# Capstone project 2: Milestone Report 1

Project Title:
Application Recommendation System for Biodegradable Polymer


## < Abstract >

The goal of this project is to make an application recommendation system that gives some hints about potential applications for a given polymer by a user. A biodegradable polymer has been chosen as a polymer kind for this project. The data were from the database of the World Intellectual Property Organization (WIPO). Data has been obtained, cleaned, and wrangled. Also, the exploratory analysis has been performed.


## < 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.

Here, I would like to provide an application recommendation system that gives some hints about potential applications for a given polymer by a user. A biodegradable polymer has been 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 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 are 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 would provide 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 have 8,182 patent information, and the columns contain 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

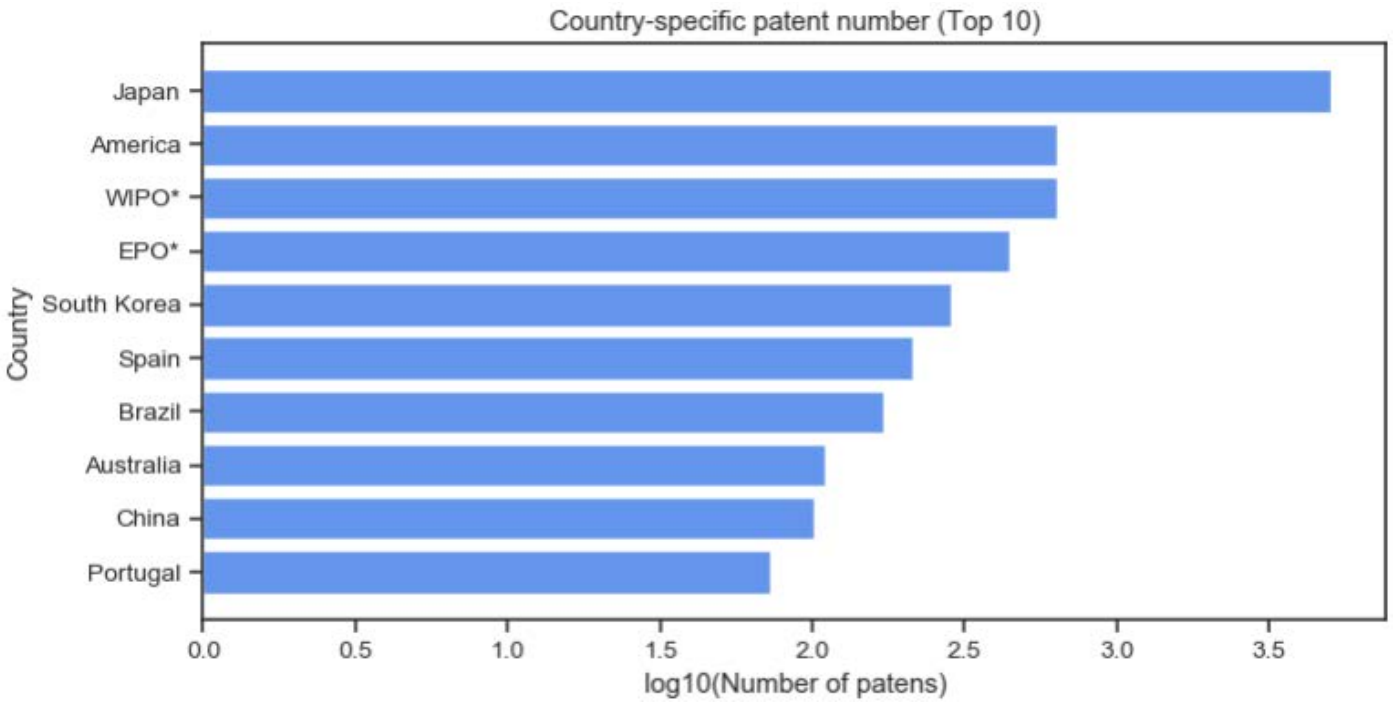
Each IPC code represents a material, application, or technology to manufacture it.

