Project 1

Background

In the knowledge-based and digital economy, the effective protection and management of Intellectual Property rights, such as patents, is essential. Patent applications are techno-legally complex and require significant experience and expertise in order to assign suitable classifications and determine its patentability during the examination process. WIPO, including other IP Offices, are exploring the use of AI/ML methods to assist practitioners (Examiners/Patent Agents) in several usecases such as automatic patent classifications, determining semantic similarity between key phrases from patent documents within the patent's contexts[1] etc.

Introduction to the datasets

I have selected the DWPI (Derwent World Patents Index) dataset containing e-waste technologies related patent applications and grants from world's patent issuing authorities. This dataset is the basis for the WIPO patent landscape report with an objective to undertake landscaping analysis of patenting and innovation activity related to e-waste technologies. The attributes of the dataset (8867x91) are as follows:

- Derwent Accession Number: This is a unique identification number assigned by Derwent that consists of the year of publication and six digit serial number.
- Basic Patent Number: This is international publication number and consists of two-letter code (such as WO"), followed by a four-digit indication of the year of publication and a serial number consisting of 6 alphanumeric digits. However, there are variations in conventions across various offices.
- DWPI Title: This provides editorially curated titles summarising the scope, use and novelty of the invention.
- Family Member Countries and Family Breadth: Each document represents a single invention or 'Patent Family'. Patent Family includes the first publication of an invention (basic) plus later published patents (equivalents) related to the invention. Family breadth includes the number of patent authorities/offices/countries where the invention has been published.
- Priority Year and Priority Country: This is the earliest filing date within a family of patent applications.
- Entity Tiers: The entities in the patent dataset are separated into four tiers (Tier-1, Tier-2, Tier-3, Tier-4) based on the number of patent families they each have.
    - Tier-1 (40 or more inventions)

[1] Google Patent Phrase Similarity, https://research.google/resources/datasets/google-patent-phrase-similarity/

    - Tier-2 (5-39 inventions)
    - Tier-3 (less than 5 inventions)
    - Tier-4 (Assigned to individuals)
- Family Contains Grant: This includes (Yes/No) whether the application has been granted.
- Technical Categories (High Level), Technical Categorisation and Technical Sub-categorisation: Each application is assigned classification w.r.t. subject matter categories. The broad thematic categories here are:
    - E-waste sources- This include devices being processed such as displays, medical devices etc. and individual components across multiple products/streams such as circuit boards, batteries
    - Processing methodologies- This include approaches such as waste logistics, chemical separation, disassembly, decontamination etc.
    - Material Recovery- This include specific recovered items such as metal recovery, plastic recycling

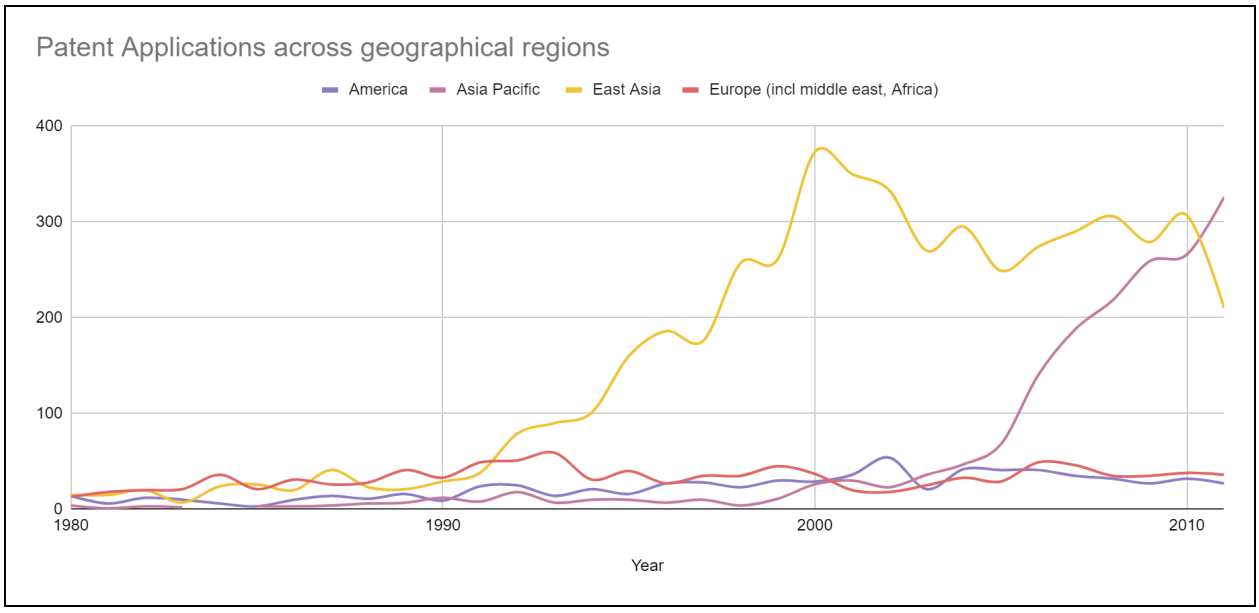
  These technical categories have been created using the existing patent classifications (such as IPC- International Patent Classification, CPC- Cooperative Patent Classification, Derwent Manual Codes and Classes) and text mining of claims, abstract.
- Count of Citations- This includes the number of citations.
- Standardised Assignee, Inventors and Sector- This includes the name of corporations and inventor's name, if not filed by individuals.

