Capstone project 2: Final Report

Project Title:
Application Recommendation System for Biodegradable Polymer

< Abstract >
The goal of this project was to make an interactive application recommendation system that recommended potential applications for a given polymer by a user. A biodegradable polymer was chosen as a polymer kind for this project.

Patent information from WIPO was used for this project. According to EDA, the data included 8,182 patents published in 37 countries and written by 19 languages (by fastText.) This wide language variety made this project more complicated, interesting, and worth pursuing to obtain multi-skills. 90% were written in English or Japanese. I decided to use English and Japanese abstracts to calculate the sentence similarities. The Japanese patents were translated into English by googletrans.

Two application recommendation systems were created. One system accepts one patent and another accepts two IPC codes as input by a user, and they provide 10 potential applications. The recommendation systems use two features (a network and sentence similarity) to extract and prioritize potential applications for maximizing the accuracy of the recommendation.

< Table of Contents >

1. Problem Statement

1.1. Problem

Developing new applications of a polymer is one of the most important and challenging steps to sell a polymer as a product. It requires a wide range of knowledge about polymer itself (physical property, formability, etc.) and products that might already exist or might not yet. Scientists would use research papers, databases, and patents as a reference. However, this is a quite painstaking process to search similar polymers' features and the applications, and combine all information to think about new applications of their polymer. I was a polymer scientist for more than 10 years. I decided to create a system that I had wanted at the time.

Here, I provided an application recommendation system that gives some hints about potential applications for a given polymer by a user. A biodegradable polymer was chosen as a polymer kind for this project. Biodegradable polymers have been researched for decades. There are mainly two aspects to get attention; to reduce the effects of plastics on the environment, and as a bioabsorbable polymer, which are degraded and absorbed in our body. Both fields are growing with raising environmental awareness and advancing in medical technology.

1.2. Client

The first clients would be chemical companies that research, manufacture, and sell biodegradable polymers as raw material, and that buy the polymers and form them for their products. Their purposes could be to accelerate the step to extract potential applications of biodegradable polymers.

2. Exploratory Data Analysis

2.1. Data

Patents were used to make the recommendation system. A patent includes information about what polymer was used, how to prepare it, for what it could be used, who invented it, and so on. The material and application information provided useful information for this project.

The patent data was acquired from the database of the World Intellectual Property Organization (WIPO) with a search word "IC:(C08L 101/16)." C08B 101/16 is the IPC code (International Patent Classification Code) for biodegradable polymers. The data had 8,182 patent information, and the columns contained Application ID, Application Date, Country, Title, Abstract, IPC Codes, Applicants, Inventors. etc.

IPC has a hierarchical structure. All technical fields are divided into eight "sections" from A to H:

- A: Human Necessities
- B: Performing Operations, Transporting
- C: Chemistry, Metallurgy
- D: Textiles, Paper
- E: Fixed Constructions

F: Mechanical Engineering, Lighting, Heating, Weapons, Blasting

G: Physics

H: Electricity

There are "classes" in each section, and each class has "subclasses". Then, there are "maingroups" under a subclass, and finally "subgroups". For example, C08L 101/16 means:

(Section) C: Chemistry, Metallurgy

(Class) C08: Organic macromolecular compounds, Preparation or chemical working-up

(Subclass) C08L: Macromolecular compounds

(Maingroup) C08L 101: Unspecified macromolecular compounds

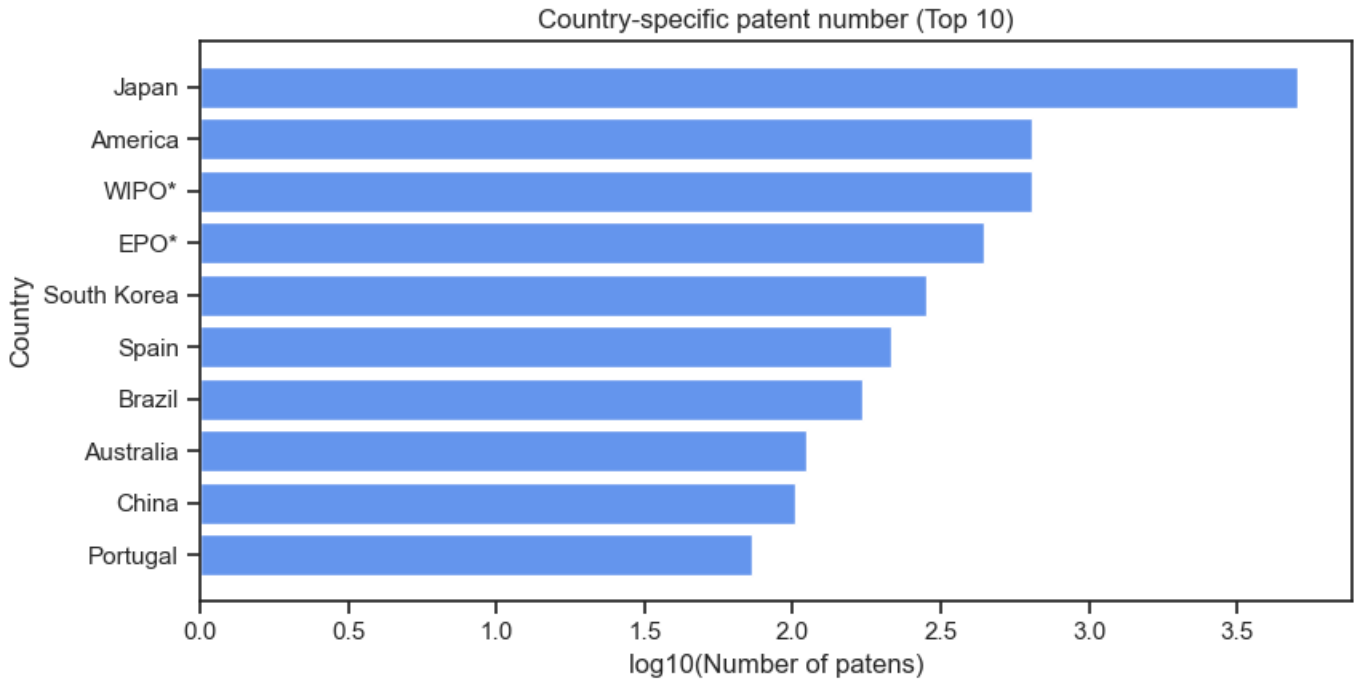
(Subgroup) C08L 101/16: Biodegradable macromolecular compounds

Each IPC code represents a material, application, or technology to manufacture it.

2.2. Countries where patents were filed

The patent information in WIPO is gathered from 193 member countries. The coverage is not exhaustive, but wide. If many patents are registered in a country, at least we would be able to say biodegradable polymers are popular in the country, and it would be connected to the potential market.

According to the analysis of the 'Country' column, the patent data were from 37 countries (and organizations). Figure 1 showed the Top 10 countries (and organizations) where the patents were registered.



* WIPO = World Intellectual Property Organization

* EPO = European Patent Office

Figure 1. Countries patents were filed in (Top 10)

Japan and the U.S. were the top 2 and followed by WIPO and EPO. Whereas they are the potentially big markets, that would affect the prediction result of the recommendation system because applications that are popular in the top countries tend to be recommended. Each country (or organization) has cultural, climate, and morbidity differences with others. The features of the top countries will tend to be captured more strongly because of the data tendency. If you want to find some applications in one of the miner countries in the data, you need to be aware of that.

2.3. Frequency of each IPC code in data

The data had a column called 'IPC.' This column showed the IPC codes the patent was categorized in. 4,344 kinds of IPC codes were included, and the frequency of each IPC code was counted. Table 1 showed the summary statistics.

Table 1. The summary statistics of the frequency of each IPC code

Min	25%	50%	75%	95%	Max	Mean	SD
1	1	2	6	35	2,830	10.6	60.4

1,665 (38%) IPC codes appeared only once in the dataset. They would not be much useful to predict applications because the network ends at the IPC codes. However, a user might input the IPC codes and want to start from one of the IPC codes to find other applications. So, I decided to keep them. On the other hand, six IPC codes appeared more than 500 times. This means there are more than 500 connections once reaching the IPC codes. That was why I decided to use the abstracts in addition to IPC codes to create the recommendation systems to prioritize the applications.

2.4. Popular polymer

Here, the IPC codes starting with 'C08L' representing polymers, were extracted. Then, the maingroups were counted to identify the popular polymers in the dataset. The maingroups of 'C08L' represent a polymer kind. For example, 'C08L 1' is 'cellulose', and 'C08L 67' is 'polyesters.' Although there are two more nested categories indicated after ' / ', the category until just before '/' was used in this section.