### 2.2. Countries patents were filed in

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) patents in the data were registered in.




* 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 will 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 has a column called 'IPC'. This column shows the IPC codes the patent is categorized in. Here, the frequency of each IPC code was counted, and 4,344 kinds of IPC codes were included. Figure 2 showed the summary statistics.

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

| Min | 25% | 50% | 75% | Max | Mean | SD |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 6 | 2830 | 10.6 | 60.4 |

1,665 (38%) IPC codes appeared only once in the dataset. It means they are not 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 have decided to keep them. On the other hand, 6 IPC codes appeared more than 500 times (Figure 2). This means once we reach one of the IPC codes, there are more than 500 connections. To prioritize the applications, using the abstracts would be a key.

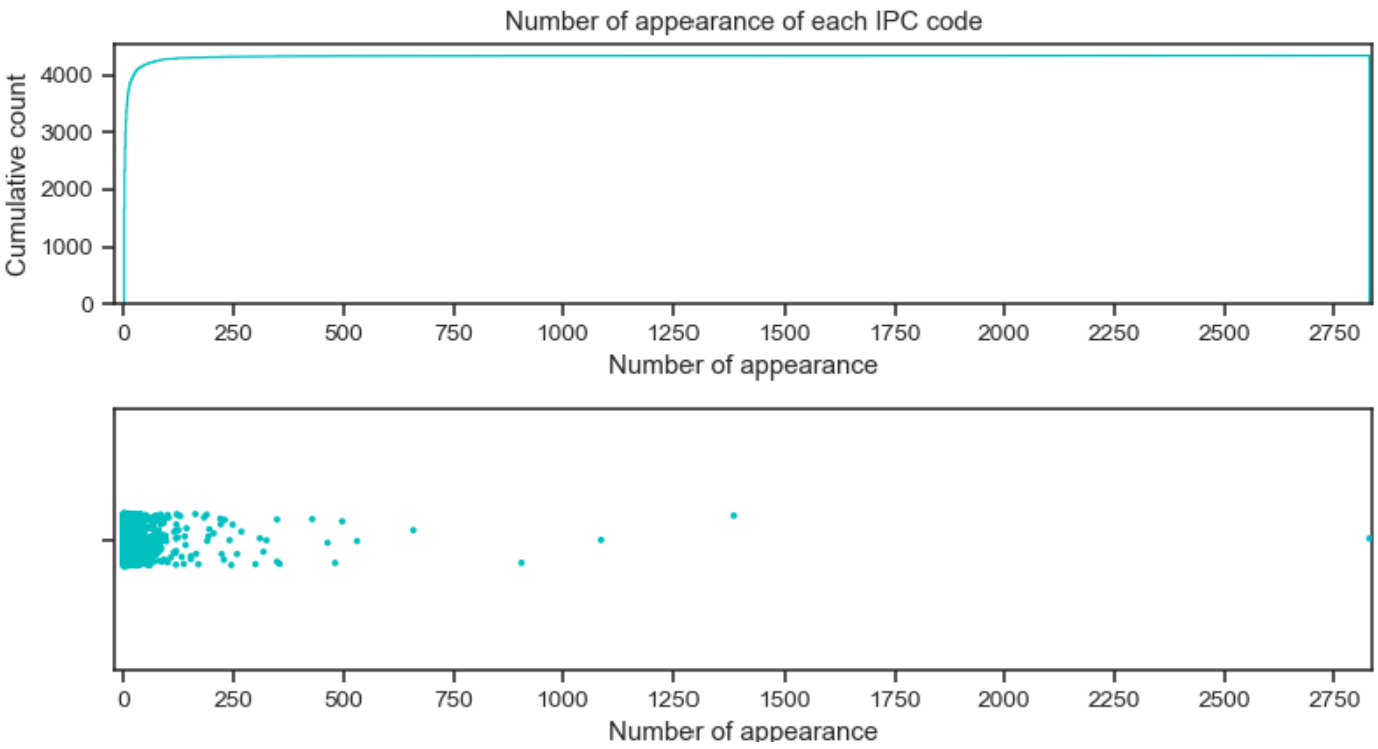



**Figure 2. Frequency of each IPC code**

### 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.

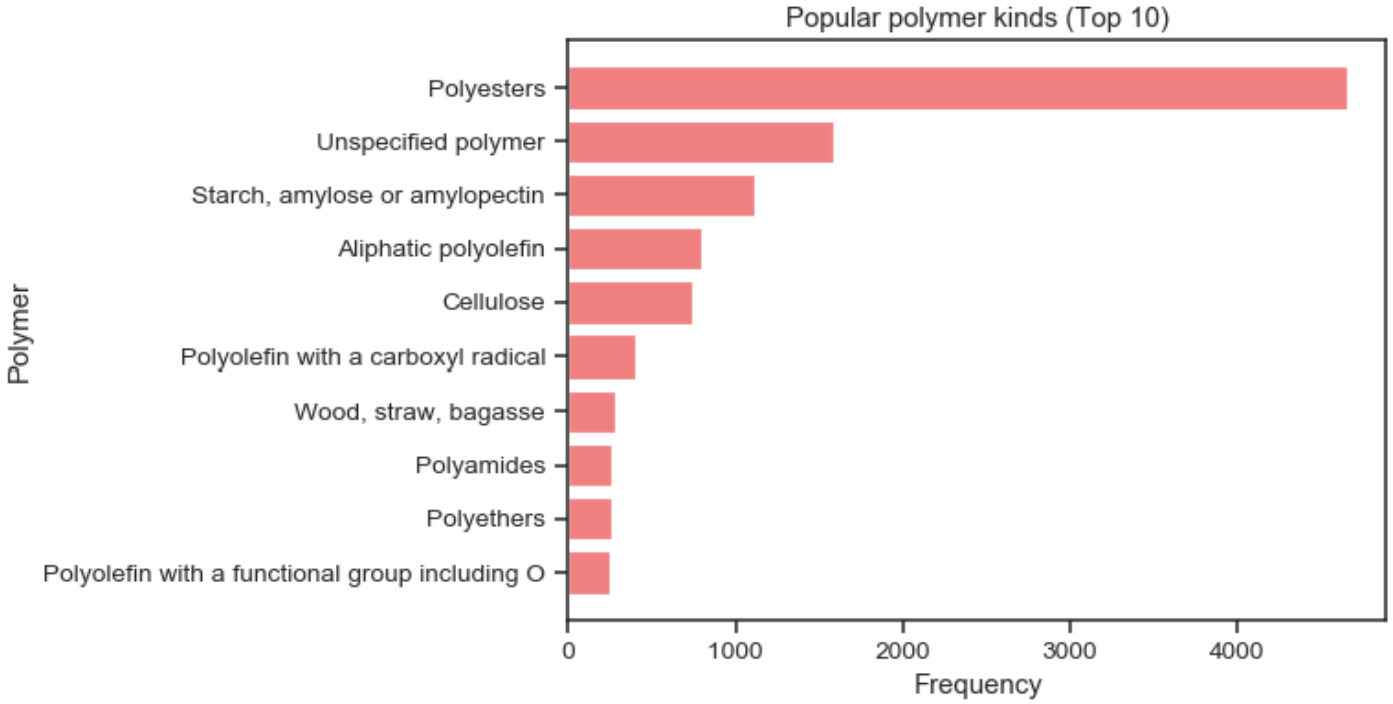



**Figure 3. Popular polymer kinds (Top 10)**

There were 49 kinds of polymers in the data, and polyesters were the most popular (figure 3).