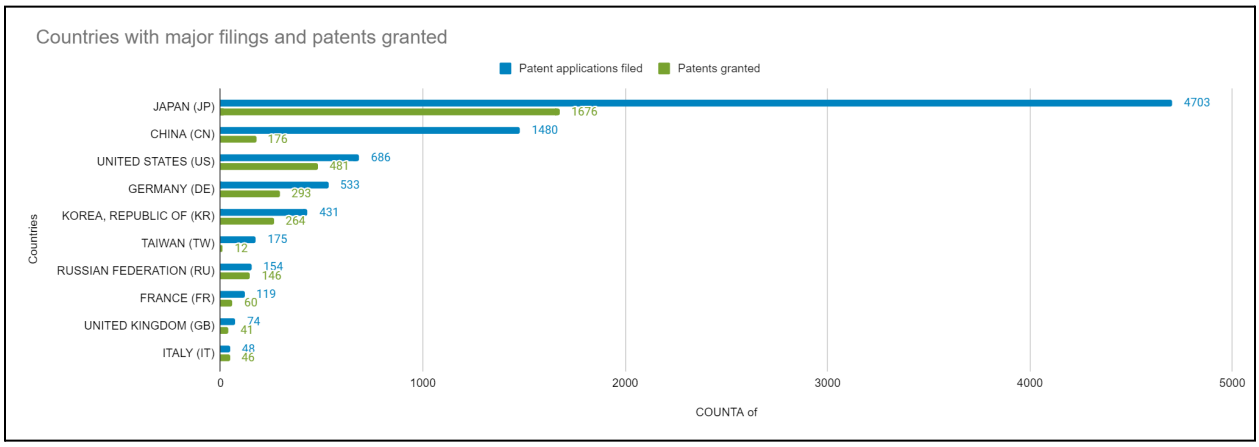
Preliminary exploration of the datasets and visualisations:

