Use Web Crawler & Doc2vector
to analysis Similarity Litigations

Bo-Hong Liu

Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan

Abstract. The general search engine is to search by keyword, although you can get a very large number of search results on the page, but still need to interpret the human to confirm whether the document content is required by the user. Legal staff in court proceedings will require the precedent as the basis for the relevant judgments, it is necessary to find similar cases to clarify the case of litigation, this study to crawl Web crawler search engine to find the relevant case articles, pre-treatment in the article Afterwards, the document value of Doc2vec is given, which can be used to make the similarity analysis of case based on the vector's directionality and vector matrix. It can provide relevant case documents of interest to the user and help the legal person to search the precedent of judgment Efficiency and accuracy.

Keywords*: Web crawler, Doc2vec, Litigation, Word2vec, Precedent*

1. Introduction

In today's era of online information development, the transmission of information is no longer based on paper records alone. It is based more on electronic records and on-line information, and even directly places information on the cloud, thus making information more transparent and readable. Government agencies and non-profit organizations are also happy to share information with people, including the court litigation online.

However, a large number of databases and sources is difficult to find the information really looking for, even using Google and other search engines still need to manually interpret, facing the massive information crawling is only the tip of the iceberg and found Of the information is true or difficult to determine, so in the text and other analysis will be inefficient problems. Therefore, in the face of a large amount of information, we can crawl the data by using the way of web crawler, and then use the doc2vec analysis method to find out the most similar articles to meet the needs of users, saving time and doing the right thing.

2. Literature review

2.1 Web Crawler

Web crawlers is an automated program that methodically scans through web pages to create an index of the data it is set to look for. This process is called Web crawling.

Web crawlers can copy all the pages they visit for later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently. Crawlers consume resources on the systems they visit and often visit sites without tacit approval. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For instance, including a robots.txt file can request bots to index only parts of a website, or nothing at all.

A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. If the crawler is performing archiving of websites it copies and saves the information as it goes. The archives are usually stored in such a way they can be viewed, read and navigated as they were on the live web. (As showing Figure 1)

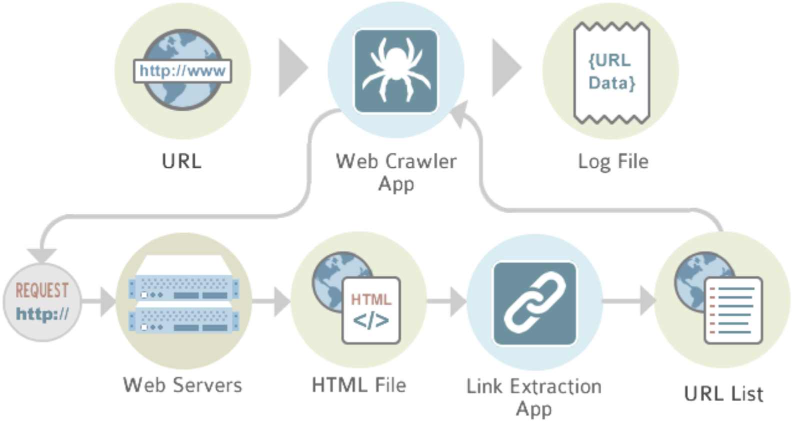


Figure 1

2.2 Natural Language Processing (NLP)

NLP is a method for computers to understand human language in a smart and useful way. By NLP, let the computer understand the meaning of a sentence or article, nouns, verbs, adjectives and adverbs are the basis of the literary meaning NLP from the text of the sentence will be non-structural features of the text presented as analytical data NLP is used to analyze text, allowing machines to understand how human’s speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more. NLP is commonly used for text mining, machine translation, and automated question answering.

NLP is characterized as a hard problem in computer science. Human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they’re linked together to create meaning. Despite language being one of the easiest things for humans to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master.

2.3 Word2vec

The word vector (also known as word embedding or representation) is a technique widely used in Natural Language Processing (NLP). Using a vector to represent each word, so that convert a sentence into a vector of words to represent it and numerically. This research use Word2vec which develop by Google to convert the words into vectors.

This tool provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These representations can be subsequently used in many natural language processing applications and for further research.