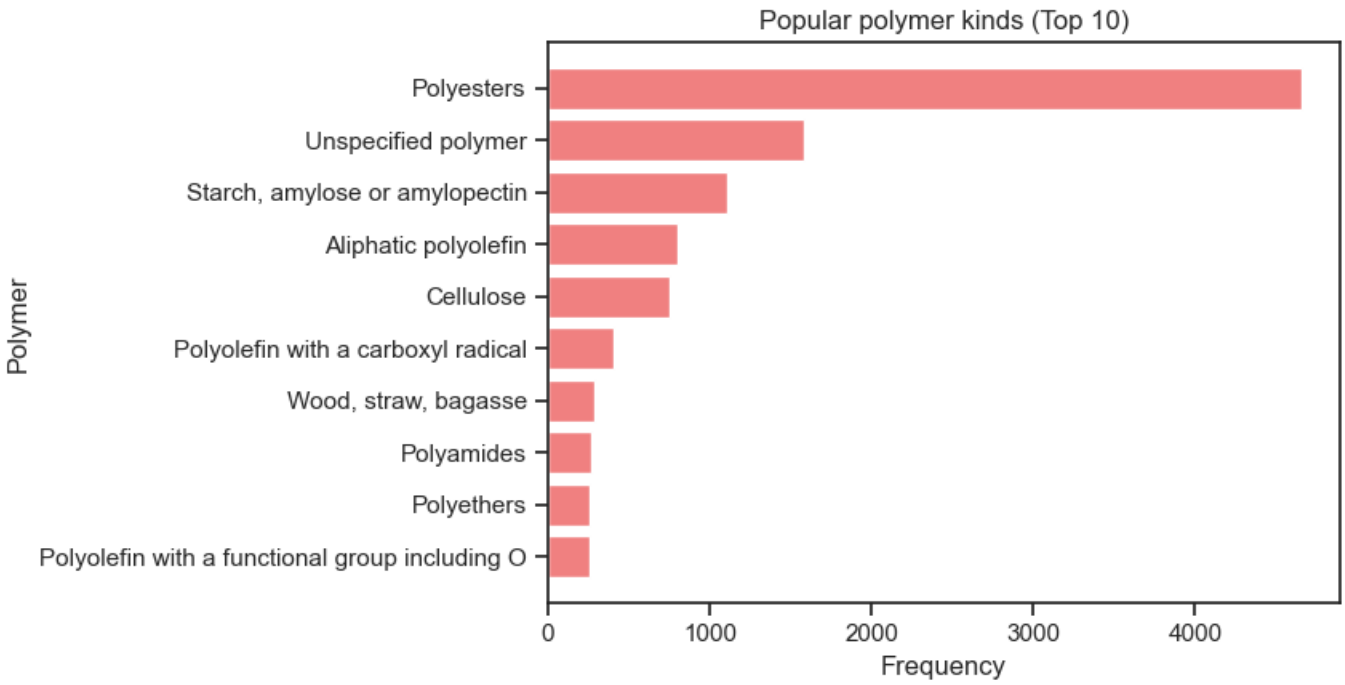


Figure 2. Popular polymer kinds (Top 10)

There were 49 kinds of polymers in the data, and polyesters were the most popular (figure 2.) This would be because polylactic acid and polyglycolic acid are very popular in biodegradable polymers.

2.5. Popular applications

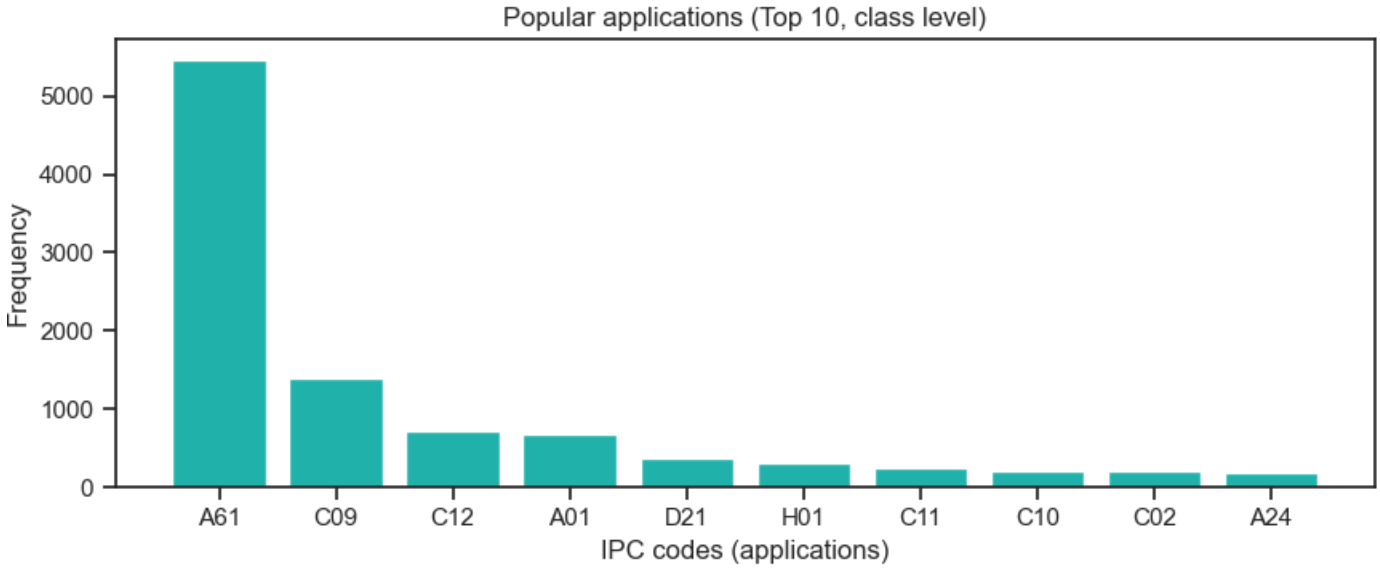
First, I needed to decide what IPC codes were treated as applications to search what were the popular applications in the data. I carefully checked the classes and extracted the following groups as applications (Table 2.) These categories were used as IPC codes of applications to make a model. As a result, applications recommended by the model are from them.

Table 2. IPC Classes for applications

Section	Class	Content
A	(all classes)	HUMAN NECESSITIES
B	09	DISPOSAL OF SOLID WASTE, RECLAMATION OF CONTAMINATED SOIL
	27	WORKING OR PRESERVING WOOD, NAILING OR STAPLING MACHINES
	28	WORKING CEMENT, CLAY, OR STONE
	31	MAKING ARTICLES OF PAPER, CARDBOARD OR MATERIAL, WORKING PAPER, CARDBOARD OR MATERIAL
	41	PRINTING, LINING MACHINES, TYPEWRITERS, STAMPS
	42	BOOKBINDING, ALBUMS, FILES, SPECIAL PRINTED MATTER
	43	WRITING OR DRAWING IMPLEMENTS, BUREAU ACCESSORIES

	44	DECORATIVE ARTS
	60	VEHICLES
	61	RAILWAYS
	62	LAND VEHICLES FOR TRAVELLING OTHERWISE THAN ON RAILS
	63	SHIPS OR OTHER WATERBORNE VESSELS, RELATED EQUIPMENT
	64	AIRCRAFT, AVIATION, COSMONAUTICS
	65	CONVEYING, PACKING, STORING, HANDLING THIN OR FILAMENTARY MATERIAL
	66	HOISTING, LIFTING, HAULING
	67	OPENING OR CLOSING BOTTLES, JARS OR SIMILAR CONTAINERS, LIQUID HANDLING
	68	SADDLERY, UPHOLSTERY
C	02	TREATMENT OF WATER, WASTE WATER, SEWAGE, OR SLUDGE
	03	GLASS, MINERAL OR SLAG WOOL
	04	CEMENTS, CONCRETE, ARTIFICIAL STONE, CERAMICS, REFRACTORIES
	05	FERTILISERS
	06	EXPLOSIVES, MATCHES
	09	DYES, PAINTS, POLISHES, NATURAL RESINS, ADHESIVES
	10	EXPLOSIVES, MATCHES
	11	DYES, PAINTS, POLISHES, NATURAL RESINS, ADHESIVES
	12	BIOCHEMISTRY, BEER, SPIRITS, WINE, VINEGAR, MICROBIOLOGY, ENZYMOLOGY, MUTATION OR GENETIC ENGINEERING
	13	SUGAR INDUSTRY
	14	SKINS, HIDES, PELTS OR LEATHER
	23	COATING METALLIC MATERIAL, COATING MATERIAL WITH METALLIC MATERIAL, CHEMICAL SURFACE TREATMENT
D	21	PAPER-MAKING; PRODUCTION OF CELLULOSE
E	(all classes)	FIXED CONSTRUCTIONS
F	(all classes)	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
G	(all classes)	PHYSICS
H	(all classes)	ELECTRICITY