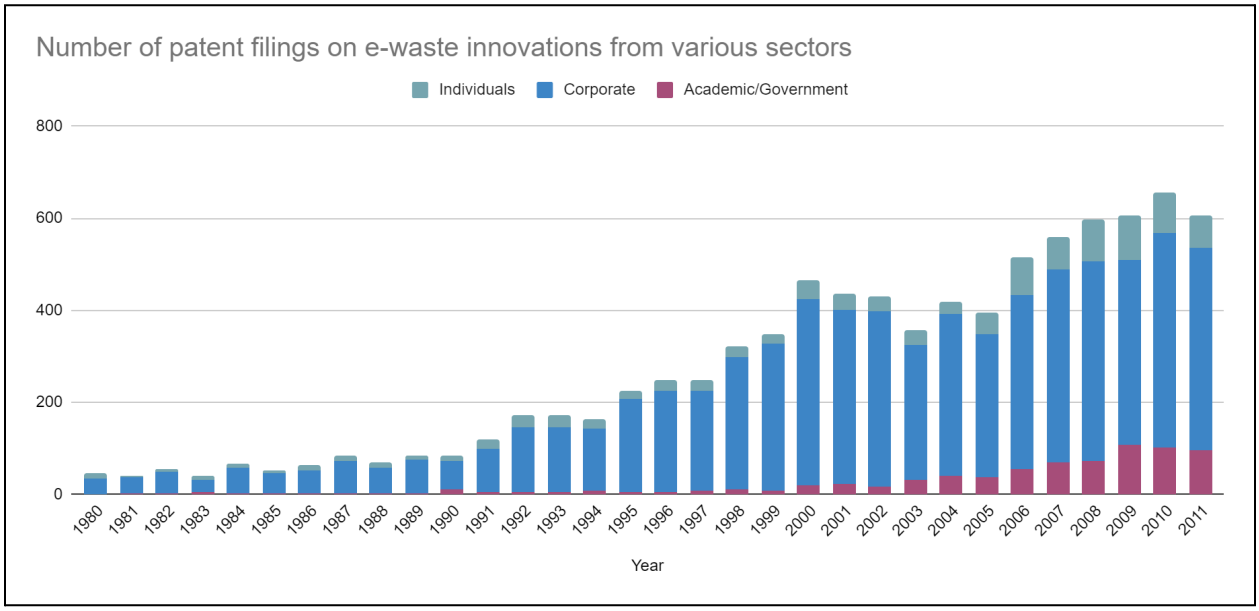
An analysis of patent literature can reveal trends, technical details and relationships and shall be helpful to guide investment decisions and policy regulations.

- *Innovation Regions and Countries*: The Patent application filings in e-waste technologies across 4 geographical regions (mapping Family Member Countries code[2]) have been analysed to sport trends/patterns. While countries in Europe and America have seen a steady e-waste patent filings in the last 30+ years, there has been a recent surge in e-waste patent filings in countries in East Asia and Asia Pacific. Top countries/patent offices with the majority of patent filings in e-waste technologies have been identified. Similarly, the number of patents filings from individuals and the academic/government sector have been looked into as well, indictating the increase from government sectors and individuals.

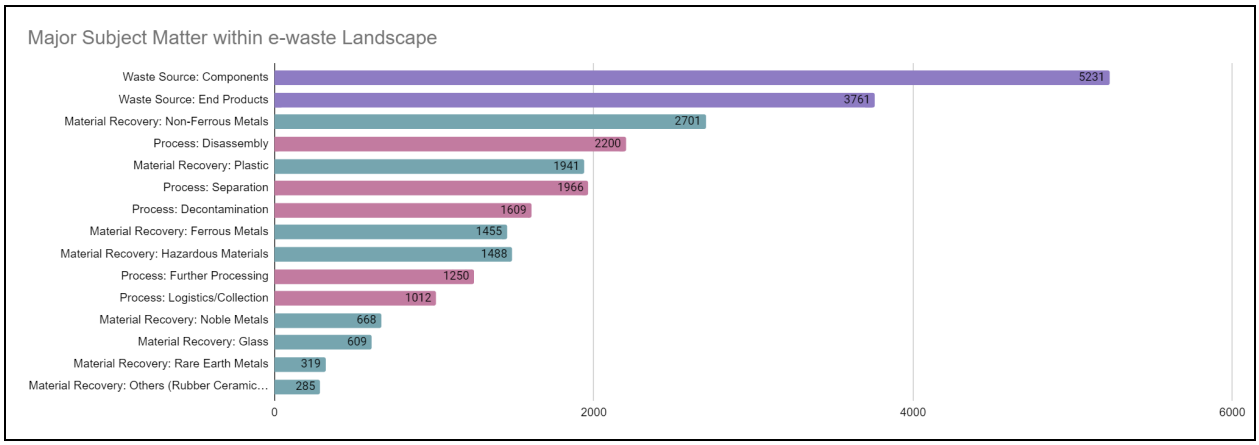
[2] https://www.uspto.gov/patents/apply/applying-online/country-codes-wipo-st3-table