Word2vec uses distributed representations of text to capture similarities among concepts. Showing as Figure 2, it understands that Paris and France are related the same way Berlin and Germany are (capital and country), and not the same way Madrid and Italy are. This chart shows how well it can learn the concept of capital cities, just by reading lots of news articles -- with no human supervision:

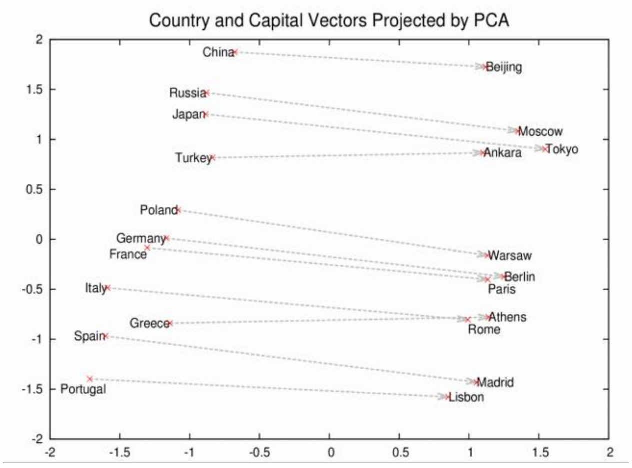


Figure 2

The model not only places similar countries next to each other, but also arranges their capital cities in parallel. The most interesting part is that we didn’t provide any supervised information before or during training. Many more patterns like this arise automatically in training.

2.4 Doc2vec

The approach for learning paragraph vectors is inspired by the methods for learning the word vectors proposed by Quoc Le and Tomas Milolov (2014). The inspiration is that the word vectors are asked to contribute to a prediction task about the next word in the sentence. So despite the fact that the word vectors are initialized randomly, they can eventually capture semantics as an indirect result of the prediction task. Users use this idea in paragraph vectors in a similar manner. The paragraph vectors are also asked to contribute to the prediction task of the next word given many contexts sampled from the paragraph. The sentence vector is unique by combining word vectors with shareable word vectors.Doc2vec attempts to predict the probability of words in the context of sentence and sentence vectors: Learning training is done via a set of documents or sentences, given the vector so that the probability of the next word appearing in the sentence vector is predicted.

In calculation, the two will be accumulated or connected, as the output layer. During the training of a sentence or paragraph, the paragraph id remains the same and contains the same paragraph vector, which is equivalent to using the semantic meaning of the sentence each time the word is predicted
In the predictive phase, through a fixed word vector to infer that a paragraph id is assigned to the sentence to be predicted by the sentence vector, the vector parameters of the word vector in the training phase are invariant, and the sentence to be predicted is trained by using the gradient descent. Finally, convergence can get the sentence to be predicted paragraph vector.

To achieve doc2vec there are two ways; namely, Distributed Bag of Words (DBOW) and Distribution Memory (DM), as shown in Figure 3 below and Figure 4, in a shorter sentence, you do not need a given word vector, the sentence vector Be trained to predict words.

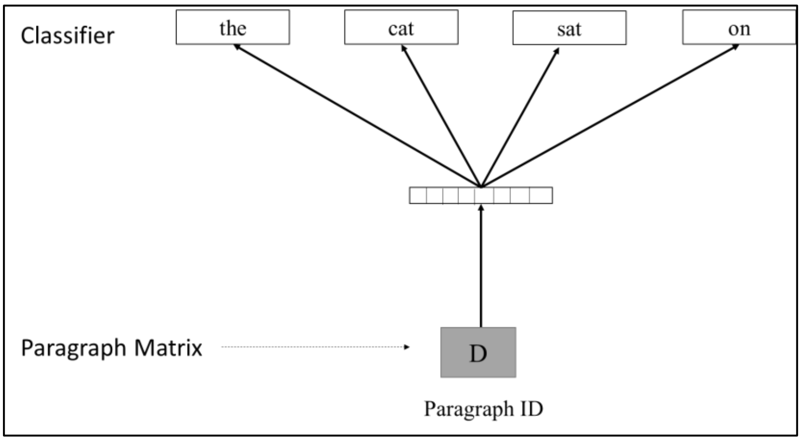


Figure 3 Distributed Bag of Words (DBOW)Model

The second is to learn a textual vector schema that predicts the fourth word ("on") in terms of three words ("the", "cat", "sat"), and the words that are typed are projected matrix w to predict words