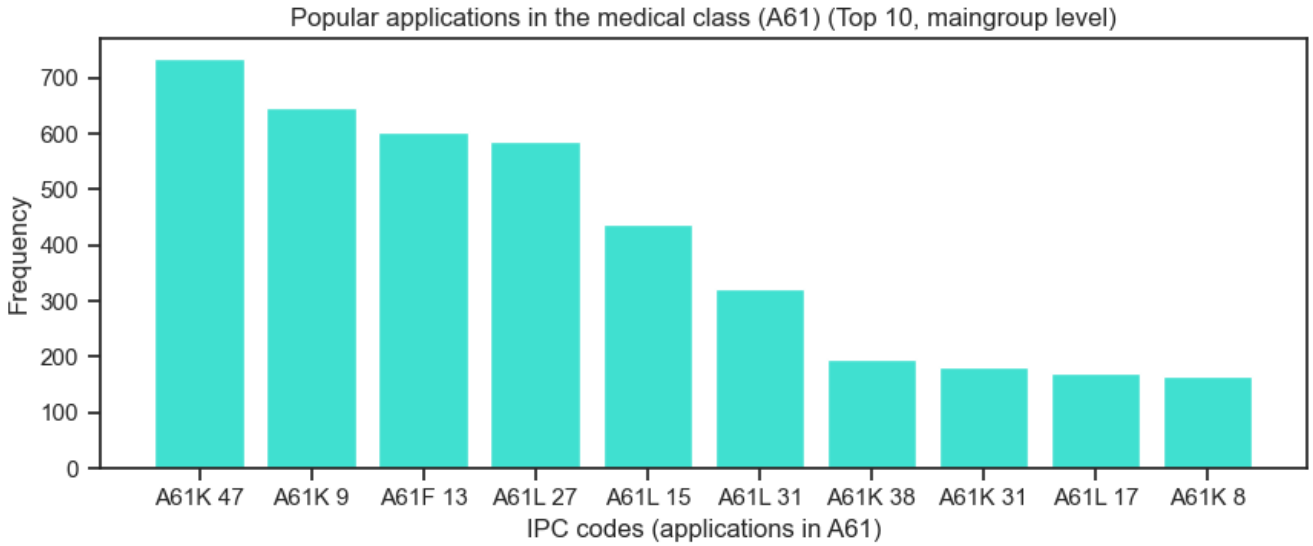
Then, each of the above IPC codes was counted at a class level, and the most frequent 10 applications were shown in Figure 3.



A61: medical or veterinary Science, hygiene
C09: dyes, paints, polishes, natural resins, adhesive
C12: Biochemistry, Microbiology, Enzymology, mutation or genetic engineering
A01: agriculture, forestry, animal husbandry, hunting, trapping, fishing
D21: paper-making, production of cellulose
H01: basic electric elements
C11: animal or vegetable oils, fats, fatty substances or waxes, detergents, candles
C10: petroleum, gas or coke industries
C02: treatment of water, waste water, sludge
A24: tobacco, cigarettes, simulated smoking devices

Figure 3. Popular application fields (Top 10)

According to the plot above, medical use (A61) was the largest group as applications of biodegradable polymers (48% (5,458/11,272)). A61 is a class. In the A61 class, there are many subclasses, and under the subclasses, there are many maingroups, which explain more detail. The maingroups in A61 were explored to determine what were the popular applications in the A61 class.



A61K 47: Medicinal preparations (the non-active ingredients used, e.g. carriers or inert additives),

targeting or modifying agents (chemically bound to the active ingredient)
A61K 9: Medicinal preparations (special physical form)
A61F 13: Bandages, dressings, absorbent pads
A61L 27: Prostheses, coating prostheses
A61L 15: Chemical aspects of materials for bandages, dressings or absorbent pads
A61L 31: Other surgical articles
A61K 38: Medicinal preparations (containing peptides)
A61K 31: Medicinal preparations (containing organic active ingredients)
A61L 17: Surgical sutures, ligature for blood vessels
A61K 8: Cosmetics or toilet preparations

Figure 4. Popular application in the medical class (A61) (Top 10)

'A61 47', 'A61 9', 'A61L 27', 'A61K 38', 'A61K 31', and 'A61L 17' could be used inside our body. It makes sense that the material would be required to biodegrade in our body. 'A61F 13', 'A61L 15', 'A61L 31', and 'A61K 8' are used outside of our body. The 'biodegrade' might mean that the material biodegrades by water, bacteria, or enzyme in a natural environment.

2.6. Language variety

According to the first five rows of the data, it included at least Spanish, Chinese, English, and Japanese. I was going to use an abstract to calculate the similarities between patents. It was important to know what languages were used and the percentage of each.

The 'Abstract' column had 578 missing values, and 569 patents out of 578 had a title. I decided to use the titles as the abstracts for them because a title had some information about a patent even if it was shorter than an abstract. Also, the 9 patents without the title nor abstract were left without change because the IPC codes could be used at least.

Here, I used a language identification model by fastText. FastText is a library developed by Facebook. The language identification model predicts the language used in a given text. It can recognize 176 languages. This time, when the prediction was more than 70% sure, the predicted language was adopted. Here, I chose 70% as a cut-off point because of the balance between the accuracy and the number of 'unsure.' The accuracy was still 100% at the 100 random samples when using 70% as a cut-off point. The languages of 130 (1.6%) patents were sure less than 70%. Some of the patents were determined by the ‘Country’ label, and the rest of them were identified by country by hand.

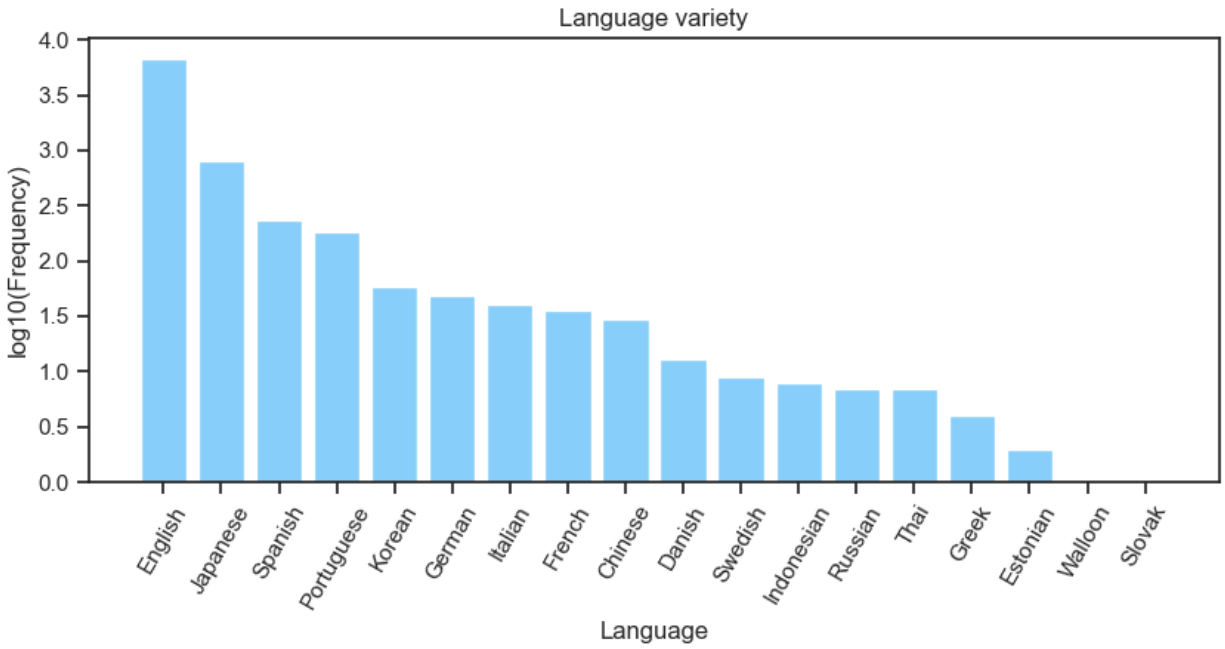


Figure 5. Language variety

19 kinds of languages were detected. 81.7% (6687/8182) of the abstracts was written in English. 9.8% (799/8182) was written in Japanese. The abstracts in English and Japanese were used to calculate the similarity of patents because I can understand the two languages. On the other hand, the IPC codes could be used to analyze the relationships between patents even if a patent was written in the other languages. So, I kept them, too.

3. Data Wrangling

The purpose of this section was to make the data ready for modeling. The 'Abstract' column was processed to prepare for calculating the similarities between patents.

