Capstone Project 2 Milestone Report

Automated Semantic Search for Patents

Summary

Semantic search on patents has been explored for years and is to discover other patents with similar contents instead of keywords. It can increase the efficiency of finding patents with similar contents for patent attorney, R&D departments and patent software organizations. This report explains a new approach of automatically discovering semantically relevant patents with AI/ML models with the small datasets. This reports takes a small manually selected seed focusing on a topic, such as “purify water” and “filter water”. Then the key information is extracted from the seed and expands seed to a larger document sets. Semi-supervised model is applied to classify this larger document sets with the small human-selected anti-seed.

Project Overview

- **Project Background**

Patent Semantic search is to search patent document based on the meaning of search words instead of keywords. Currently, the rule-based software engineering is the most common approach in real products. However, it requires very sophisticated techniques such as software engineering, computing linguistics and domain experience. To keep minimal manual efforts and improve efficiency in search, some data scientists are also leveraging Neural Network in supervised machine learning for semantic search. For example, Hamel Husein uses Deep Learning to build semantic search for arbitrary objects. But the huge amount of labeled data are needed, which also prevent AI/ML from further use in semantic search.

- **Project Scope**

Inspired by Google’s research work, we can create a AI/ML semi-supervised model to automatically explore semantic patent search with similar schema. Given that patent contains many technical jargon, we select a topic “purify water” and “filter water” to simplify the project. A small seed dataset containing “purify water” and “filter water” are manually selected and

labelled. And synonym phrases would be extracted from this dataset. Then these phrases, as keyword in patent abstract, would extract more patents.

- **Project Benefits**

This approach can utilize the advantage of less manual efforts and dependency on domain expertise. Moreover, it requires smaller labelled data and improve the data collections. It will reduce the barrier of applying ML/AI in patent semantic search.

- **Targeted Customer**

This can improve the efficiency and user' experience in patent searching. Targeted Customer includes Google Patent Search Group, Patent Search Software Company or other organization heavily related to patent search.

- **Approach**

1. Data Wrangling.

   This step is really fundamental to the whole process. Unlike other data science project, the project retrieves key information from seed analysis and search more data in Google Patent Public dataset. Ultimately, we have three datasets: seed (labelled), anti-seed (labelled), expanded seed(unlabelled)

   Seed: Starting from seed data (classified patents including patent number and title only), we extract abstract through Google Bigquery from Google Patent Database for further analysis. So the seed patents are added to 1285 including "purify water" and "filter water".

   Anti-seed: The anti-seed dataset, opposite to seed dataset, is also manually selected from patents containing "algorithm" without "purify water" or "filter water". These dataset are manually labelled.

   Information Extraction. Three methods: Bigram(tfidf), Doc2Vec and Word2Vec are tried to extract key information. The result from Bigram seems more relevant to the ones from Doc2Vec. From the top 20 features of tf-idf can be used for keywords for expanding data in Bigquery. Another option is that we can leverage pre-trained word2vec model from 5.7M (by Google) to obtain the similar words to "purify", "filter" and "water". These words can be also used as keywords to expand data.

   Seed Expansion. The expanded data can be retrieved based on keywords in abstract. The expanded data are treated as unlabelled data.

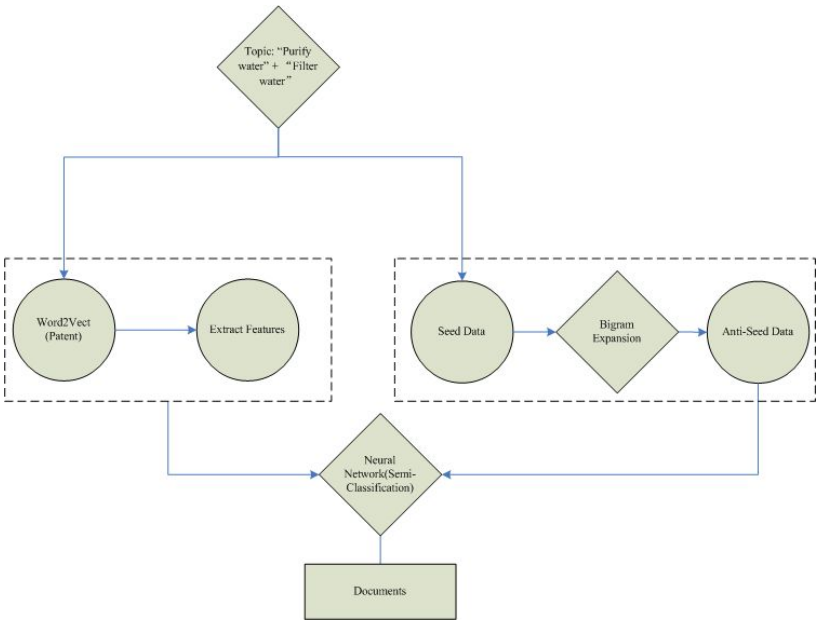
2. Data Analysis. Since the key information from seed dataset is a phrase with two words, unigram and bigram are our focus in EDA. And the length and words of abstract are also visualized.

