Project Report

Author: Niraj Dev Pandey
Date: 04 June 2020

Introduction : The search for existing patent can be easier job for them, who knows the search keywords for designated search fields. However, it will be very difficult for the naive user who is not a subject matter expert. One solution for such scenario could be to use machine learning (ML) based search engine. Where user will enter whatever information they have with them and retrieve results which matches the closest to the query. ML have capability to understand the meaning of your words/sentences and can relate them with the data base. Even when your vocabulary isn't the same as patent documents. We know such techniques with the name "*Word Embedding*".

Data-set : The data we are going to use for this project comes from European patent office. The whole data is about 4 gigabyte. As this whole data can't be processed with small machine, such as mine. I am going to work with a small portion of data. Remember, the more data the better ML systems can be build. Less data can still be okay for first prototype though.

The data comes in .XML format. Where, we have information like, author name, patent id, date of publication, title, claim and description etc. We are aiming to provide a search engine where, one can put any of these information (in any order) and find relevant documents based on it.

Approach : The following steps were taken through-out the development phase.

1. Download huge corpus from EPO website
2. Unzip it and take a portion of this data (According to your computational power)
3. Load taken data into python (use *beutifulSoup* package)
4. Performed some analysis (find relevant data)
5. Filter target text within those XML files
6. Save filtered text into a text file
7. Find open source English Word Embedding (we can train ours too)
8. Use word moving distance (WMD) to find similarity between two documents (query and text files)
9. Test model performance and improve accordingly

A bit on the terminologies.

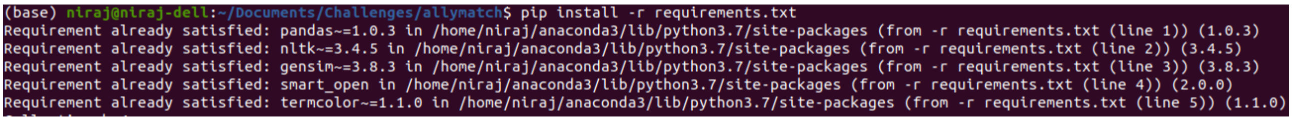
Word Embedding: We want to provide a search engine which understand the meaning of user input and should not be just looking for *literal word matching* in the patent documents. There comes the word embedding into play. Embedding is basically vector representation of words. It is achieved by a neural network model that learned to map a set of words or phrases in a vocabulary to vectors of numerical values. This helps words or phrases to be closer to each other in vector space. In other words term "king" would be close to "Queen" and words alike.

Word moving distance: Word Mover's Distance (WMD) is a promising new tool in machine learning that allows us to submit a query and return the most relevant documents. Both query and documents will be word embedded. In other words, user will enter their query regarding patent search and WMD model will return best matched patent from whole data-set. It allows us to assess the "distance" between two documents in a meaningful way, even when they have no words in common.

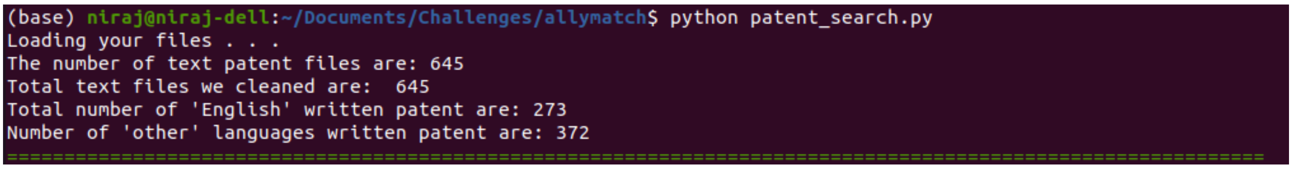
Analysis : Data insights always helps to make better prediction models. I found that some of the patent is written in different language: `['de', 'en', 'fr']`. To find out, we utilize library called *langdetect*. We are going to work with English *patents* "only for this prototype". The reason being that the ML models and word embedding will be negatively influenced if different languages are in the same document. Although, there are bilingual embedding too (in case we want to include more languages in our smart search prototype). Just to make it clear, the more languages are not at all a problem. It's just about getting huge data to train your word embedding for that particular language. We chose 645 patents for this prototype and out of these only 273 is in English. Rest can be French or German.

