied on the abstract text in a sequence, they are as follow: Sentence Tokenizer, POS Tagging, Chunking, Chinking and Stop Word Removal. All these techniques remove the irrelevant information from the text and exposes the potential keywords that capture the inner meaning of the text, which are used to hit the API, which provides us with the patents that have the relevant keywords in them.

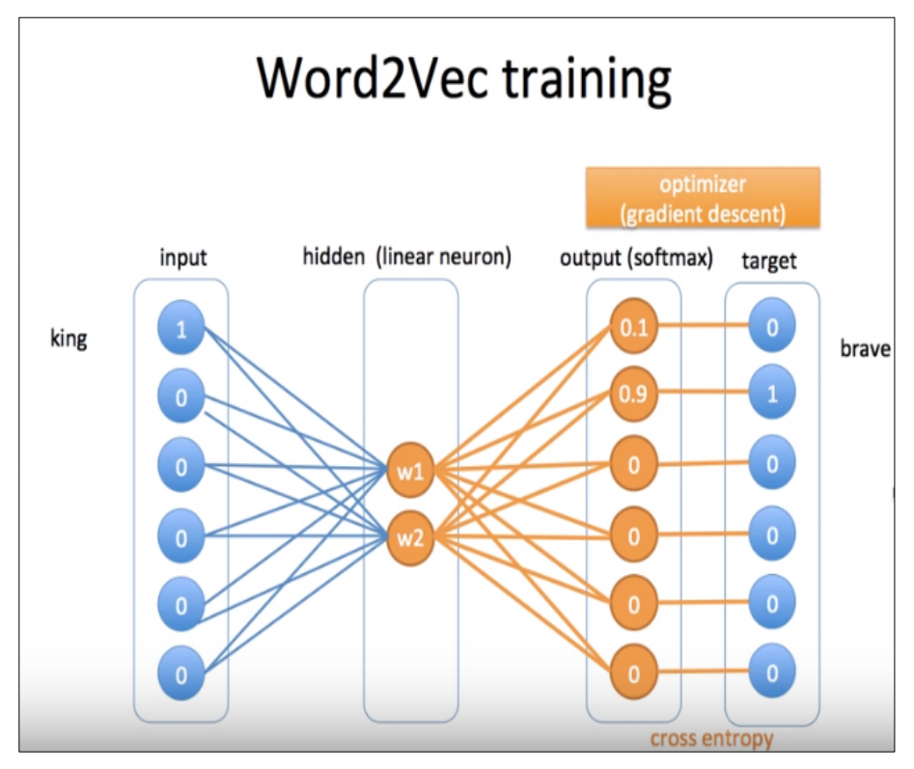
Phase 2: Document Similarity Matching

This phase deals with matching the given abstract with the abstract of patents present in the patent repository and ranking them according to their relevancy. Relevancy in this context deals not only on how strong the patents match with each other but also on the firm's previous patent purchase history and financial status. A probabilistic weighted score is then allocated to each measure which is the relevancy score. This score clearly defines how similar the target abstract is with the instigating abstract.

TFIDF – Cosine Similarity

To create this model, a term frequency matrix is created which is cross-checked with document frequency matrix, and a mathematical calculation is performed on these two matrixes to yield a TFIDF matrix, which gives a high score to those keywords which occur mostly in all documents. Cosine Distance, which is similar to the distance formula of Euclidean is used on this TFIDF – matrix, is performed to figure the similarity between two documents. This measure is dominated by the occurrences of keywords in the documents. Below is a diagram depicting the same [3]:





LSI – Latent Semantic Analysis

LSI model using singular value decomposition methodology on the keyword matrices is performed to come up with a metric that validates the semantic association between the words. This methodology is extended to documents to be used in this problem context. Semantic analysis between the words is generated, which is then used to aggregate at the document level and then generate a document level semantic score. This semantic score is used to find the relevancy in content with the other documents. This measure is dominated by the semantics of the document.

Word2Vec Model

This method uses a neural network to create a numerical array of representation for a given word. This neural network is designed in a way that after training the neural net on a large data set, the numerical representation of any given word is almost similar to a similar contextual word as can be seen in the figure above [4].

Methodology

Combining these three measures gives us relevant information pertaining to keywords, context and semantic. A weighted average scoring system is then used to determine the relevancy score.

However, patent relevancy score alone is not sufficient enough to match patent buyer with the company that is most likely to buy the patent. To make a sound recommendation, additional information such as the firm's frequency of patent purchase and other financial information is gathered as well.

All the above information, along with the patent similarity score, is used to determine the likelihood of a firm buying the patent. This likelihood score is used to sort the firms in descending order and provide a recommendation to the user.