### 2.5. Popular applications

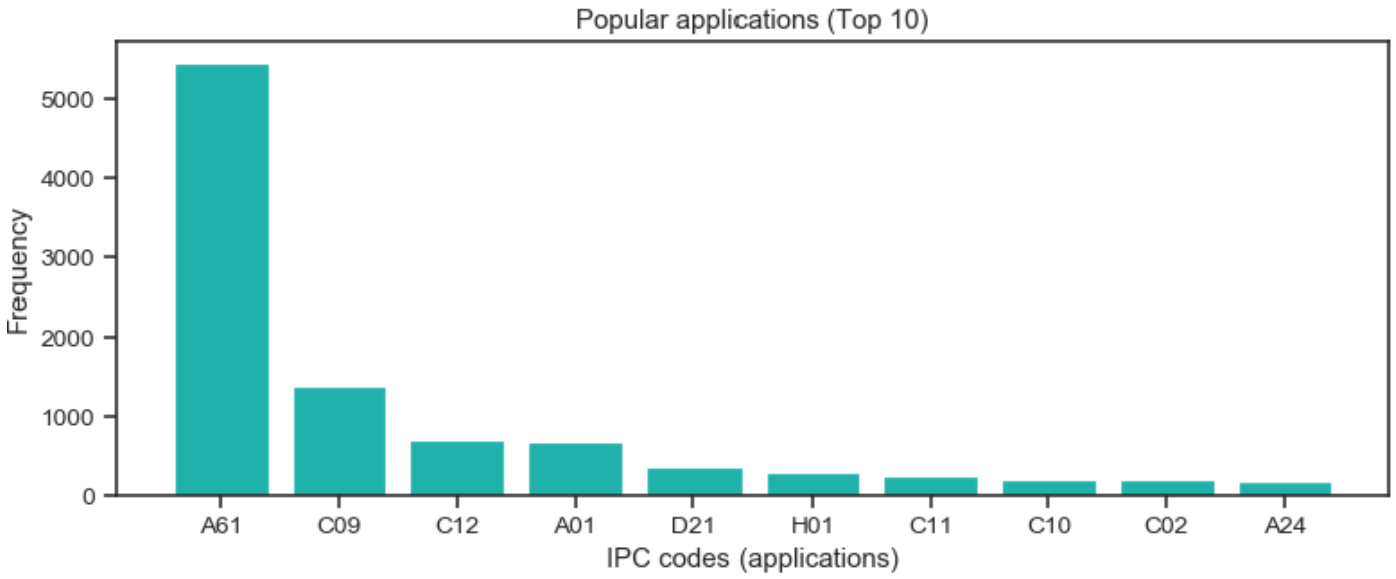
First, I needed to decide what IPC codes were treated as applications. I carefully checked the classes and extracted the following groups as applications (Table 2). These categories are going to be used as IPC codes of applications to make a model. As a result, applications recommended by the model will be 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 |

