Textual Methods for quantifying patent importance

Gaurav Pathak, Bryan Kelly, Dragomir Radev

Yale University, New Haven, CT

LILY Lab

Introduction

Patents are an important store of modern inventions and innovations.From a ‘method of swinging on a swing’ to CRISPR, patents document a multitude of novel/important technologies and ideas.
In the modern economy, patents are more important as a way to protect one’s invention so that one can extract an economic rent from the usage of the content of the patent. The ‘patent stock’ of a company is an important asset and lawsuits’ settlements can go into the billions of dollars for patent infringement as awarded to companies like Apple and Polaroid.
Despite the apparent economic value of a patent, it is quite difficult to build a system to determine how valuable a certain patent is. The difficulty arises from the fact that the economic outcome of patents held, i.e. revenue growth, stock price etc. is noisy; a function of not only the underlying technology but also of marketing, the economy etc.
In this work I try to use textual methods to quantify how useful patents are. First, I see how well a Doc2Vec model captures patent information by building a patent acceptance classfier. Then I propose a dataset which can be thought of as a noiseless approximator for patent related company outcomes to perform experiments on

Testing/Result

The two different Doc2Vec models were tested for performance. The distributed memory doc2vec, one that does skip-gram word vector training appeared to perform better than the distributed bag of words model in a variety of training situations. (Different vector representation sizes, train-test split etc.)
Next, two different parts of the patents were tested to see how useful they are in terms of the accuracy in predicting whether a patent application was successful or not.
The first was the claims section of the document. In general it displayed a classification accuracy of around 75%.
Using the ‘Description’ section, however, significantly improved performance, with accuracy going to around 90%

	Class	precision	recall	f1-score	support
	Success	0.86	1	0.93	19
	Failure	1	0.75	0.86	12
micro avg		0.9	0.9	0.9	31
macro avg		0.93	0.88	0.89	31
weighted avg		0.92	0.9	0.9	31

Table 1: Prediction accuracy =90% for using patent Description along with a Distributed memory model of Doc2Vec

	Column2	precision	recall	f1-score	support
accuracy 0.7419354838709677					
	Success	0.72	0.95	0.82	19
	Failure	0.83	0.42	0.56	12
micro avg		0.74	0.74	0.74	31
macro avg		0.78	0.68	0.69	31
weighted avg		0.76	0.74	0.72	31

Table2: Prediction accuracy using only the claims information of the patent application

Materials and Methods

Pharmaceutical and Biotechnology companies rely to a great extent on patents for their survival and profitability. In order to look up the patents provided to a certain company, the United States Patent and Trademark Office allows a search by assignee name. Biotechnology companies patents and applications that did not end up becoming patents were scraped form the search website.
In order to determine whether textual methods can capture the information of a patent, I set up a task where a the objective to determine whether an application ended up becoming a patent or not. To do this, I used two variants of the Doc2Vec to build a feature vector. The feature vector was subsequently used to classify documents by using logistic regression
In terms of the documents themselves, I tested the system on only the ‘claims’ section and the ‘description’ section.

Conclusion/Next Steps

In this work there appears to be some validation that textual representations do have some predictive ability in the task set for classifying patent applications. More work needs to be done on how this may be validated to a larger corpus of companies and whether the results hold at that level.

With respect to the problem of quantifying the economic value of patents or patent features, the approach one can use is to use the funding success of early stage pharmaceutical/bio-tech companies as a function of the patent features. It can be a binary ‘raised next round’ variable or an indication of how it changed the valuation of the company. Trying to predict these with features such as those created by the Doc2Vec algorithm can be useful in creating a model that is able to demonstrate what the economic value of a patent was and what features contributed to it. (The features portion is not straightforward as it is quite difficult to interpret elements of a doc2vec vector)