3. Model Building.
   Feature Engineering: The input would be texts from patent abstract. So we would convert them to numbers by tfidf (unigram, bigram) or word2vec.
   Model Algorithm. Semi-supervised model is applicable to this problem due to small labelled dataset and large unlableled datasets. The algorithm is designed as two steps: The first step is to build a classifier with labelled data (seed and antiseed). It is very similar to supervised model. And the second step is to label the expanded patent (unlabelled data) with the classifier. We will start from Naive Bayes as baseline classifier and try other classifiers as XGBoost and Neural Network (RNN and LSTM) with/without pre-trained word-embedding model.

4. Model Evaluation.Two Metrics can evaluate the performance of model algorithm. The first metric is AUC ROC to evaluate the classifier with labelled data. The second metric is the classification accuracy for sampled data from unlabelled dataset.

5. Hyperparameter Optimization. After the models would be compared with metrics, we will select the best model and optimize the model's hyperparameters.

The drafted schemas follows as the figure below:



- **Data Source**

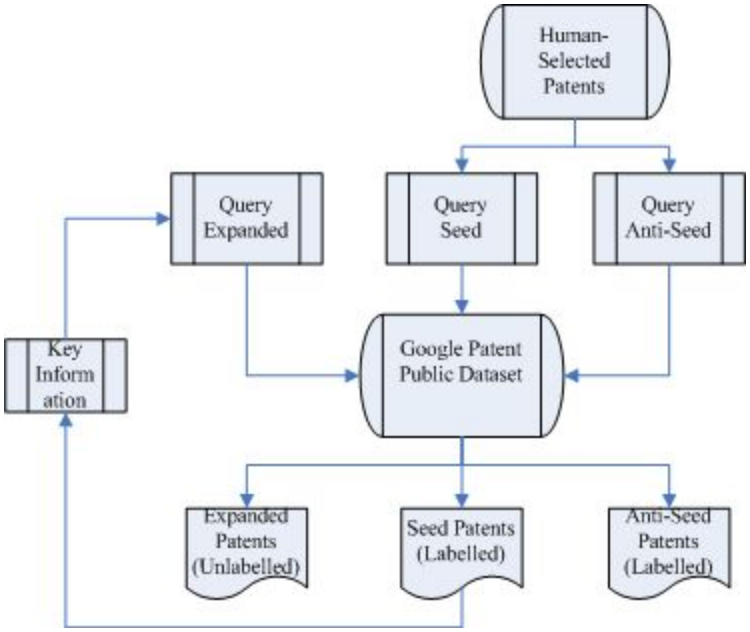
The data mainly comes from Google Public Data. We are using Bigquery to extract data**.** All patents are strictly limited in US patent database. Abstracts in patents are only analyzed since this project focuses on main ideas of patents,. Here is the link.

- **Deliverables:**

A Final Report, PowperPoint Slides, Source Code at Github. I might also publish it at my medium blog.

Data Wrangling

In this project, data wrangling are repeated to extract data for EDA analysis and modelling. This fundamental step is to expand human-selected patents and results in three kinds of datasets: Seed, Anti-seed and Expanded Patents. The first two are human- labelled and the last one is unlabelled. The steps are shown in the figure below.



Here are details on every step.