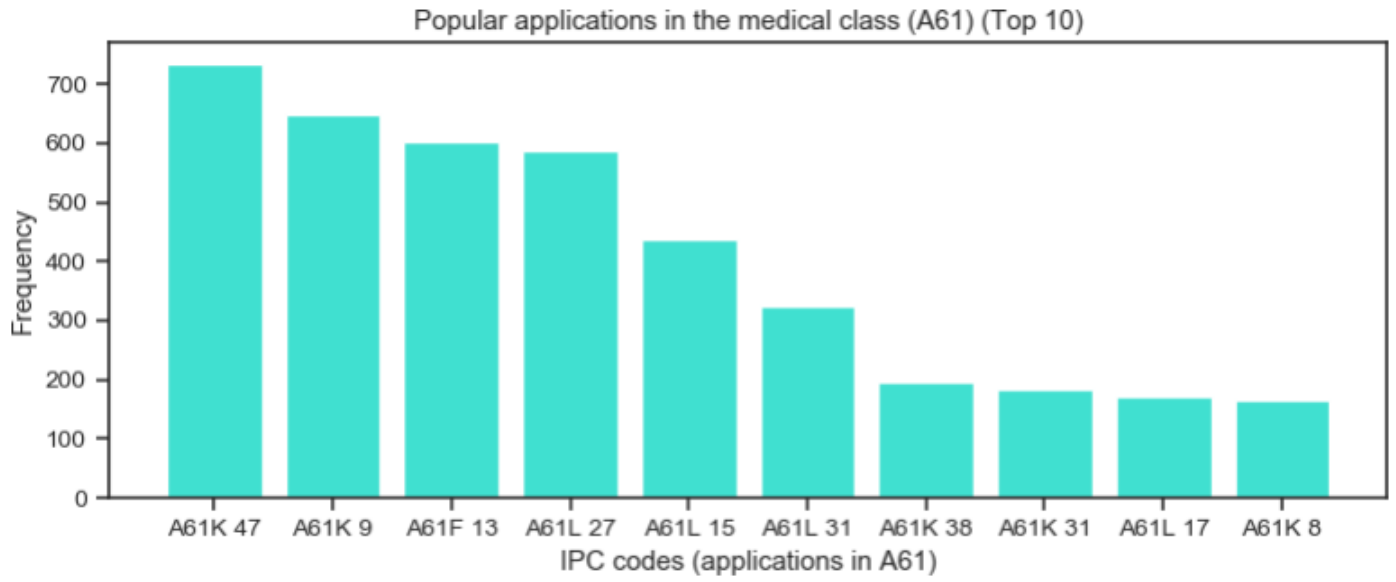
Then, each of the above IPC codes was counted at a class level, and the most frequent 10 applications were shown in Figure 4.




A61: MEDICAL OR VETERINARY SCIENCE, HYGIENE
C09: DYES; PAINTS, POLISHES, NATURAL RESINS, ADHESIVES
C12: BIOCHEMISTRY, BEER, SPIRITS, WINE, VINEGAR, 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, FATTY ACIDS THEREFROM, DETERGENTS, CANDLES
C10: PETROLEUM, GAS OR COKE INDUSTRIES, TECHNICAL GASES CONTAINING CARBON MONOXIDE, FUELS, LUBRICANTS, PEAT
C02: TREATMENT OF WATER, WASTE WATER, SEWAGE, OR SLUDGE
A24: TOBACCO, CIGARS, CIGARETTES, SIMULATED SMOKING DEVICES, SMOKERS' REQUISITES

**Figure 4. 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 mainclasses 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 5. 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 includes at least Spanish, Chinese, English, and Japanese. I'm going to use an abstract to calculate the similarities between patents. It's important to know what languages are 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 as is 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. The languages of 165 (2%) patents were sure less than 70%. Because the rate was low, they were just removed from the language variety plot (Figure 6).