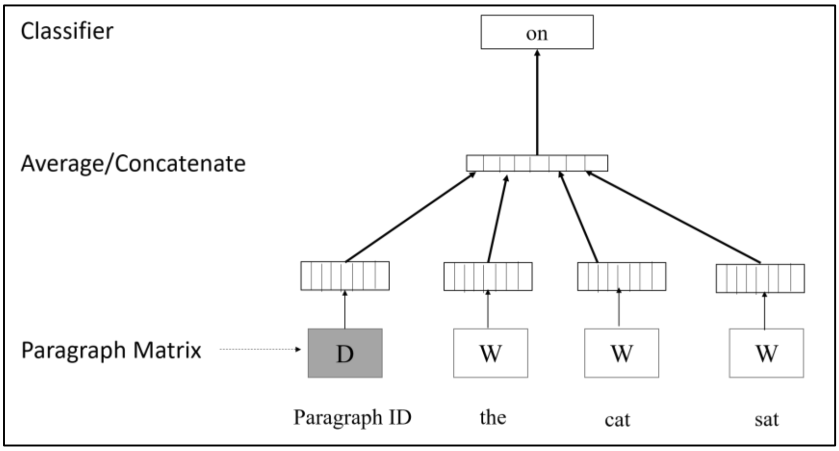


Figure 4 Distributed Memory(DM) Model

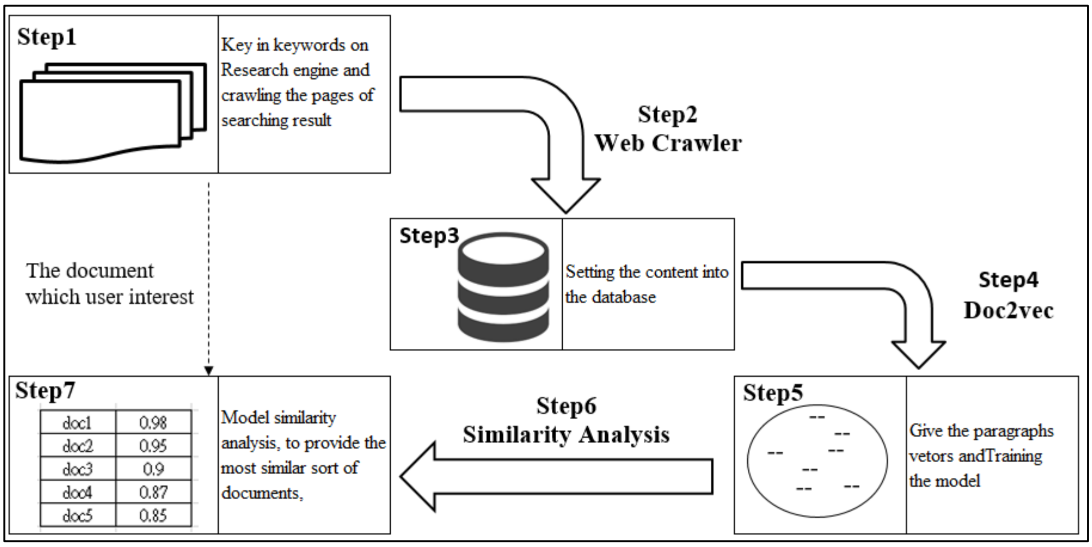
2.5 Programing Language-Python

Python has become a popular programming language in recent years. It is a powerful, fast, easy-to-read programming language. This report will rely heavily on python for web crawlers and dec2vec tools.

3. Methodology

It is difficult to parse a large number of documents on these networks. Therefore, it must be done in stages.

Showing as Figure 5, first, web crawlers should be implemented on the website to crawler pages, and the simulation package can be used to automatically crawler and put the contents of the documents into the database. After the establishment of the database, it can set the base as the dictionary. And then through word processing to eliminate redundant words and unnecessary noise characters, through Doc2vec to create a training module, you can give each document vector, and you can use the module for a critical comparison that Document the relationship.


Step1
Key in keywords on Research engine and crawling the pages of searching result
Step2
Web Crawler
Step3
Setting the content into the database
Step4
Doc2vec
The document which user interest
Step5
Give the paragraphs vetors andTraining the model
Step6
Similarity Analysis
Step7
Model similarity analysis, to provide the most similar sort of documents,

doc1	0.98
doc2	0.95
doc3	0.9
doc4	0.87
doc5	0.85

Figure 5

By using this method instead of using manual identification, not only increase the horizon and depth of the data, but also save a lot of time. For professional analysts, they can view documents more systematically.

4. Case Study

4.1Backround (As-is)

When handling legal proceedings, general law and order officials will make similar decisions in the light of the so-called precedent precedents and the application of previous norms in the light of similar precedent and circumstances as evidence in the face of similar legal cases or facts. Following the precedent as the basis of the precedents, therefore, it is necessary to find similar precedents as an important basis for legal practitioners in hearing or assisting litigation to clarify the offensive and defensive.

Through the court litigations online, official and unofficial network provide search engine for people and law politicians more conveniently search the relevant judgments of the document. Although network offer the massive data, tradition manual way to find relevant case is difficult and waste time.(Showing the Figure 6) Most of the general search engine for keyword research, but do not have the actual semantic analysis. Users still have to find the page from the keyword in accordance with the title to read, and then enter the page to confirm whether the content is consistent, it will take a lot of time to do artificial confirmation. Search by one page can also cause the following page can’t be seen. Keyword search is word search, but there is no way to ensure the importance and understand the relationship between document and document.