3.1. Deal with missing values and duplicate data

This dataset had 8,182 rows (patents) and 9 columns; ‘Application_Id’, ‘Application_Number’, ‘Application_Date’, ‘Country’, ‘Title’, ‘Abstract’, ‘IPC’, ‘Applicants’, and ‘Inventors.’ There was no duplicate data. The six columns (‘Application_Number’, ‘Application_Date’, ‘Title’, ‘Abstract’, ‘Applicants’, and ‘Inventors.’) had missing values. As we already discussed the ‘Abstract’ column, the missing values were filled by the title if the patent had a title. The ‘Application_Date’, ‘Title’, ‘Applicants’, and ‘Inventors’ columns are used for result output of the recommendation systems. In this case, there is no problem to have any missing values because the result provides additional information about a patent and not for the essential structure or function of the recommendation systems.

3.2. Preprocess the abstracts

9.8% of the abstracts were written in Japanese. First, those Japanese abstracts were translated into English to use them to calculate the similarities between patents. I used a translation library called googletrans, which was developed by Google. It was free, but there was a restriction that was a maximum of 15k words in 24 hours from one IP address. I divided those abstracts into two parts so that each of them had less than 15k words. (If you want to avoid the

restriction, there is a paid version.) To evaluate the translation, 80 samples were randomly picked up and checked manually. Sometimes the translations were not straightforward, but the meanings were mostly fine. I assumed the rest of the translations were also accurate.

Second, the English abstracts (including the abstracts translated from Japanese) were extracted (7,486 abstracts, 91.5%), and the punctuations were removed. Compound names generally have numbers and punctuations (e.g., poly (1,5-dioxepan-2-one).) Because I would like to remain the sets of numbers and structure names, the punctuations were removed before the tokenization.

Next, the abstracts were tokenized. Then, the case was lowered. The stop words and tokens having only numbers were removed. Lastly, the rest of the words were transformed into the base forms by stemming.

As a result, 14,487 unique tokens remained. The top 100 frequent words were shown in Figure 6. Some words related to polymer kinds (e.g., polyest(er), starch, and cellulos(e)), to forms (e.g., film, particle, and fiber), and to applications (e.g., organ.)

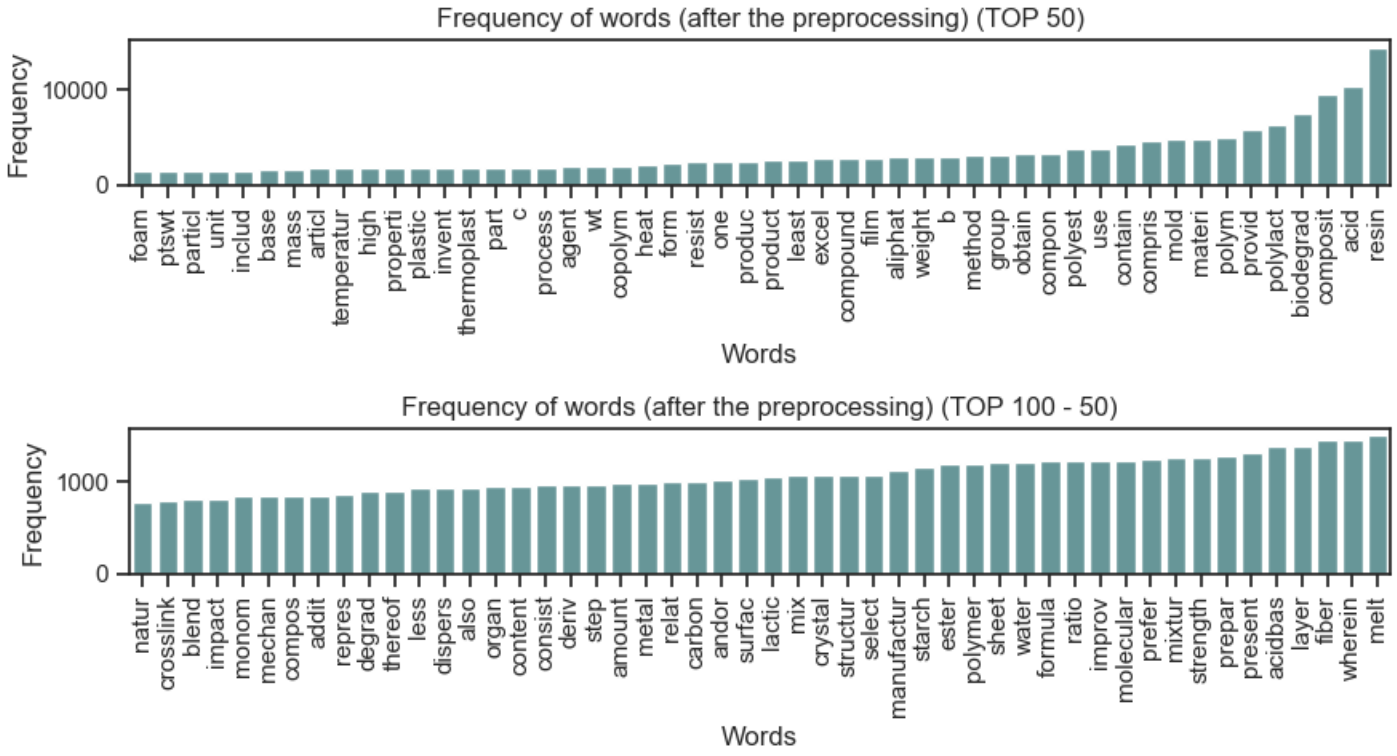


Figure 6. Frequency of words (TOP 100)

4. Modeling

Network analysis and a sentence similarity were combined for the recommendation systems to maximize the accuracy of recommendation. A network allowed us to easily extract patents sharing IPC codes with an input patent. Although the extracted patents were guaranteed to share at least one feature (IPC code), the number of the extracted patents could be large (because some patents had many IPC codes.) That was why the sentence similarity of the abstracts

was used to prioritize them.

4.1. Sentence similarity

Tf-idf vectorizer was chosen to transform the tokens into a vector so that the importance of each word was standardized. A unigram, bigram, and trigram were used because compound names were often composed of several words. For example, methyl methacrylate, poly (1,5-dioxepan-2-one), or poly (ethylene -co- acrylic acid.) They should be considered in the calculation. To reduce the complexity, the tokens appearing less than three patents were ignored. The tokenized abstracts of 7,355 patents (90%) that had an English or Japanese abstract were vectorized on the conditions. As a result, the vectors had 45,438 dimensions.

Next, the sentence similarities were calculated by cosine similarity to the vectors. Cosine similarity is often used to measure the similarities between texts. To evaluate the accuracy of the similarities, the number of common IPC codes was counted. IPC codes are not exhaustively assigned, but the tendency would be shown. All of the patents (7,355) were sorted by the similarity scores to each of the patents. Then, the scores were averaged at each rank. The top 20 were shown in Figure 7. The higher similarity scores, the more common IPC codes. The similarity calculation tended to be accurate. The similarity scores were used for the recommendation system.

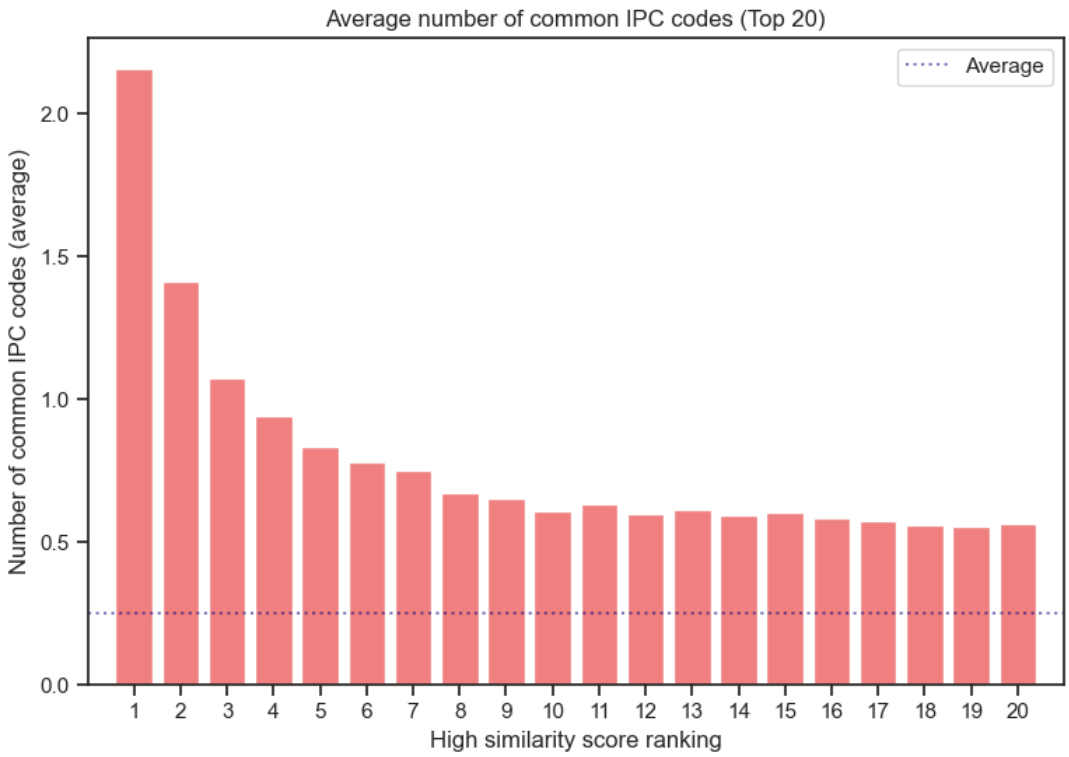


Figure 7. Average number of common IPC codes (Top 20)