- Seed: Fundamental dataset contains “purify water” or “filter water”. At first, we tried 87 patents relevant to “purify water”, which is insufficient to extract key information for next step. Then the seed is increased to 1285 and human-labelled dataset. Through Bigquery, we find abstracts based on patent number and title. When the seed data is larger, it can cover more topics to extract more information. Seed dataset is labelled as “1”.

	publication_number	title	abstract
0	US-6652816-B2	Apparatus for generating ozone and anion	An apparatus for generating ozone and anion ca...
1	US-4655910-A	Liquid filtering device	Liquid filter devices, particularly for irriga...
2	US-3996136-A	Pump-filter for bilge water	A housing for a bilge pump-filter has a lower ...
3	US-6036178-A	Device for mixing air and water in a water pur...	A device for mixing air and water in a water p...
4	US-5813245-A	Pressure relief circuit for refrigerator conta...	A refrigerator having a water filtration and d...

- Anti-seed: This dataset has the opposite polarity to seed. Patents including “algorithm” are searched to increase polarity. Through Bigquery, the keyword “algorithm” in abstract is searched and exclude any containing “purify water” or “filter water” to minimize the overlapping with seed. Google’s work recommends the range of anti-seed number from 10,000 to 40,000. So we select 20,000 for anti-seed datasets and labelled as “0”
- Key Information Extraction: The key information would be extracted from the seed datasets. Three approaches for extraction were tried: Tfidf (bigram), Word2Vec,

Doc2Vec. The results show Tfidf can produce the expected results. So we choose top 10 phrases. More phrases would have broader coverage on results but need more time for searching patents.

```
['water purification',
 'reverse osmosis',
 'waste water',
 'water filter',
 'purified water',
 'filter cartridge',
 'filter water',
 'present invention',
 'water treatment',
 'filter element']
```

We can see that the keywords are expanded from “purify water” or “filter water” to other words such as “water treatment”. Meanwhile, there are also unwanted information such as “present invention”. Since little human-intervention is involved, we still use these keywords to find patents and have machine learning algorithm prune the results.

Pre-trained Word2Vec model for patents trained by Google can match single words only. For example, top 10 of words close to “purify” found from Word2Vec are

```
[{'word': 'purifying'},
 {'word': 'purification'},
 {'word': 'purified'},
 {'word': 'adsorbent'},
 {'word': 'adsorbing'},
 {'word': 'adsorbed'},
 {'word': 'adsorb'},
 {'word': 'purifier'},
 {'word': 'treat'},
 {'word': 'purity'}]
```

Top 10 of words close to “water” are:

```
[{'word': 'soluble'},
 {'word': 'dispersible'},
 {'word': 'hot'},
 {'word': 'sea'},
```

```
{'word': 'tank'},
{'word': 'steam'},
{'word': 'jet'},
{'word': 'cold'},
{'word': 'swimming'},
{'word': 'purified'}]
```

We can combine these similar words by the order of “purify water” as keywords to expand more patents with containing.

We also tried Doc2Vec. But it does not provide the ideal results.

In this project, tfidf for patent expansion is mainly discussed and used. The technique of Word2Vec for expansion is very similar.

- Seed Expansion:

  With the extracted keywords from the previous step, Bigquery searched more patents. Patents as expanded patents are expanded 1285 to 1176846. Among them, most patents are apparently irrelevant to seed patents, such as the topic containing “present invention”.

Three datasets are prepared for the following steps:

Dataset Name	Dataset Size	Labelled
Seed Patents	1285	Yes
Anti-seed Patents	20,000	Yes
Expanded Patents	1,176,846	No