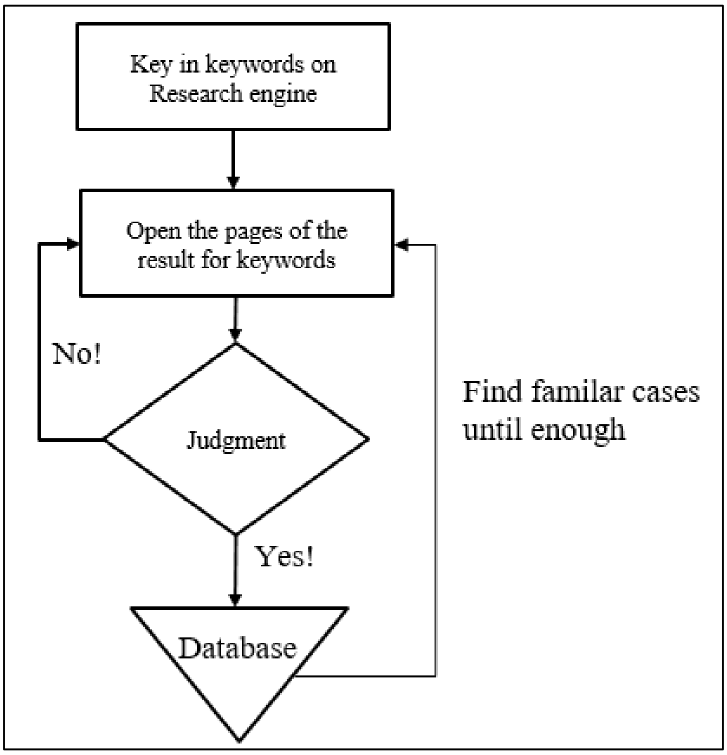

Key in keywords on Research engine
Open the pages of the result for keywords
No!
Judgment
Yes!
Database
Find familar cases until enough

Figure 6

There are some major reasons that searching problems.

1. Man-made judgments time-consuming and laborious

2. Limited to man-made interpretation of the restrictions that only can find a smaller scope

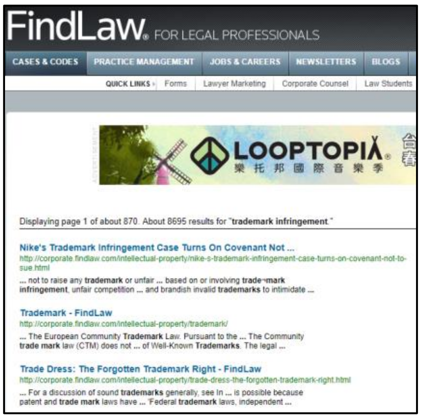
3. To judge the searched cases of relevance need to try &error again and again until find the similar ones.

4.2 To-be model

The paper provide the semantic analysis to figure out the similarity of the specific litigations. By programming language, Python, to do web crawler and setting database, then convert the documents into paragraph vectors by doc2vec, after that training the model which can comparison the similarity

Step1: User typing the keywords which interest in the litigation search engine

(As showing Figure 7) (Web provide: Find law http://lp.findlaw.com/)



FindLaw® FOR LEGAL PROFESSIONALS

Displaying page 1 of about 870. About 8695 results for "**trademark infringement**."

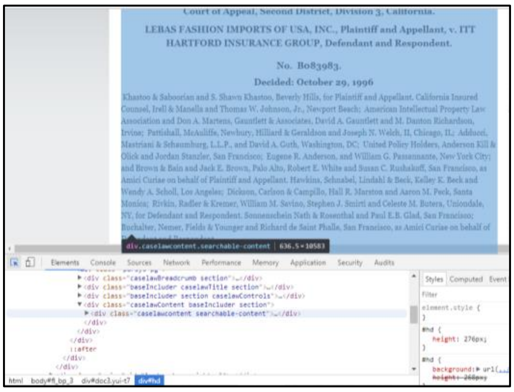
Nike's Trademark Infringement Case Turns On Covenant Not ...
http://corporate.findlaw.com/intellectual-property/nike-s-trademark-infringement-case-turns-on-covenant-not-to-sue.html
... not to raise any **trademark** or unfair ... based on or involving **trade-mark infringement**, unfair competition ... and brandish invalid **trademarks** to intimidate ...

Trademark - FindLaw
http://corporate.findlaw.com/intellectual-property/trademark/
... The European Community **Trademark** Law. Pursuant to the ... The Community **trade mark** law (CTM) does not ... of Well-Known **Trademarks**. The legal ...