How to run : Follow below step-by-step process to use this project

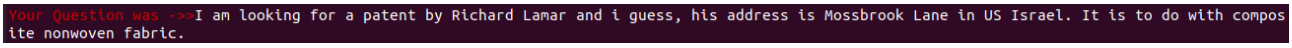
1. Download Embedding (1.5 GB) from here. Unzip it in the same directory.
2. Now you can see a python project named patent_search.py
3. I assume that you already have python & pip installed on your machine
4. The second thing you need is, the libraries which were used to develop this project
5. Find them in a file named : requirements.txt
6. Open terminal (command prompt) in the same repository
7. Type: pip install -r requirements.txt
8. This will install all required packages. (let me know, if there is still missing)
9. Type: python patent_search.py
10. Done (refer below)



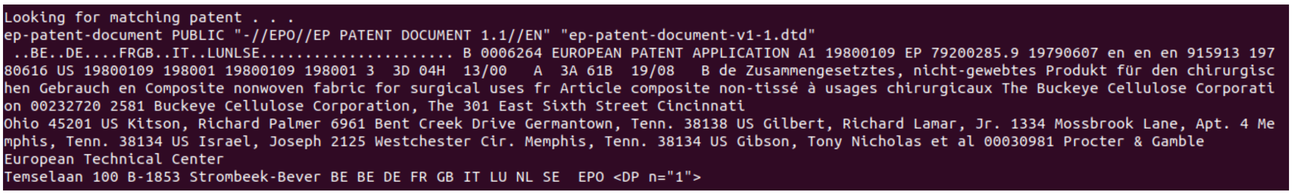
```
(base) niraj@niraj-dell:~/Documents/Challenges/allymatch$ pip install -r requirements.txt
Requirement already satisfied: pandas~=1.0.3 in /home/niraj/anaconda3/lib/python3.7/site-packages (from -r requirements.txt (line 1)) (1.0.3)
Requirement already satisfied: nltk~=3.4.5 in /home/niraj/anaconda3/lib/python3.7/site-packages (from -r requirements.txt (line 2)) (3.4.5)
Requirement already satisfied: gensim~=3.8.3 in /home/niraj/anaconda3/lib/python3.7/site-packages (from -r requirements.txt (line 3)) (3.8.3)
Requirement already satisfied: smart_open in /home/niraj/anaconda3/lib/python3.7/site-packages (from -r requirements.txt (line 4)) (2.0.0)
Requirement already satisfied: termcolor~=1.1.0 in /home/niraj/anaconda3/lib/python3.7/site-packages (from -r requirements.txt (line 5)) (1.1.0)
```



```
(base) niraj@niraj-dell:~/Documents/Challenges/allymatch$ python patent_search.py
Loading your files . . .
The number of text patent files are: 645
Total text files we cleaned are:  645
Total number of 'English' written patent are: 273
Number of 'other' languages written patent are: 372
==============================================================================================================
```



```
Your Question was ->>I am looking for a patent by Richard Lamar and i guess, his address is Mossbrook Lane in US Israel. It is to do with compos
ite nonwoven fabric.
```



```
Looking for matching patent . . .
ep-patent-document PUBLIC "-//EPO//EP PATENT DOCUMENT 1.1//EN" "ep-patent-document-v1-1.dtd"
 ..BE..DE....FRGB..IT..LUNLSE...................... B 0006264 EUROPEAN PATENT APPLICATION A1 19800109 EP 79200285.9 19790607 en en en 915913 197
80616 US 19800109 198001 19800109 198001 3  3D 04H  13/00   A  3A 61B  19/08   B de Zusammengesetztes, nicht-gewebtes Produkt für den chirurgisc
hen Gebrauch en Composite nonwoven fabric for surgical uses fr Article composite non-tissé à usages chirurgicaux The Buckeye Cellulose Corporati
on 00232720 2581 Buckeye Cellulose Corporation, The 301 East Sixth Street Cincinnati
Ohio 45201 US Kitson, Richard Palmer 6961 Bent Creek Drive Germantown, Tenn. 38138 US Gilbert, Richard Lamar, Jr. 1334 Mossbrook Lane, Apt. 4 Me
mphis, Tenn. 38134 US Israel, Joseph 2125 Westchester Cir. Memphis, Tenn. 38134 US Gibson, Tony Nicholas et al 00030981 Procter & Gamble
European Technical Center
Temselaan 100 B-1853 Strombeek-Bever BE BE DE FR GB IT LU NL SE  EPO <DP n="1">
```

Remarks: Now, the program will run and let you know every steps in the progress. It will take some time to finish. Like 2 to 3 minutes. Thereafter, you will see a colorful text. Which reads "Your Question". You can type any information you have with yourself and let the model find best suited patent for your query. You can go on until you pleases. Enter "stop" to finish the program. This is ML language model, so it usually expect you to provide long information. If you just provide the name of author, this model will choose the documents where author name is found anywhere in all doc. It will pick the one with best accuracy. Such models can also be built, however, for this prototype, we are playing with more than just keyword search.

Thanks a lot. Please don't hesitate if there is something you need to run it successfully. I am always at your disposal. I am eagerly looking forward to hearing from you. Stay safe and healthy.