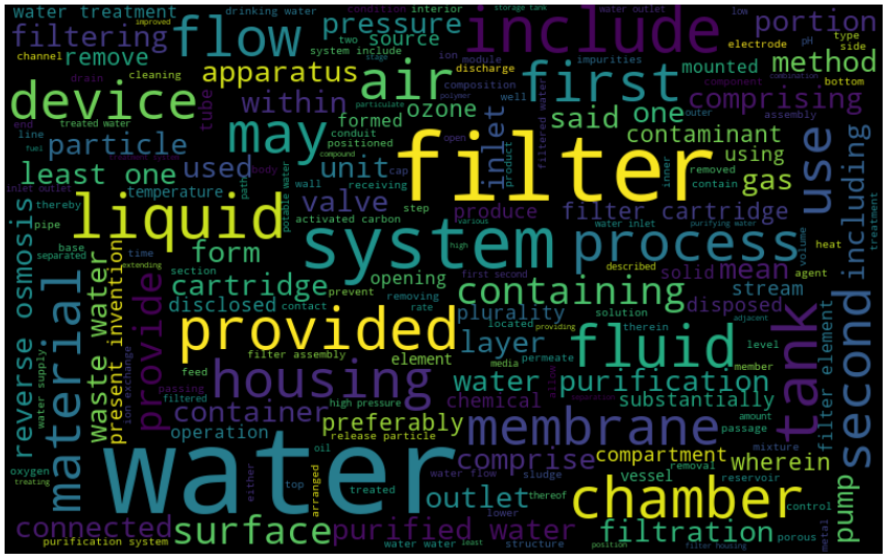
Data Cleansing:

The three datasets we are dealing with are all texts in abstracts of patents and our purpose is to recognize and classify the expanded patents. Because the extraction of three datasets are performed individually, there may be duplicates among those three datasets. Those duplicates are removed. The capital letters are changed to lower cases. Patent often contains non-alphabetic characters. So non-alphabetic characters (number and punctuation) are

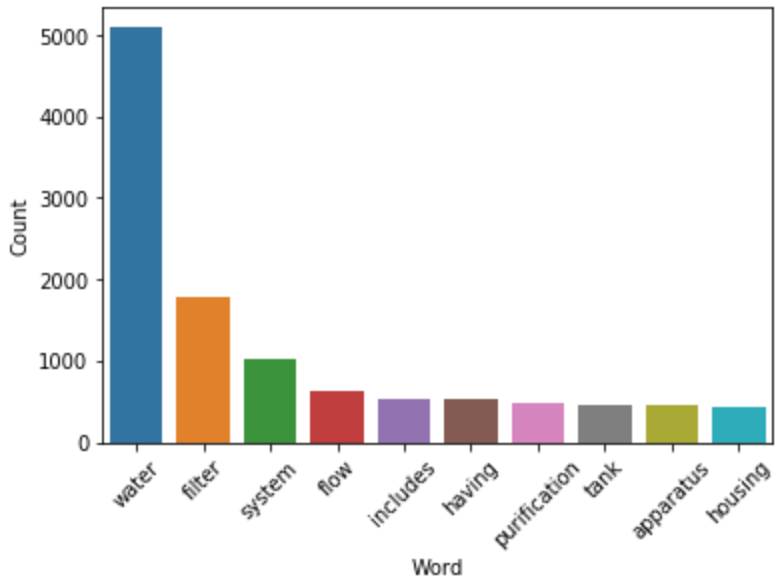
removed and stop words such as “am”, “is”, “to ” etc. are removed. This applies to all three datasets.

Data Visualization:

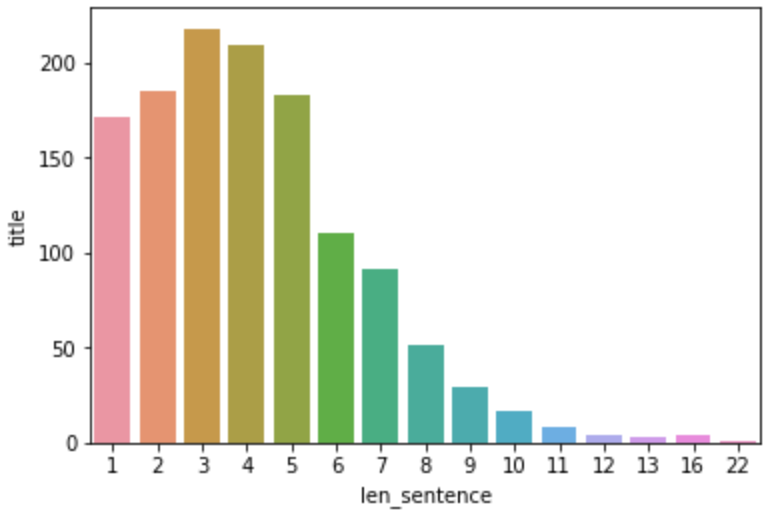
Let’s dig deeper about the dataset visualization. We are taking seed patent as an example here. Technique would be similar and you can find complete results in the powerpoint slides in this folder. The below chart indicates the most commonly words in texts in WordCloud.