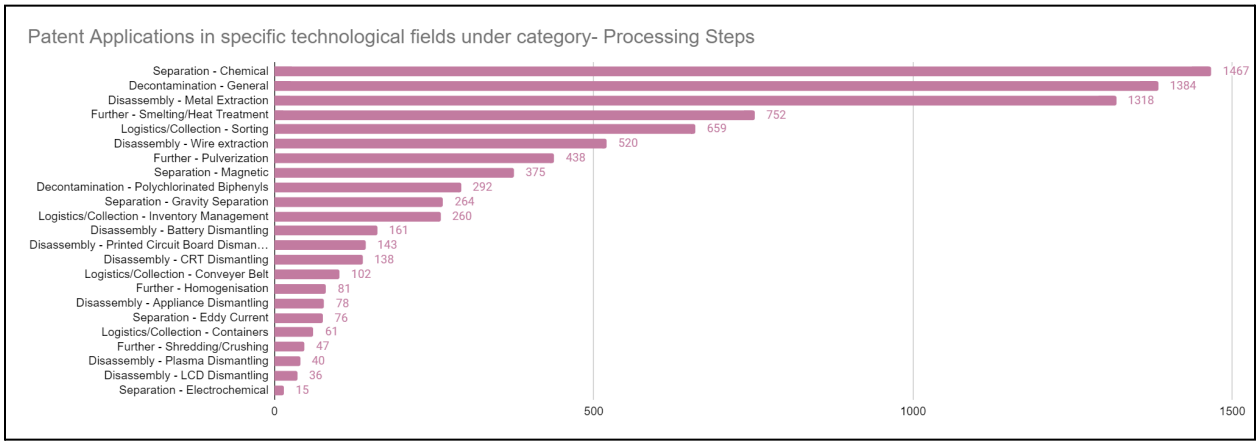

Patent Applications across geographical regions
America
Asia Pacific
East Asia
Europe (incl middle east, Africa)
400
300
200
100
0
1980
1990
2000
2010
Year

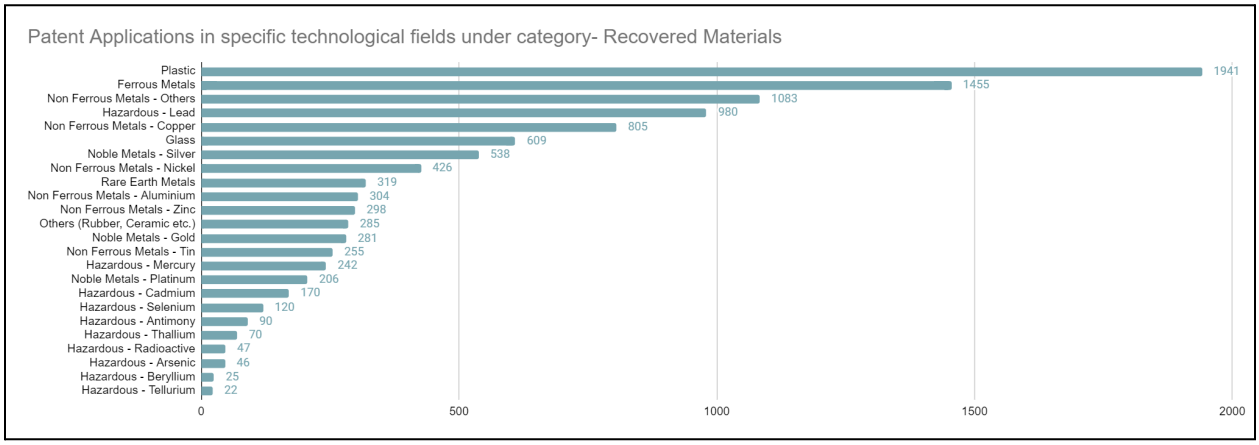

Countries with major filings and patents granted
Patent applications filed
Patents granted
JAPAN (JP)
1676
4703
CHINA (CN)
176
1480
UNITED STATES (US)
481
686
GERMANY (DE)
293
533
KOREA, REPUBLIC OF (KR)
264
431
TAIWAN (TW)
12
175
RUSSIAN FEDERATION (RU)
146
154
FRANCE (FR)
60
119
UNITED KINGDOM (GB)
41
74
ITALY (IT)
46
48
Countries
0
1000
2000
3000
4000
5000
COUNTA of


