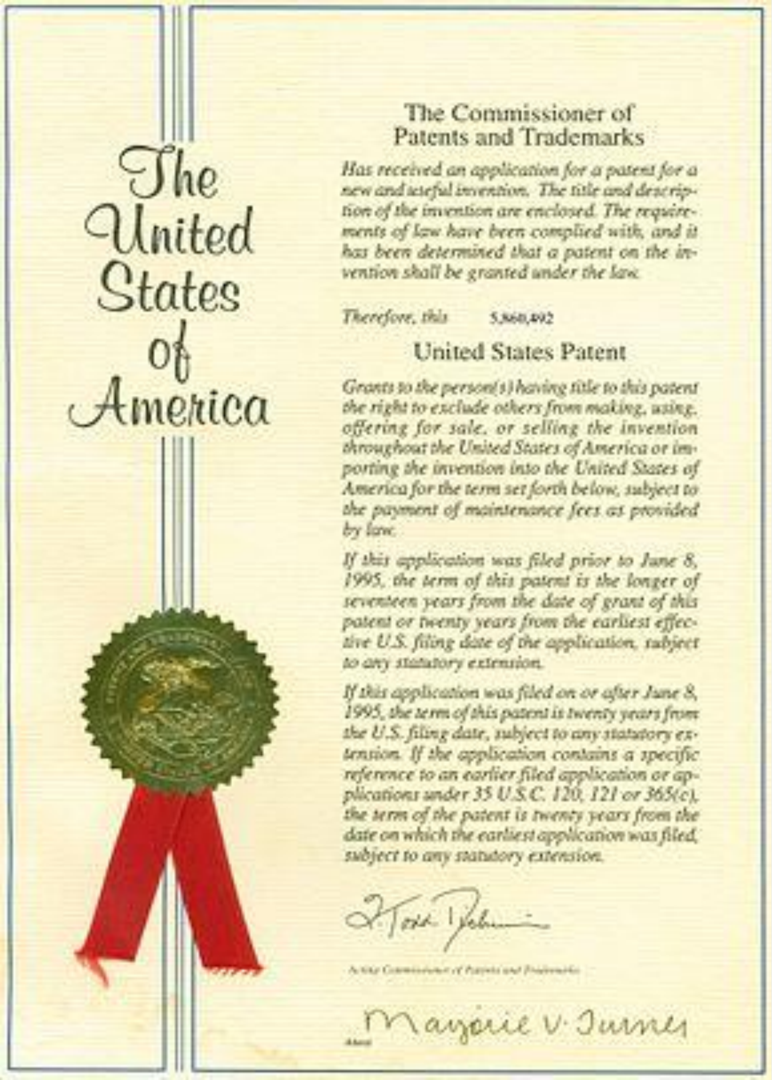
Analyzing Patent Data

Jasraj Singh Bedi
Mandeep Kaur



The United States of America

The Commissioner of Patents and Trademarks

Has received an application for a patent for a new and useful invention. The title and description of the invention are enclosed. The requirements of law have been complied with, and it has been determined that a patent on the invention shall be granted under the law.

Therefore, this 5,860,492

United States Patent

Grants to the person(s) having title to this patent the right to exclude others from making, using, offering for sale, or selling the invention throughout the United States of America or importing the invention into the United States of America for the term set forth below, subject to the payment of maintenance fees as provided by law.

If this application was filed prior to June 8, 1995, the term of this patent is the longer of seventeen years from the date of grant of this patent or twenty years from the earliest effective U.S. filing date of the application, subject to any statutory extension.

If this application was filed on or after June 8, 1995, the term of this patent is twenty years from the U.S. filing date, subject to any statutory extension. If the application contains a specific reference to an earlier filed application or applications under 35 U.S.C. 120, 121 or 365(c), the term of the patent is twenty years from the date on which the earliest application was filed, subject to any statutory extension.

Q. Todd Dickinson

Acting Commissioner of Patents and Trademarks

Attest Marjorie V. Turner

What is Patent?

- A legal protection which gives an inventor the right to exclude others from performing certain activity in the country of issuance.