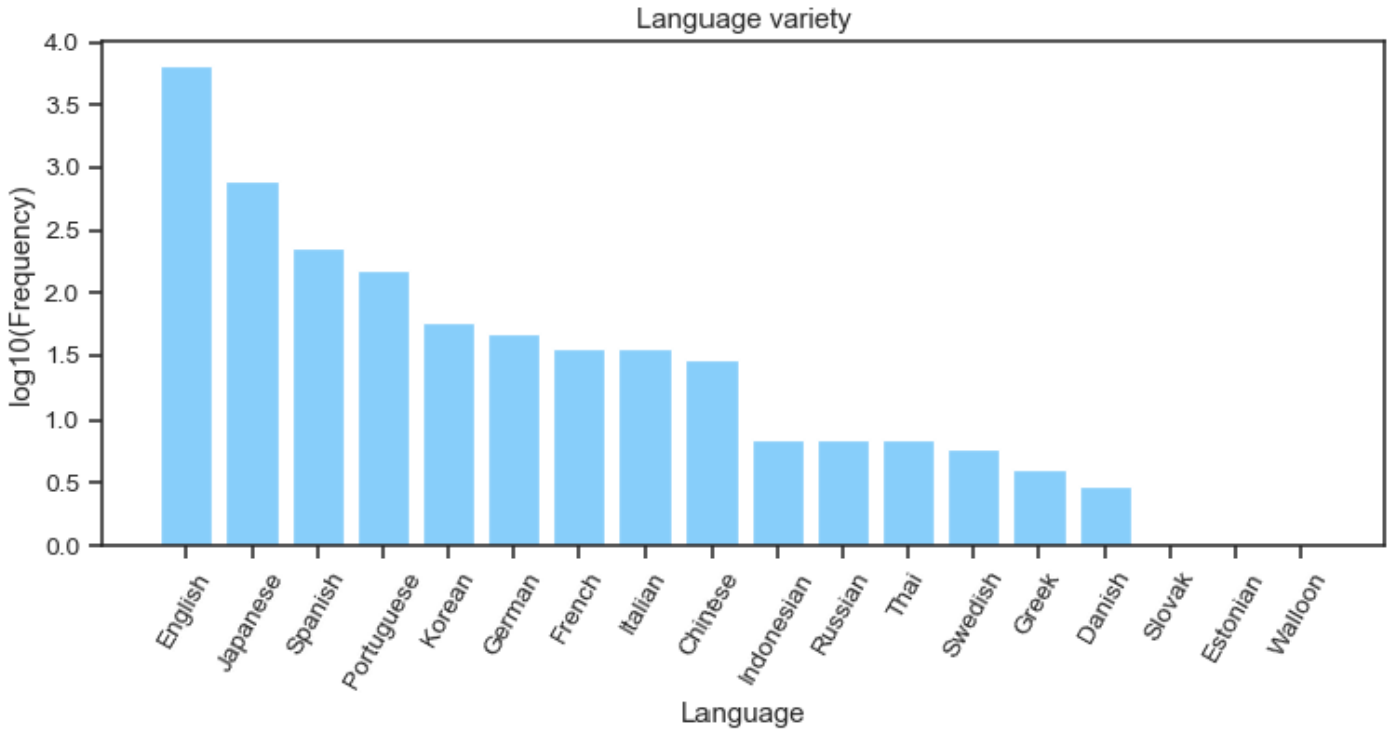



**Figure 6. Language variety**

18 kinds of languages were detected. 81.7% (6687/8182) of the abstracts is written in English. 9.8% (799/8182) is written in Japanese. The abstracts in English and Japanese will be used to calculate the similarity of patents. On the other hand, the IPC codes can be used to analyze the relationships even if a patent is 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. I deal with the 'Abstract' column to prepare for calculating the similarities between patents.

### 3.1. Dealing 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. I am planning to use the five columns out of the six columns other than the ‘Application_Number’ column for this project. 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 going to be used as metadata for a patent node. In this case, there is no problem to have any missing values because the metadata will be used to provide additional information about a patent and not for the essential structure or function of the recommendation system.

### 3.2. Preprocessing of 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 is developed by Google. It is free, but there is a restriction that is 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. The meanings were mostly fine.

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, 14,489 unique tokens remained. The top 100 frequent words were shown in Figure 7. 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.