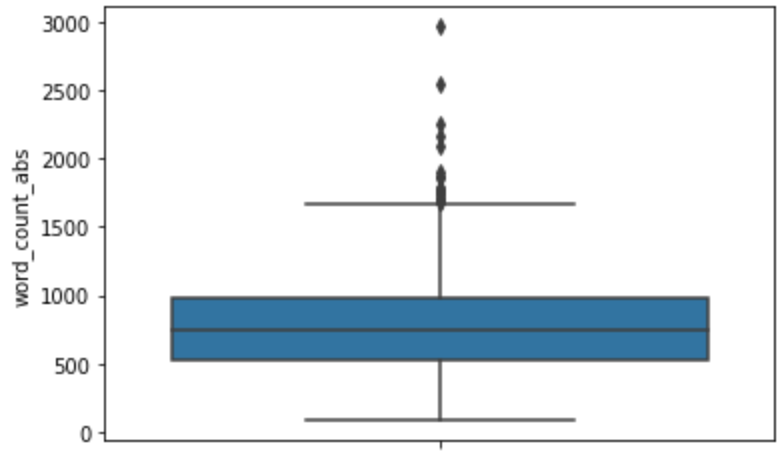
The chart below shows the top 10 words in abstracts. We can see the word “water” is top 1.



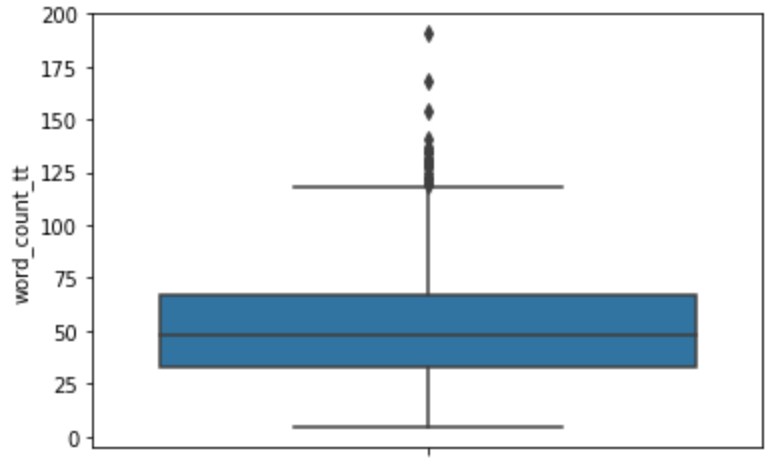
We can see most abstracts would be less than six sentences.



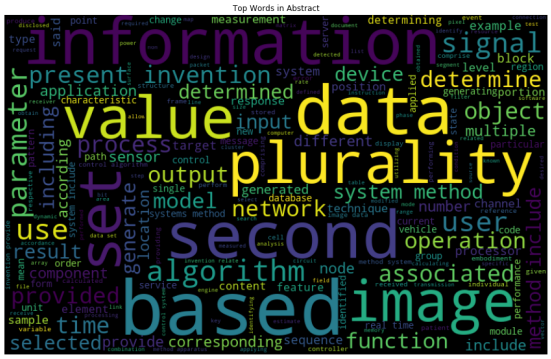
The median of word number in abstracts is around 750.

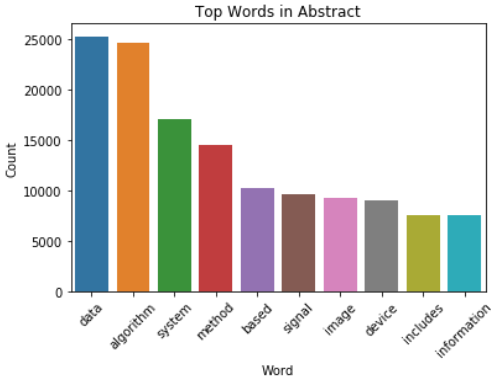


The median of word number in title is around 50.



Antiseed focuses mainly on data and algorithm





Expanded dataset is very unbalanced.