Trade Dress: The Forgotten Trademark Right - FindLaw
http://corporate.findlaw.com/intellectual-property/trade-dress-the-forgotten-trademark-right.html
... For a discussion of sound **trademarks** generally, see In ... is possible because patent and **trade mark** laws have ... 'Federal **trademark** laws, independent ...

Figure 7

Step2: Use Python to crawl all the text of the litigation to climb out into database. Obtain network source code, are generally html format files, as long as the study of html tag (tag) structure, and then parse, you can get the required information. With the package "BeautifulSoup" which base on the properties of html attributes and retrieve the text to find the web page source code. (Showing Figure 8)



Court of Appeal, Second District, Division 3, California.

LEBAS FASHION IMPORTS OF USA, INC., Plaintiff and Appellant, v. ITT HARTFORD INSURANCE GROUP, Defendant and Respondent.

No. B083983.

Decided: October 29, 1996

Khastoo & Saboorian and S. Shawn Khastoo, Beverly Hills, for Plaintiff and Appellant. California Insured Counsel, Irell & Manella and Thomas W. Johnson, Jr., Newport Beach; American Intellectual Property Law Association and Don A. Martens, Gauntlett & Associates, David A. Gauntlett and M. Danton Richardson, Irvine; Pattishall, McAuliffe, Newbury, Hilliard & Geraldson and Joseph N. Welch, II, Chicago, IL; Adducci, Mastriani & Schaumburg, L.L.P., and David A. Guth, Washington, DC; United Policy Holders, Anderson Kill & Olick and Jordan Stanzler, San Francisco; Eugene R. Anderson, and William G. Passannante, New York City; and Brown & Bain and Jack E. Brown, Palo Alto, Robert E. White and Susan C. Rushakoff, San Francisco, as Amici Curiae on behalf of Plaintiff and Appellant. Hawkins, Schnabel, Lindahl & Beck, Kelley K. Beck and Wendy A. Scholl, Los Angeles; Dickson, Carlson & Campillo, Hall R. Marston and Aaron M. Peck, Santa Monica; Rivkin, Radler & Kremer, William M. Savino, Stephen J. Smirti and Celeste M. Butera, Uniondale, NY, for Defendant and Respondent. Sonnenschein Nath & Rosenthal and Paul E.B. Glad, San Francisco; Buchalter, Nemer, Fields & Younger and Richard de Saint Phalle, San Francisco, as Amici Curiae on behalf of

div.caselawcontent.searchable-content | 636.5×10583

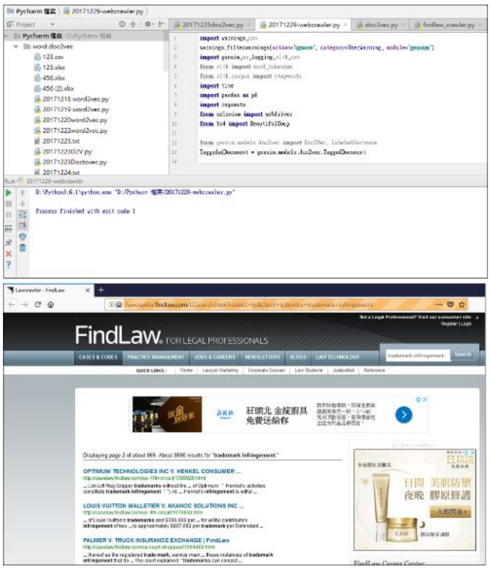
```
    ▶<div class="caselawBreadcrumb section">…</div>
    ▶<div class="baseIncluder caselawTitle section">…</div>
    ▶<div class="baseIncluder section caselawControls">…</div>
    ▼<div class="caselawContent baseIncluder section">
      ▶<div class="caselawcontent searchable-content">…</div>
     </div>
    </div>
   </div>
   ::after
  </div>
 </div>
```

```
element.style {
}
#hd {
  height: 276px;
}
#hd {
  background:▶ url(…
  height: 268px;
```

html body#fl_bp_3 div#doc3.yui-t7 div#hd

Figure 8

Through the package "selenium", it can simulate the user's behavior to automatic web crawler and switch pages. (As showing Figure 9)



```
import warnings,csv
warnings.filterwarnings(action='ignore', category=UserWarning, module='gensim')
import gensim,os,logging,nltk,csv
from nltk import word_tokenize
from nltk.corpus import stopwords
import time
import pandas as pd
import requests
from selenium import webdriver
from bs4 import BeautifulSoup

from gensim.models.doc2vec import Doc2Vec, LabeledSentence
TaggedDocument = gensim.models.doc2vec.TaggedDocument
```

```
D:\Python3.6.1\python.exe "D:/Pycharm 檔案/20171229-webcrawler.py"

Process finished with exit code 1
```

Displaying page 2 of about 869. About 8690 results for "trademark infringement."