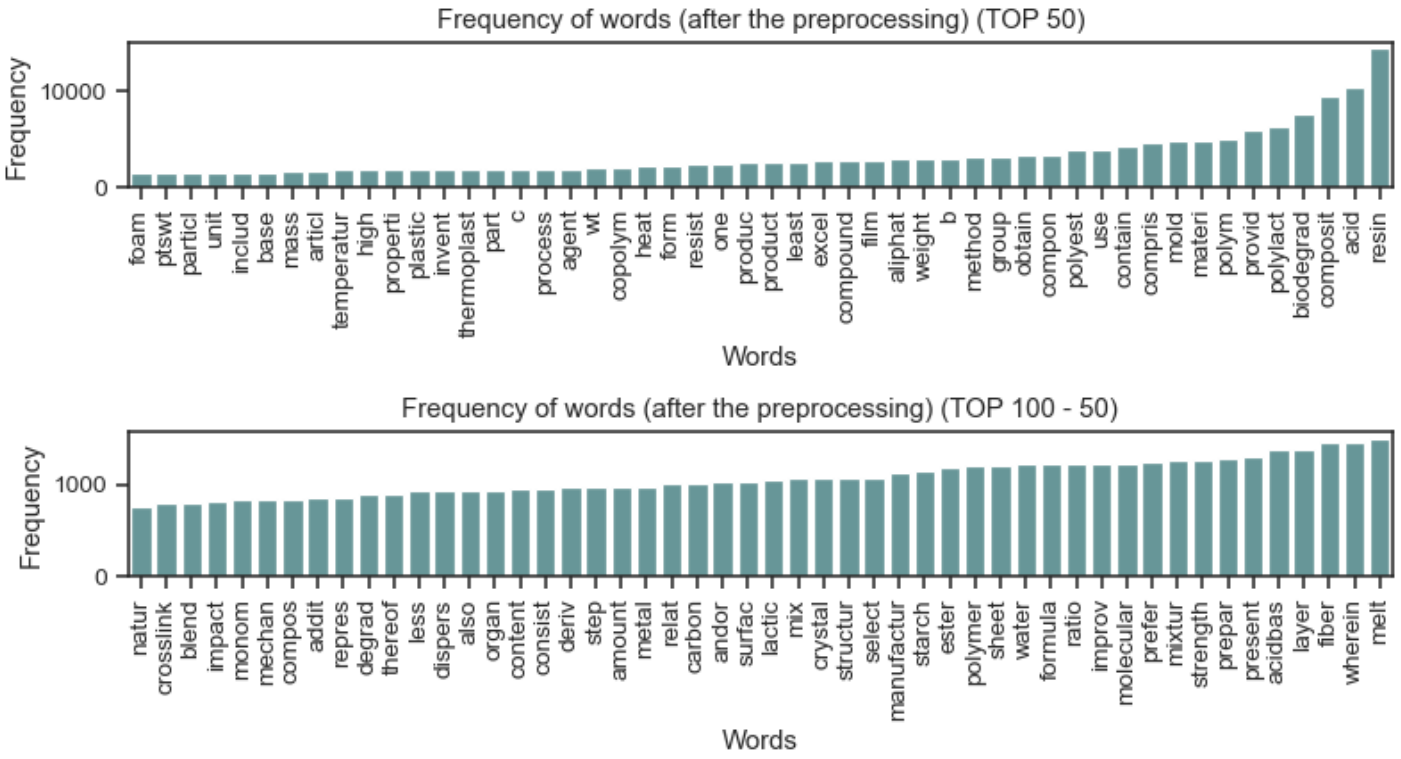



**Figure 7. Frequency of words (TOP 100)**

## 4. Next steps

The IPC codes are going to be used to build the network. A patent will be one kind of node, and an IPC code will be another kind of node. A patent-node will connect with some IPC-code-nodes (the number of connections depends on the number of IPC codes the patent has). Because the system will have a lot of patents, the connections become a network form; a patent shares some code-nodes with other patents. This network will represent the relationships between patents and IPC codes.

Next, the abstracts will be used to calculate the similarities between patents. The information will be stored, and loaded when the recommendation system needs to prioritize potential applications.

On the third step, a system extracting IPC codes close to a given patent (or IPC codes) will be created. When this system is given a patent (or IPC codes) by a user, this system will extract the nearest patents to the given one using the network. Then, the nearest patents will be ordered by the similarities of their abstracts. Finally, the IPC codes, which represent applications (Table 2), of some most similar patents will be provided to the user as the potential applications.