Number of patent filings on e-waste innovations from various sectors
Individuals
Corporate
Academic/Government
800
600
400
200
0
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Year

- *Innovation Categories and Subject Areas*: The preliminary analysis has been conducted to list subject areas within the e-waste technologies landscape that are witnessing innovations. This has been further done at the level of Technical Categories (Waste Source, Process, Material Recovery). A majority of e-wastepatent applications are addressing Waste Source category. This analysis helps explore technological trends in e-waste technologies.

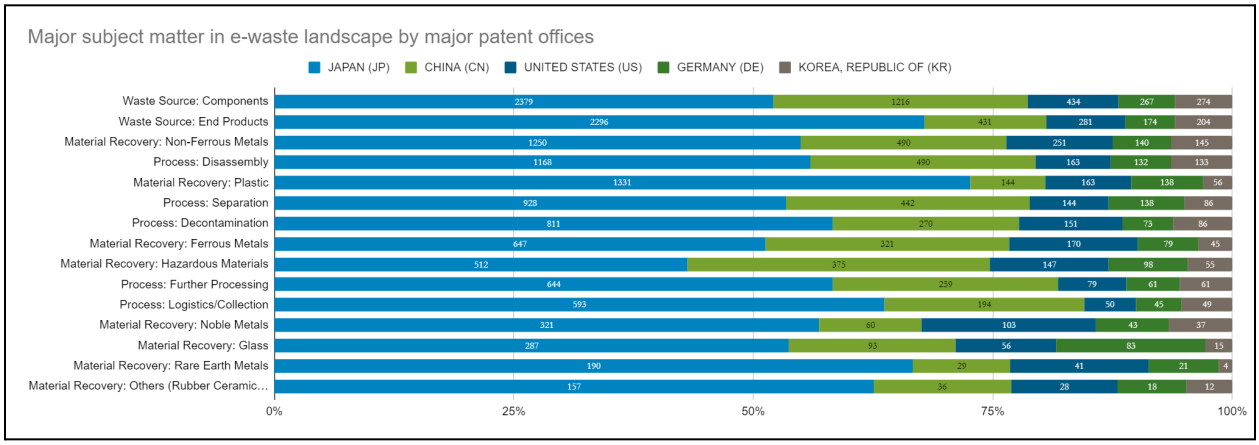






- *Innovation Categories and Countries:* In top 5 leading patent countries, the patent applications have been analysed subject categories wise. While Japan is leading in

most of the subject categories, China is gaining grounds in Material Recovery (Hazardous, ferrous and non-ferrous metals).



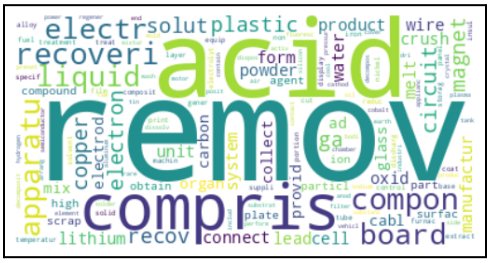
<u>Advanced analysis:</u>

The objective is general exploration of the core inventive aspects of the patent application by extracting information (title/abstract etc.). This is done using the natural language toolkit (nltk) and topic modeling framework (gensim). The method is the following:

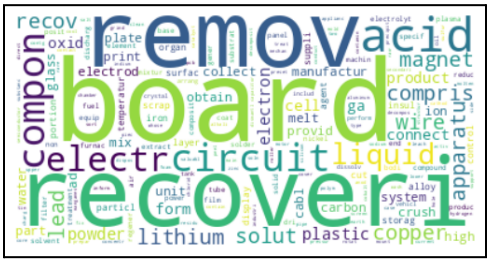
- Pre-processing the text data (titles) in Patents applications by
    - Tokenising the “DWPI Title”, Removing the “stop words” and punctuations etc, Stemming the words
    - Building a term dictionary that associate each stemmed word with a particular index number and converting documents-term matrix using TF-IDF algorithm
- Classifying the documents into “Topic Areas” using the LSI model or LDA model. Here, I iterated the model for a number of topic areas.
- Visualising the word cloud for all applications and by topic areas. THis has been done by removing certain words (such as 15 most frequent words).