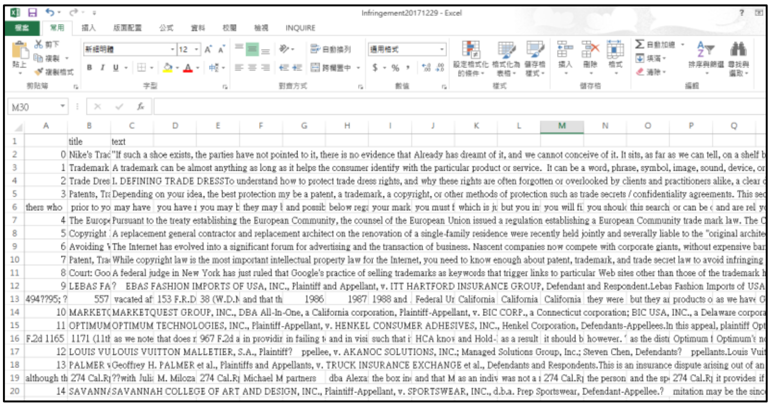
OPTIMUM TECHNOLOGIES INC V. HENKEL CONSUMER ...
http://caselaw.findlaw.com/us-11th-circuit/1200928.html
... Lok-Lift Rug Gripper trademarks without the ... of Optimum. " Henkel's activities constitute trademark infringement " "). Id. ... Henkel's infringement is wilful ...

LOUIS VUITTON MALLETIER V. AKANOC SOLUTIONS INC ...
http://caselaw.findlaw.com/us-9th-circuit/1578893.html
... of Louis Vuitton's trademarks and $300,000 per ... for willful contributory infringement of two ... to approximately $807,692 per trademark per Defendant ...

PALMER V. TRUCK INSURANCE EXCHANGE | FindLaw
http://caselaw.findlaw.com/ca-court-of-appeal/1064463.html
... thereof as the registered trade mark, service mark ... those instances of trademark infringement that do ... The court explained: "trademarks can consist ...

Figure 9

Step3: Capture the text into the database: Put the data into the database, and turning the data frame as Csv type. (As showing Figure 10) With the dataframe, it is more easily to manage the whole data to analysis in next step.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		title	text														
2	0	Nike's Trac	"If such a shoe exists, the parties have not pointed to it, there is no evidence that Already has dreamt of it, and we cannot conceive of it. It sits, as far as we can tell, on a shelf b														
3	1	Trademark	A trademark can be almost anything as long as it helps the consumer identify with the particular product or service. It can be a word, phrase, symbol, image, sound, device, or														
4	2	Trade Dres	I. DEFINING TRADE DRESSTo understand how to protect trade dress rights, and why these rights are often forgotten or overlooked by clients and practitioners alike, a clear c														
5	3	Patents, Tra	Depending on your idea, the best protection my be a patent, a trademark, a copyright, or other methods of protection such as trade secrets / confidentiality agreements. This sec														
6	thers who	prior to yo	may have	you have r	you may b	they may l	and possib	below rega	your mark	you must f	which is ju	but you in	you will fi	you shoulc	this search	or can be c	and are rel
7	4	The Europe	Pursuant to the treaty establishing the European Community, the counsel of the European Union issued a regulation establishing a European Community trade mark law. The C														
8	5	Copyright	A replacement general contractor and replacement architect on the renovation of a single-family residence were recently held jointly and severally liable to the "original archite														
9	6	Avoiding \	The Internet has evolved into a significant forum for advertising and the transaction of business. Nascent companies now compete with corporate giants, without expensive bar														
10	7	Patent, Tra	While copyright law is the most important intellectual property law for the Internet, you need to know enough about patent, trademark, and trade secret law to avoid infringing														
11	8	Court: Goo	A federal judge in New York has just ruled that Google's practice of selling trademarks as keywords that trigger links to particular Web sites other than those of the trademark h														
12	9	LEBAS FA	? EBAS FASHION IMPORTS OF USA, INC., Plaintiff and Appellant, v. ITT HARTFORD INSURANCE GROUP, Defendant and Respondent.Lebas Fashion Imports of USA														
13	494?795; ?	557	vacated af	153 F.R.D	38 (W.D.N	and that th	1986	1987	1988 and	Federal Ur	California	California	California	they were	but they ar	products o	as we have
14	10	MARKETQ	MARKETQUEST GROUP, INC., DBA All-In-One, a California corporation, Plaintiff-Appellant, v. BIC CORP., a Connecticut corporation; BIC USA, INC., a Delaware corpora														
15	11	OPTIMUM	OPTIMUM TECHNOLOGIES, INC., Plaintiff-Appellant, v. HENKEL CONSUMER ADHESIVES, INC., Henkel Corporation, Defendants-Appellees.In this appeal, plaintiff Opt														
16	F.2d 1165	1171 (11th	as we note	that does n	967 F.2d a	in providin	in failing t	and in visi	such that i	HCA know	and Hold-	as a result	it should b	however. '	as the distr	Optimum t	Optimum's
17	12	LOUIS VU	LOUIS VUITTON MALLETIER, S.A., Plaintiff? ppellee, v. AKANOC SOLUTIONS, INC.; Managed Solutions Group, Inc.; Steven Chen, Defendants? ppellants.Louis Vuit														
18	13	PALMER v	Geoffrey H. PALMER et al., Plaintiffs and Appellants, v. TRUCK INSURANCE EXCHANGE et al., Defendants and Respondents.This is an insurance dispute arising out of an														
19	although th	274 Cal.Rp	??with Julia	M. Miloza	274 Cal.Rp	Michael M	partners	dba Alexa	the box inc	and that M	as an indiv	was not a r	274 Cal.Rp	the person	and the sp	274 Cal.Rp	it provides
20	14	SAVANNA	SAVANNAH COLLEGE OF ART AND DESIGN, INC., Plaintiff-Appellant, v. SPORTSWEAR, INC., d.b.a. Prep Sportswear, Defendant-Appellee.? mitation may be the sinc														