4.2. Build a network

The second feature of the recommendation systems was a network. The network was composed of nodes and edges. There were two kinds of nodes: patent nodes and IPC code nodes. A patent node was connected to IPC code nodes by edges if the patent had the IPC codes. There were 8,182 patents and 4,344 IPC codes in the data, then the total

number of nodes was 12,526. The number of edges was 46,012 in total.

Some IPC classes had been chosen as applications (Table 2.) Each IPC code node was labeled using the list, and 2,153 out of 4,344 IPC code nodes were categorized as ‘application.’ Because the recommended applications would be chosen from the existing application IPC codes in the data, the system would show 10 out of 2,153 applications each time the system runs.

As an example, I introduce a part of the network around 'JP271085664.' The patent 'JP271085664' had four IPC codes ('A61L 31/00', 'C08K 5/521', 'C08L 101/00', and 'C08L 67/00'), and the four IPC codes were connected to 2,313 patents in total. Because it's still too big to visualize the partial network, I picked up one IPC code ('A61L 31/00'), and visualized the network having nodes of 'JP271085664', the four IPC codes and the neighbor patents of 'A61L 31/00' (Figure 8.) 'JP271085664' (red) was connected to the four IPC codes (blue.) Then, 'A61L 31/00' had a connection with 97 patents. According to the network plot, 'C08L 67/00' shared 13 patents with 'A61L 31/00' (purple.) This network was a part of the original graph. The whole graph would be more highly connected with one another.

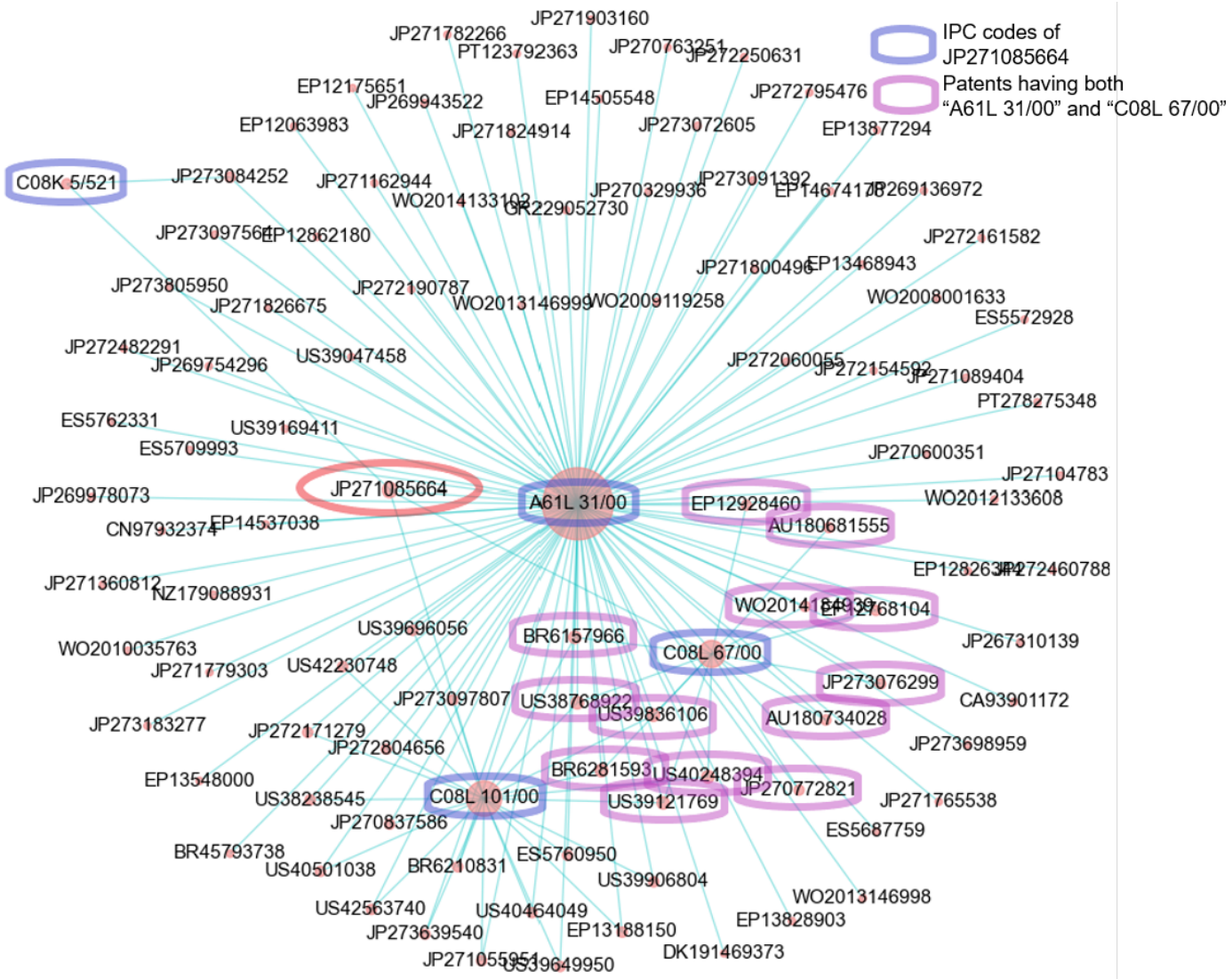


Figure 8. Network around JP271085664

The graph G had two kinds of nodes (patent nodes and IPC code nodes.) Next, I made a new graph having only patent nodes. The new graph was made by projecting the relationships between patent nodes and IPC code nodes. For example, if a patent (P1) had some IPC codes (IPC1 and IPC2) and another patent (P2) had some IPC codes (IPC1, IPC3), it was said that P1 and P2 were indirectly connected through IPC1. In this case, the new graph had an edge between P1 and P2. Because the new graph had only the patent nodes, there were 8,182 nodes.

Degree centrality is a rate of the number of edges to the potential number of edges. For example, if it is 1, the node is connected to all the other nodes, and if it is 0.5, the node is connected to the half of the other nodes. The degree centrality distribution of the patent projection was visualized in Figure 9.