The similar analysis has been done by subsetting datasets based on 3 technical categories (e-waste products, Processes; Recovered Materials).

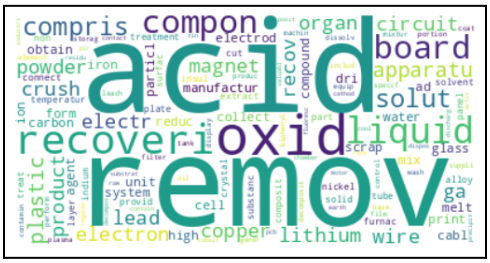
The word cloud analysis for the technical categories indicate the relevant words (such as electrod, circuit board, wire etc. for Technical Category-1(e-waste products); liquid, oxidation, solution etc. for Technical Category-2 (Process); copper, lead, plastic for Technical Category-3 (Recovered Materials). The word cloud analysis also confirms that there is overlap across these categories as some of the words from Technical Category#3 (Recovered Materials) are also frequent in Technical Category#2 (Process).



All Applications



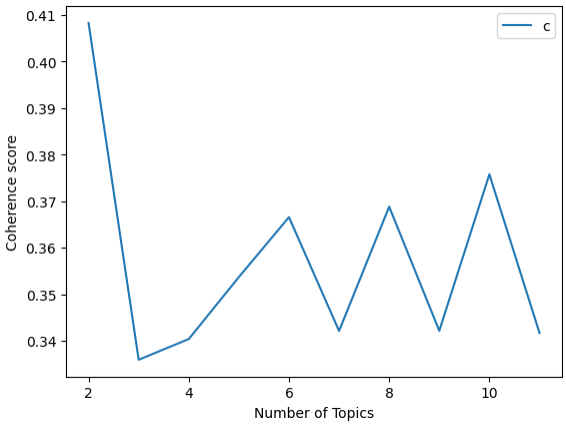
Technical Categorisation01: e-waste products



Technical Categorisation02: Processes

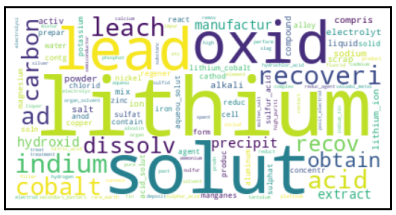
Technical Categorisation03: Recovered Material

I've used LSI model for topic modelling. In order to determine the number of topics in the data, the 'Coherence Score' metric has been used to determine the relative distance between words within a topic. Models were trained for a finite number of topics and the 'Coherence Scores' were computed for comparison. The number of topics=10 is finally selected. However, Nearly 75% of documents are in Topic Area#1. This has to be corrected.

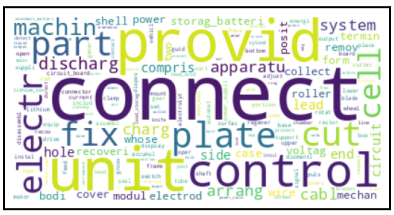


The word clouds for Topic Areas have been visualised excluding the first one. Word Clouds can help visualize top words in a topic for further analysis and identifying the subject area. As it seems, these topic areas may further been categorised into the Technical Categories:

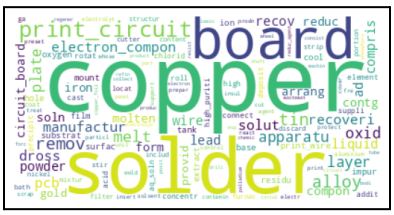
- Technical Category-01 (e-waste products): Topic6, Topic 5
- Technical Category-02 (Products): Topic8, Topic2, Topic3, Topic 4
- Technical Category-01 (Recovered Materials): Topic9; Topic7; Topic1