Figure 10

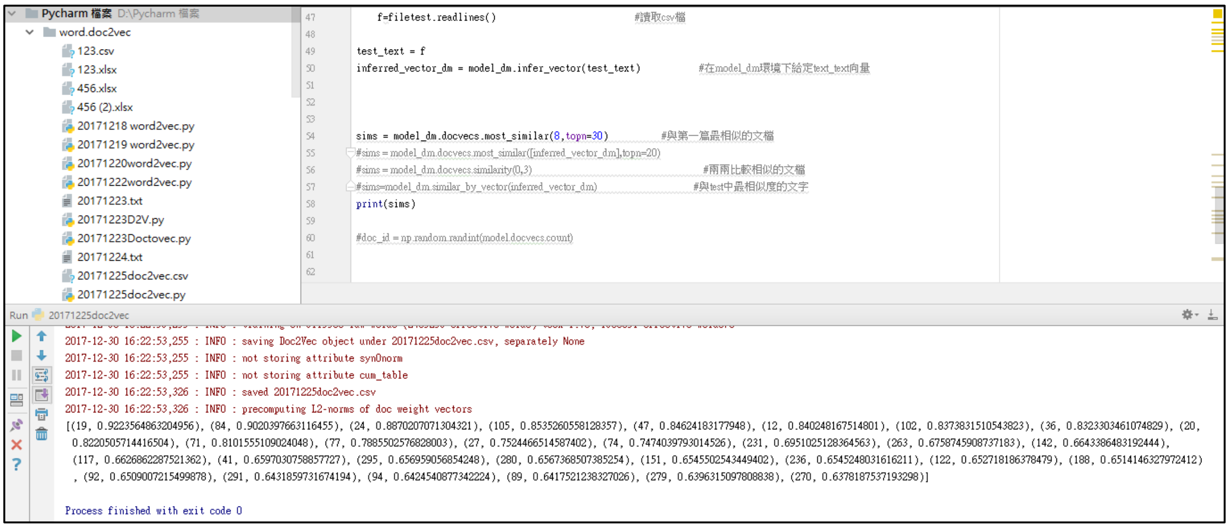
Step4: The use of Python for semantic, word meaning analysis, the establishment of the database model. Before that, it is important that doing text pre-processing which include wordiness, punctuation mark and specific words or signal. After that, setting each document a unique tagged id, and training as a model.

Step5: According to user’s needs to find similar documents

Users can interpret their own to find out the case of interest to the module, and do similarity analysis of the whole module. Finally, module provides the highest degree of similarity to the user.

In this case, assume that users are interested in case No. 9 and want to find out the documents related to them, and then use similar analysis to find similar documents. While the similarity is compared with the vector angle. The equal $\cos\theta$ is to 1, the higher the similarity is, so the relationship can be obtained according to the arrangement. If there is a low similarity among this document and other documents, the relevance of the document is low or can be changed Search again for new results.