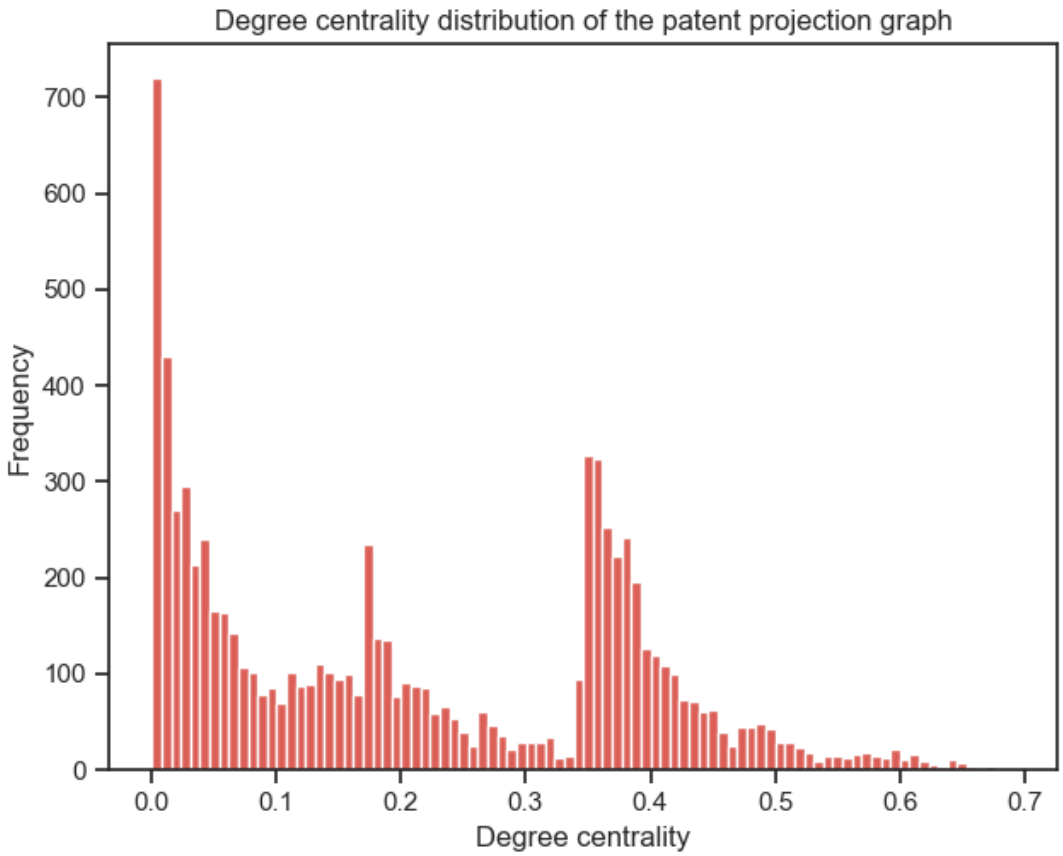


Figure 9. Degree centrality distribution of the patent projection graph

There were three peaks around 0, 0.18, and 0.36. Because the graph had 8,182 nodes, 0.18 meant more than 1,470 connections and it was a lot. That was why the sentence similarity was used to prioritize them.

On the other hand, if a degree centrality is very close to 0, the patent has only a few edges (connections with the other patents). Then, the neighbor patents might or might not be connected to the other patents. If not, the network is closed. There were 46 subgraphs in the graph (one huge subgraph and 45 small subgraphs.) The small subgraphs had only one component, that is, each small subgraph was composed of one patent. They did not have any neighbor patents, and it was impossible to extract the applications from the neighbors. On the other hand, all patents were connected by 'C08L 101/16' actually. Then, in the case, recommended applications would be extracted from patents having the most similar abstract using the sentence similarities.

4.3. Create recommendation systems

Two versions of the recommendation systems were created. The first one accepts one patent and the second one accepts two IPC codes from a user and they recommend up to 10 applications.

4.3.1. Create a recommendation system accepting one patent as input (system 1)

When a user already published a patent in the biodegradable polymer field and wants to know other potential applications for their polymer, this recommendation system would be used. The patent a user inputted should be in the data.

For example, suppose that Company_A published Patent_1 about a biodegradable suture using polyglycolide, and they want to know that their polyglycolide could be used for the other applications. They would input Patent_1 in the system and get 10 potential applications. The input patent could be a competitor's patent.

Two cases had to be recognized to create the system successfully. The first case was that an input patent did not have any neighbor patent. Remember that 45 patents formed individual subgraphs. In this case, the network could not be used to extract similar patents, but the sentence similarity would be. Another case was that an input patent did not have an English or Japanese abstract (10%.) In this case, the extracted neighbor patents by the network could not be prioritized. I decided to show a warning and use the unsorted patent list for the following steps. These two cases could be combined. The handling was shown in Table 3.

Table 3. How to deal with exceptional cases

Neighbor patents	English or Japanese abstract	Sign	Handling
Yes	Yes	-	Extract neighbor patents using the network and prioritize them by the sentence similarity
No	No	Error 2	Exit the system
No	Yes	Warning 1	Extract similar patents in similar order by the sentence similarity
Yes	No	Warning 2	Extract neighbor patents using the network and use the neighbor patents without sorting

The workflow of the recommendation system was shown in Figure 10. When a patent is given by a user, the neighbor patents are extracted. The neighbor patents are sorted by the sentence similarities of the abstracts. When prioritizing the neighbor patents, some patents might not have an English or Japanese abstract. In this case, the average score of the input patent is used as the similarity score. Then, the application IPC codes are extracted from the top patents. If 10 applications are not collected at one cycle, the neighbor patents of the neighbor patents would be extracted from the network, and the new neighbor patents are used to extract more applications.

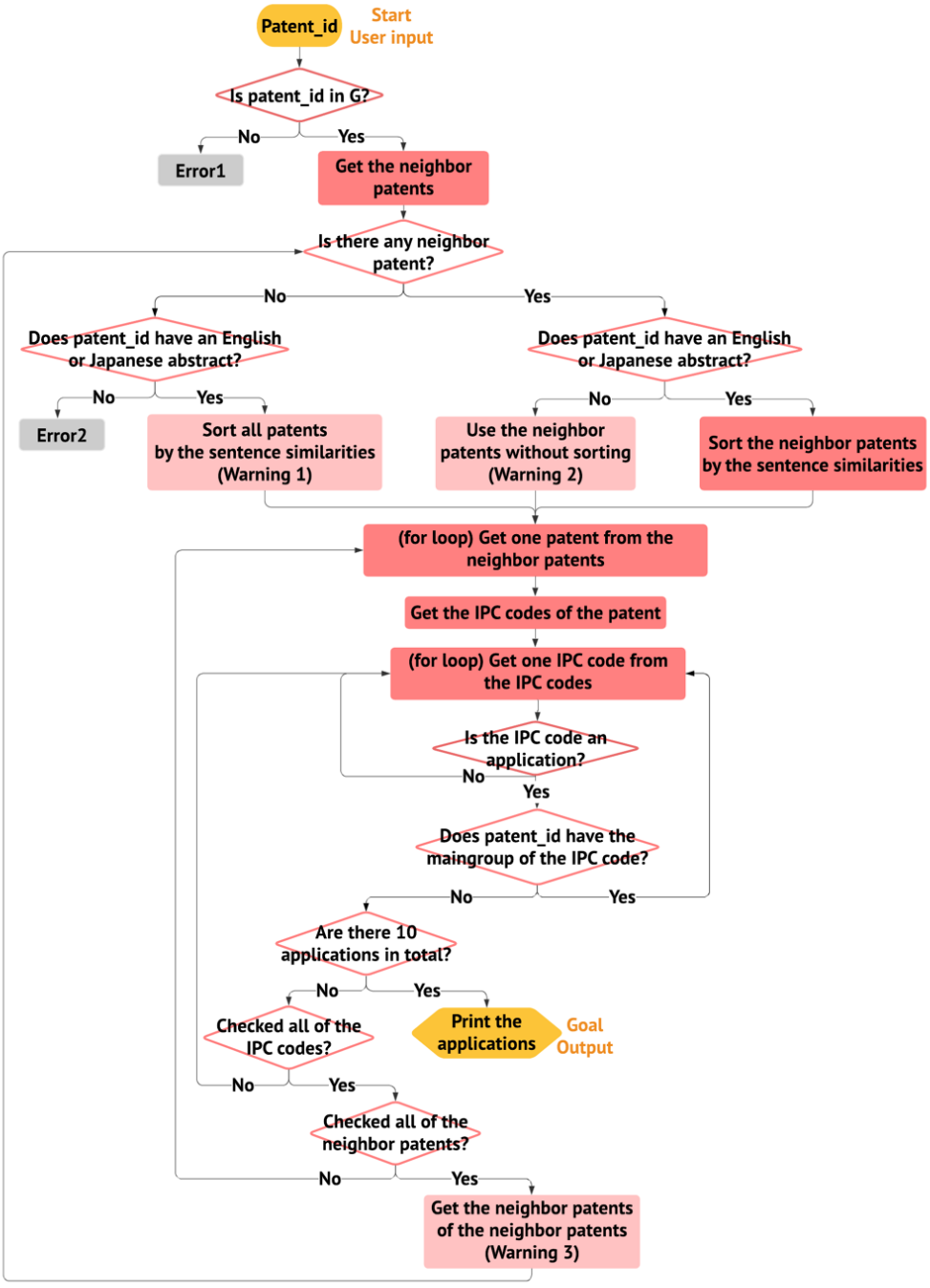


Figure 10. Workflow of the recommendation system accepting one patent as input

As you can see in Figure 10, this system might raise two kinds of errors or three kinds of warnings. Unit tests were prepared for them to confirm if the system raised appropriate messages (errors and/or warnings) in the needed cases.

4.3.2. Create a recommendation system accepting two IPC codes as input (system 2)

When a user does not have any target patent and wants to search for some applications from IPC codes, this second recommendation system would be used. The IPC codes inputted by a user should be in the data.

For example, suppose that Company_B is developing a biodegradable suture using polyglycolide (they do not have the patent yet), and wants to know other potential applications for their polymer. They would search the IPC codes of polyglycolide (C08G 63/06) and a biodegradable suture (A61L 17/06, A61L 17/08, or A61L 17/12 (IPC is a hierarchic structure. I recommend to try several IPC codes that are related to what you want to search).) Then, you input two IPC codes (e.g., C08G 63/06 and A61L 17/08) into the system and would get up to 10 potential applications. The workflow of the second recommendation system was shown in Figure 11.