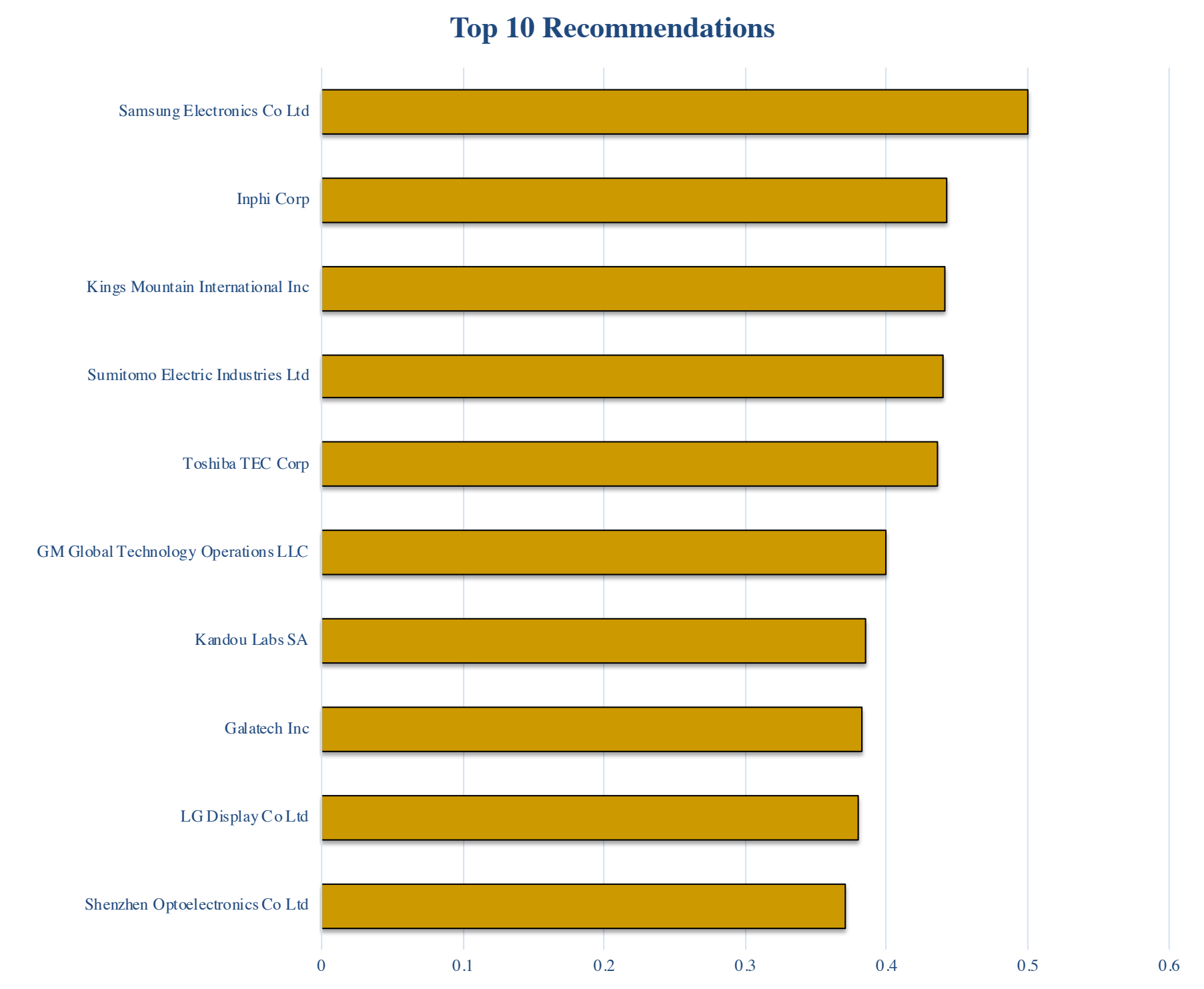
Results

The aim of our study was to match patents with a list of companies which may be interested in buying the patent off from the publisher's hand. This is essential for certain smaller companies to raise cash immediately without worrying about uncertainties in their future. Once the instigating company obtains a patent, our algorithm will be used to find target companies, based on an assigned score which is the probability of these companies holding similar patents. These target companies will then be listed in descending order and provided to the instigating company. The instigating company can then approach the target companies to strike a deal to sell their patents if need be.

Given below is an example of an instigating abstract, which when searched using our algorithm provides the results shown:

Instigating Abstract

Communication system, apparatus and method An example remote control device includes an inertial sensor; one or more manipulable input devices; wireless communication circuitry; and control circuitry for controlling the wireless communication circuitry to communicate information about the inertial sensor and the input devices to an electronic device. The information is communicated at one or more communication intervals using a data format which permits a value associated with the inertial sensor and sampled at a given sampling time to be communicated in at least first and second different communications.



Conclusion

The purpose of this study was to design a methodology wherein given a document, keywords could be extracted which can be used in a search engine to find similar results. Once the results are obtained, our algorithm can be used to test the relevancy and score the results to find the top ten documents which are an exact match. This algorithm has immense potential in any area which requires matching documents. One application of our algorithm would be to find prospective candidates for job by scanning resumes of the applicants and comparing it with the job requirements. Document matching can also be used in recommendation systems and in the academic space, to find if there exists papers to help you with what you're working on. In this study, we used our algorithm to find companies that would most likely buy a patent, based on their previous patents. With this algorithm in place, we expect Loci.io to make signifant increases in their revenue by taking a percentage for every patent sold.

Assumptions

- The list of target companies generated is heavily dependent on the text provided in the instigating company's patent. As the text in the abstract is what will be used to gather keywords and phrases to search for target companies and to match documents to obtain a score for the target companies, we assume that the abstract is worded in a way to mention all necessary keywords which would be required.
- We assume that only those companies who hold patents similar to the one being searched will be interested in purchasing the patent.

References

[1] https://cs224d.stanford.edu/reports/PoulosJackson.pdf
[2] https://arxiv.org/pdf/1801.00644.pdf
[3] https://www.machinelearningplus.com/nlp/cosine-similarity/
[4] https://www.maryville.edu/as/wp-content/uploads/sites/9/2019/01/Data-Science-Project-Sample-C-SENTIMENT-ANALYSIS.pdf