(As Showing Figure 11)



```
    f=filetest.readlines()                              #讀取csv檔

test_text = f
inferred_vector_dm = model_dm.infer_vector(test_text)          #在model_dm環境下給定text_text向量



sims = model_dm.docvecs.most_similar(8,topn=30)          #與第一篇最相似的文檔
#sims = model_dm.docvecs.most_similar([inferred_vector_dm],topn=20)
#sims = model_dm.docvecs.similarity(0,3)                          #兩兩比較相似的文檔
#sims=model_dm.similar_by_vector(inferred_vector_dm)              #與test中最相似度的文字
print(sims)

#doc_id = np.random.randint(model.docvecs.count)
```

```
2017-12-30 16:22:53,255 : INFO : saving Doc2Vec object under 20171225doc2vec.csv, separately None
2017-12-30 16:22:53,255 : INFO : not storing attribute syn0norm
2017-12-30 16:22:53,255 : INFO : not storing attribute cum_table
2017-12-30 16:22:53,326 : INFO : saved 20171225doc2vec.csv
2017-12-30 16:22:53,326 : INFO : precomputing L2-norms of doc weight vectors
[(19, 0.9223564863204956), (84, 0.9020397663116455), (24, 0.8870207071304321), (105, 0.8535260558128357), (47, 0.84624183177948), (12, 0.840248167514801), (102, 0.8373831510543823), (36, 0.8323303461074829), (20,
 0.8220505714416504), (71, 0.8101555109024048), (77, 0.7885502576828003), (27, 0.7524466514587402), (74, 0.7474039793014526), (231, 0.6951025128364563), (263, 0.6758745908737183), (142, 0.6643386483192444),
 (117, 0.6626862287521362), (41, 0.6597030758857727), (295, 0.656959056854248), (280, 0.6567368507385254), (151, 0.6545502543449402), (236, 0.6545248031616211), (122, 0.652718186378479), (188, 0.6514146327972412)
 , (92, 0.6509007215499878), (291, 0.6431859731674194), (94, 0.6424540877342224), (89, 0.6417521238327026), (279, 0.6396315097808838), (270, 0.6378187537193298)]

Process finished with exit code 0
```

Figure 11

For the similarity document of the 9th litigation, the similarity of the litigations is (123, 0.8105), (201, 0.792), (15, 0.784), (37, 0.763), (64, 0.752), (215, 0.751), (192, 0.7432620525360107), (143, 0.732), and so on. The ninth issue concerns the acquisition of a company and the confusion of consumers with the trademark. As a result of a manual review of similarities in litigation, it would be found that the first five articles are all litigation with trademark infringement and corporate acquisitions, and therefore can satisfy users' needs. (As showing Table1)

Litigation	Similarity	Type	Cause
9(test)	**1**	Trademark confusion	The Acquisition contract of acquisition causes the reverse confusion
123	85.8%	Trademark confusion	Advertise of reverse confusion.
201	78.6%	Trademark confusion	Similar trademarks of the two wineries cause trademark infringement.
15	78.3%	Trademark confusion	Trademark dilution caused by internet website.
192	77.9%	Trademark confusion	Similar trademarks of the two wineries cause trademark infringement.
143	76.0%	Trademark confusion	Trademark license dispute.

Table1

Compare the As-is and To-be model with several search efficiency indicators，(Showing as Table 2)

Index	As-Is	To-Be
Search Method	Manual	Computer assist
Search Scope	less than 5%	100%
Search Efficiency	Low	High
Search Accuracy	uncertain	Systematic
Time Cost	high	lower

Table 2

Table 2 shows that As-is Model traditionally performs manual interpretation, which is less efficient and accurate than To-be model, and takes more time. In the other hand, program analysis is also more systematic, which can use the similarity relationship to search for similar cases.

5. Conclusion

With regard to the speed of receiving information and the convenience of searching, people's lives have been surrounded by a great deal of data and information. The data analysis and quantification of words can predict the behavior of human beings.

However, in the face of a large amount of information, Find what user really want, Big data is not equal to Right data, when access to large amounts of data, but also how to make the right use of the choice of kinds of methods and tools. In this case study, though there is a wealth of documentation on the web for reference, it is better than the dark days of paper space and confined space, but the attendant is how to view the information users want. Web crawler can quickly crawl more than 300 documents, and by doc2vec to analyze the similarity between the various document vectors to sort out the document recommended to the user, it can greatly reduce the user to view one by one, but also increase the breadth of the scope of the document, for the control of litigation can be more perfect. Do the right thing is better than do the thing right.

References

1. http://pushwindersingh.com/what-is-web-crawlers/

2. https://code.google.com/archive/p/word2vec/

3. Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (pp. 1188-1196).

4. http://cs224d.stanford.edu/lectures/CS224d-Lecture2.pdf

5. https://rare-technologies.com/word2vec-tutorial/

6. 玩轉社群：文字大數據實作 謝邦昌、鄭宇庭 五南出版社

7. Python 初學特訓班 文淵閣工作室 碁峰出版