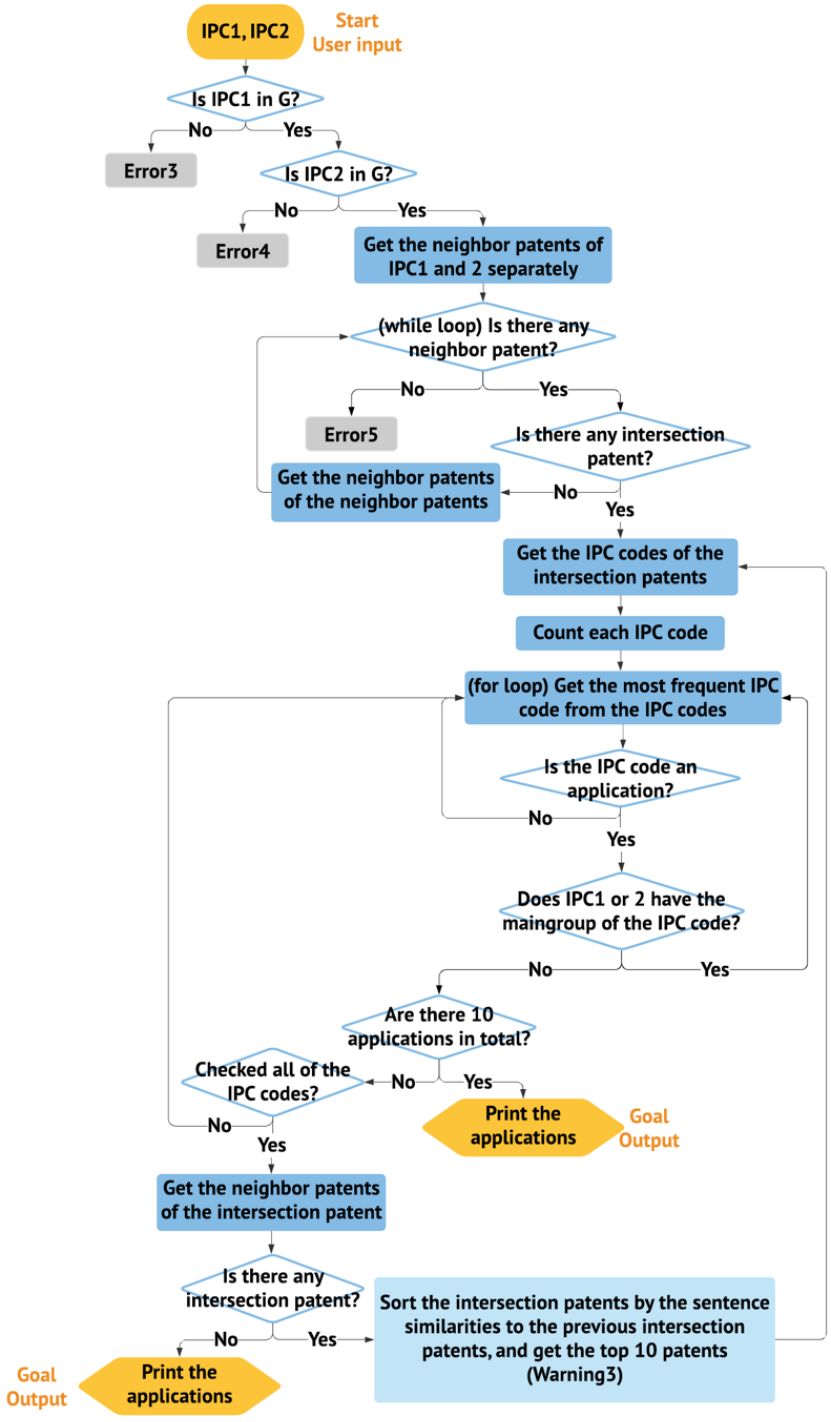


Figure 11. Workflow of the second recommendation system accepting two IPC codes as input

When two IPC codes are given by a user, the neighbor patents of each IPC code are separately searched, and the intersection patents are extracted. This time, user inputs are IPC codes, and the intersection patents cannot be sorted by the similarities. I decided to use a majority voting. All of the IPC codes of the intersection patents are searched and counted, and the IPC codes are checked if it is an application from the most frequent IPC codes to least. When 10 application IPC codes are acquired, the system outputs the result. If 10 applications are not collected at one cycle, the neighbor patents of the neighbor patents are extracted from the network, and the new intersection patents are sorted by the sentence similarities to the previous intersection patents. Then, the top 10 intersection patents are used for the next cycle.

Unit tests were prepared for the function of system 2 as well to confirm if it raised appropriate messages (errors and/or warnings) in specific cases.

5. Recommend applications (using the systems)

The interfaces of the recommendation systems were created. A user can input a patent or two IPC codes through an input function, and the systems return an error, warning(s), and/or the applications. The applications are provided as IPC codes. The examples were shown in the following sections.

5.1. Example 1: use the recommendation system 1

I picked up a patent “JP274783873” as an example for system 1. The patent was about a polylactic acid (PLA) composition having heat-resistance property (Table 4.) According to the patent body, the polymer can be used as a coffee capsule for a coffee service system.

Table 4. A patent “JP274783873”

	Patent_Id	Application_Number	Application_Date	Country	Title	Abstract	IPC	Applicants	Inventors
100	JP274783873	2018097974	22.05.2018	JP	POLYMER COMPOSITION CONTAINING PLLA AND PDLA	PROBLEM TO BE SOLVED: To provide a polymer having high heat resistant shape stability, biodegradability, and based on an organism-derived raw material and a molded article obtained from the polymer. SOLUTION: There is provided a polymer composition containing following components, a. 15 to 70 wt.% of PLLA, b. 0.1 to 15 wt.% of PDLA, c. 5 to 40 wt.% of polyester and d. 5 to 40 wt.% of an organic or inorganic filler, based on total weight of the polymer composition. Such kind of polymer composition can be biodegradated, mainly can contain bio-based carbon and can have high heat resistance. Further such kind of polymer composition can be used in a special method for manufacturing a molded article, a film or a fiber, and the molded article, the film or the fiber can be used as a container for a coffee service system because they are high in heat resistant shape stability. SELECTED DRAWING: Figure 1 COPYRIGHT: (C)2018,JPO&INPIT	C08L 67/04; C08K 3/013; C08L 67/02; C08J 5/00; C08J 5/18; C08L 101/16; D01F 6/62	BIO-TEC BIOLOGISCHE NATURVERPACKUNGEN GMBH & CO KG; バイオーテック ビオローギッシュ ナチューフェアパックンゲン ゲーエムベーハー ウント コンパニ カーゲー	SCHMIDT HARALD; ハラルド シュミット; CHRISTOPH HESS; クリストフ ヘース; WOLFGANG FRIEDEK; ウルフガング フリーデク; BECKMANN RALF; ラルフ ベックマン

Suppose that the company wants to make a new product using their PLA for something else other than food containers. When they execute the system, they would see an interactive window (Figure 12). They would input JP274783873 and get the result (Figure 13.)

```
# Accept input from a user
user_input = input('Input a patent ID (e.g., JP273590701):')

# Input 'JP274783873'

# Run the recommendation system 1
messages, appIpc_refPatents_dict = recommend_app_from_patent(user_input)
df_applications, df_references = make_dataframes(appIpc_refPatents_dict)
show_result(messages, df_applications, df_references, user_input1=user_input)
```

```
Input a patent ID (e.g., JP273590701): JP274783873
```

Figure 12. An interactive window when executing system 1

```
================================================================================================================================
User input: JP274783873 ,

Recommended application codes and the reference patents:
--------------------------------------------------------------------------------------------------------------------------------
  Application IPC Code Reference Patent
0            A61J 3/07      JP274736750
1           A61L 27/00      JP271388842
2            A61F 2/84      JP271388842
3           A61L 17/00      JP271388842
4           A01G 13/02      JP270713333
5            A01G 9/14      JP270713333
6            B27N 5/00       EP13094850
7            B27N 3/02       EP13094850
8            A01G 9/10       EP13094850
9            A61K 6/10       US39070606
--------------------------------------------------------------------------------------------------------------------------------

The reference patents:
--------------------------------------------------------------------------------------------------------------------------------
        Patent_Id Application_Date                                    Title
4691  JP270713333       19.06.2003              ポリ乳酸系重合体組成物、その成形品、および、フィルム
6569   EP13094850       25.05.1998          Biodegradable molding material
3220  JP271388842       30.11.2006  -HYDROXY ACID POLYMER COMPOSITION AN...
439   JP274736750       23.12.2015                              成形部品の生産方法
6753   US39070606       10.03.1997  Cross-linkable or curable polylacton...
--------------------------------------------------------------------------------------------------------------------------------
================================================================================================================================
```

Figure 13. The result output of example 1

The 10 applications were (this page was used to look up the IPC codes.):

0. A61J 3/07 = Capsules or small containers for oral use medicines,
1. A61L 27/00 = Materials for prostheses or for coating prostheses,
2. A61F 2/84 = Devices providing patency to, or preventing collapsing of, tubular structures of the body (e.g., stents),
3. A61L 17/00 = Materials for surgical sutures or ligaturing blood vessels,
4. A01G 13/02 = Protective coverings for plants, devices for laying-out coverings,
5. A01G 9/14 = Greenhouses,
6. B27N 5/00 = Manufacture by dry processes of non-flat articles made from wood (or other lignocellulose) particles

or fibers,

7. B27N 3/02 = Manufacture of substantially flat articles (e.g., boards) from wood particles,
8. A01G 9/10 = Receptacles for seedlings, and
9. A61K 6/10 = Compositions for taking dental impressions.