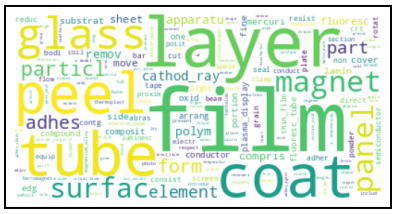
Topic1 (n=299)



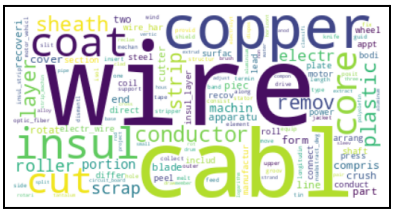
Topic2 (n=884)



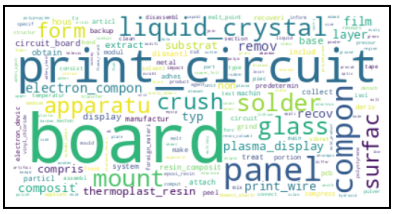
Topic3 (n=166)



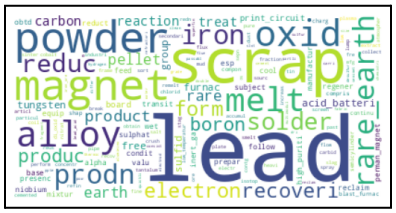
Topic4 (n=79)



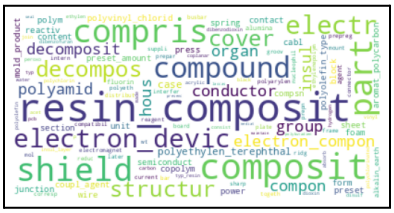
Topic5 (n=264)



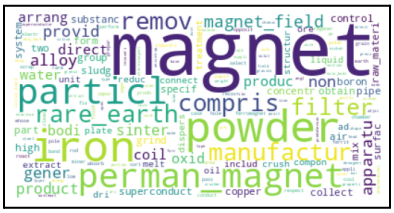
Topic6 (n=170)



Topic7 (n=63)



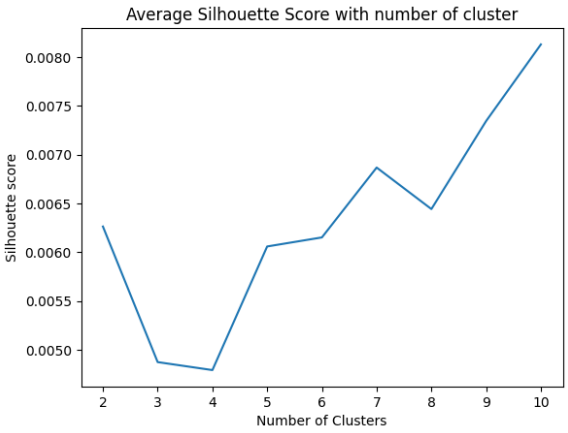
Topic8 (n=22)



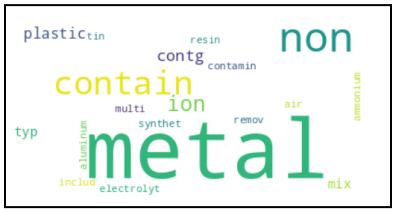
Topic9 (n=248)

Within these documents belonging to the Topic Area, the cosine similarity may be computed to pick the outlier ones. This may be pursued as further work.

K-means algorithm also has been used to divide into documents clusters with n=10. The quality of clustering has been analysed again using word clouds.



Cluster1



Cluster2

Conclusions

The sheer increase in the number of patents over years makes a case to facilitate patent classification using AI/ML methods to assist the Patents Office efficiently manage patent applications.

Future Work

- Pre-processing: How to pre-process data for Patents ? Since its a technical document, how to do that using language models such as SciSpacy etc.
- Topic Modeling- How to decide the optimal number of topics that may exist and can be identified from the data?

The analysis may be done for source of patent applications (government, corporations, individuals) as well.