1 to 3 are medical materials retained in the body using PLA (JP271388842). 4, 5, and 8 are for agriculture using PLA (JP270713333, EP13094850). 6 and 7 are for molding of wood materials using PLA (EP13094850.) 0 and 9 are medical materials using different polymers (polyhydroxyalkanoates, polycaprolactone.) I think 1 to 8 could be manageable using the PLA compound of JP274783873 (the inputted patent), and the heat-resistance property could work well because thermal stability generally improves moldability (although there is another discussion if it is the best to use the material). On the other hand, it would need a big modification to use the PLA compound for 0 or 9 purpose because the expected features are very different. In this case, 8 out of 10 recommended applications were prospective. It would be enough.

5.2. Example 2: use the recommendation system 2

Suppose that my company sells cellulose nanofibers as a material, and wants to develop an end-product using the cellulose nanofibers to increase profitability. I would google some patents using cellulose nanofibers (if you find a patent that is very similar to what you want to do, the patent can be used on system 1), search the IPC codes listed on them here, and get two IPC codes, “C08L 1/00” (cellulose) and “B82Y 30/00” (nanotechnology for materials.) I would input the two IPC codes on system 2 (Figure 14) and get the result (Figure 15.)

```
# Accept input from a user
user_input_1 = input('Input the first IPC code (e.g., A61J 3/07):')
user_input_2 = input('Input the second IPC code (e.g., A61J 3/07):')

# Input 'C08L 1/00'
# Input 'B82Y 30/00'

# Run the recommendation system 2
messages, appIpc_refPatents_dict = recommend_app_from_2ipcs(user_input_1, user_input_2)
df_applications, df_references = make_dataframes(appIpc_refPatents_dict)
show_result(messages, df_applications, df_references, user_input1=user_input_1, user_input2=user_input_2)
```

```
Input the first IPC code (e.g., A61J 3/07):C08L 1/00
Input the second IPC code (e.g., A61J 3/07): B82Y 30/00
```

Figure 14. An interactive window when executing system 2

```
================================================================================================================================
User input: C08L 1/00 , B82Y 30/00

** Rough estimate **
The second-tiers after 3 [warning3]

Recommended application codes and the reference patents:
--------------------------------------------------------------------------------------------------------------------------------
  Application IPC Code Reference Patent
0           A61K 47/38    {JP289824828}
1           A61K 47/36    {JP289824828}
2            A23L 5/00    {JP289824828}
3            A24D 3/10     {EP12753518}
4            D21C 3/00     {EP13533016}
5          C09D 201/00     {EP13533016}
6            D21B 1/36     {EP13533016}
7            A01C 1/06     {EP13533016}
8            D21B 1/04     {EP13533016}
9          C09D 101/02     {EP13533016}
--------------------------------------------------------------------------------------------------------------------------------

The reference patents:
--------------------------------------------------------------------------------------------------------------------------------
        Patent_Id Application_Date                                  Title
6984   EP12753518       01.12.1995  Cellulose ester compositions and sha...
6545   EP13533016       24.07.1998  CELLULOSE FIBER BASED COMPOSITIONS A...
48    JP289824828       05.12.2018                  セルロースナノファイバー及び澱粉を含む組成物
--------------------------------------------------------------------------------------------------------------------------------
================================================================================================================================
```

Figure 15. The result output of example 2

The result showed warning 3. This warning means the applications from No.3 to 9 might not be as useful as No.0 to 2 because No.0 to 2 were extracted at the first cycle of the system (from the closest patent to both IPC codes) but No.3 to 9 were from the second cycle.

The 10 applications were:

0. A61K 47/38 = Medicinal preparations using cellulose (as the non-active ingredient (e.g., carriers or inert additives), or targeting or modifying agent chemically bound to the active ingredient),
1. A61K 47/36 = Medicinal preparations using polysaccharides (as the non-active ingredient (e.g., carriers or inert additives), or targeting or modifying agents chemically bound to the active ingredient),
2. A23L 5/00 = Preparation or treatment of foods,
3. A24D 3/10 = Tobacco smoke filters (e.g., filter-tips, filtering inserts), filters specially adapted for simulated smoking devices, or mouthpieces for cigars or cigarettes using cellulose,
4. D21C 3/00 = Pulping cellulose-containing materials,
5. C09D 201/00 = Coating compositions based on unspecified macromolecular compounds,
6. D21B 1/36 = Fibrous raw materials or their mechanical treatment by dividing raw materials into small particles (e.g., fibers) by defibrating by explosive disintegration by sudden pressure reduction
7. A01C 1/06 = Seed coating or dressing,
8. D21B 1/04 = Fibrous raw materials or their mechanical treatment by dividing raw materials into small particles (e.g., fibers), and

9. C09D 101/02 = Coating compositions based on cellulose

All of the three reference patents were about cellulose fibers. The recommended applications covered a broad range (medicinal carriers, preparation of foods, cigarette filters, seed coating, and coating compositions.) Because this recommendation system provides not only the applications but also the reference patents, a user can compare their materials with the competitor's materials by reading the patents and think about their superiority. Moreover, using the reference patents on system 1 would give additional information. Try “JP289824828” on system 1 (Figure 16.)

```
==========================================================================================================================
User input: JP289824828 ,

Recommended application codes and the reference patents:
--------------------------------------------------------------------------------------------------------------------------
  Application IPC Code Reference Patent
0            A61K 9/70      JP274091683
1           A61L 27/00      JP274091683
2            A61K 9/06      JP274091683
3            D21C 3/00       EP13533016
4          C09D 201/00       EP13533016
5            D21B 1/36       EP13533016
6            A01C 1/06       EP13533016
7            D21B 1/04       EP13533016
8          C09D 101/02       EP13533016
9            A23L 1/00       EP13533016
--------------------------------------------------------------------------------------------------------------------------

The reference patents:
--------------------------------------------------------------------------------------------------------------------------
        Patent_Id Application_Date                                    Title
6545   EP13533016       24.07.1998  CELLULOSE FIBER BASED COMPOSITIONS A...
534   JP274091683       08.05.2015  BIODEGRADABLE CELLULOSE NANOFIBER MI...
--------------------------------------------------------------------------------------------------------------------------
==========================================================================================================================
```

Figure 16. The result output of JP289824828 on system 1

The system provided one new reference patent about cellulose nanofibers and four new applications. If you try the other reference patents on system 1, you will get more information in this field and know about who are the competitors, what are their technologies, and what are their scopes of patent claims.

6. Summary

Two recommendation systems were created for biodegradable polymers using patent information. One system accepts one patent and another accepts two IPC codes as input from a user, and they provide 10 potential applications and the reference patents. The recommendation systems use two features to select potential applications: a network and sentence similarity. The network is composed of 8,182 patents and 4,344 IPC codes as nodes and 46,012 edges that represent the connections between patents and IPC codes. The network is used to extract the neighbor patents of an input patent (or two IPC codes.) Sentence similarities of patent abstracts were calculated by cosine similarity. The similarities are used to prioritize the neighbor patents so that the system can pick up the most similar patents and therefore the most prospective applications.

The recommendation systems recommended useful applications on the tests. The recommended applications were

extracted from reference patents sharing some features (e.g., a polymer kind, form, function) with the input information. I am sure that these recommendation systems would help researchers think about how to apply their wonderful polymers.

Last words:

I was a polymer scientist for more than 10 years. I created what I had wanted at the time. I truly hope these recommendation systems help some researchers and their polymers and products will blossom to make people's lives better.