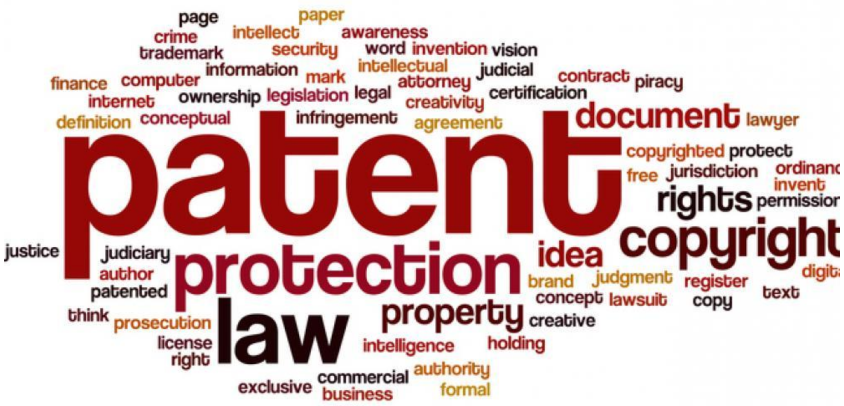
- Sanctioned monopoly for a set number of years in exchange for disclosure to the public.

- Does not give the inventor the right to make, use or sell the patented invention

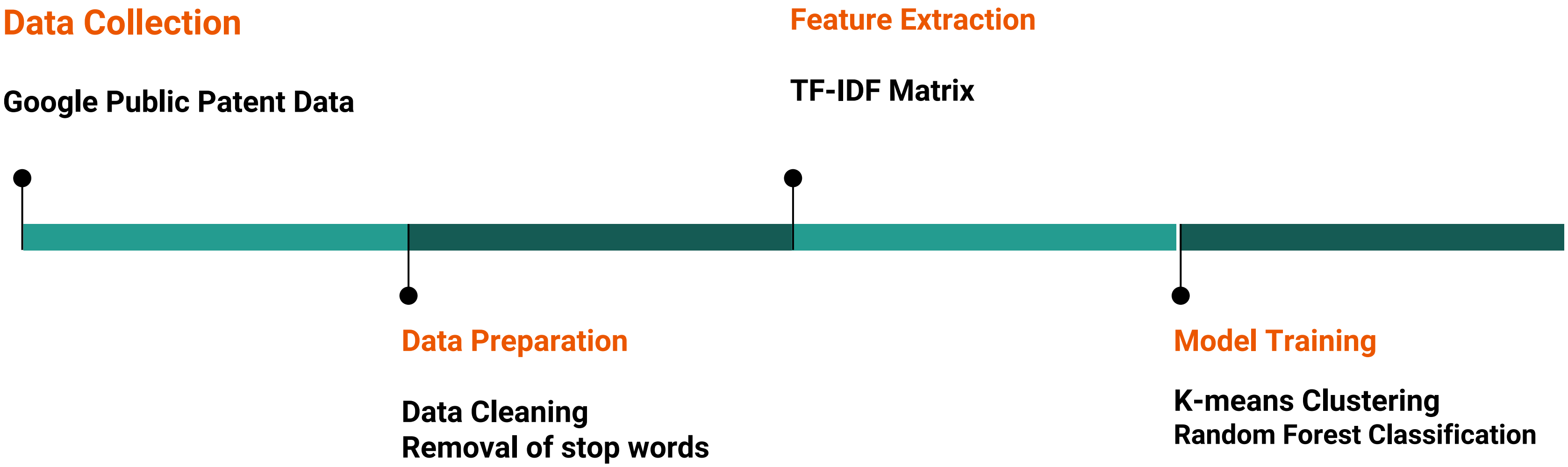
Problem Statement

Our project aims at analyzing patent data and getting useful insights out of it such as,

- Top Potential Competitors
- Technology in Trend (2018)
- Similar Patents



Approach



Data Collection

Google Public Patent Data

Data Preparation

Data Cleaning
Removal of stop words

Feature Extraction

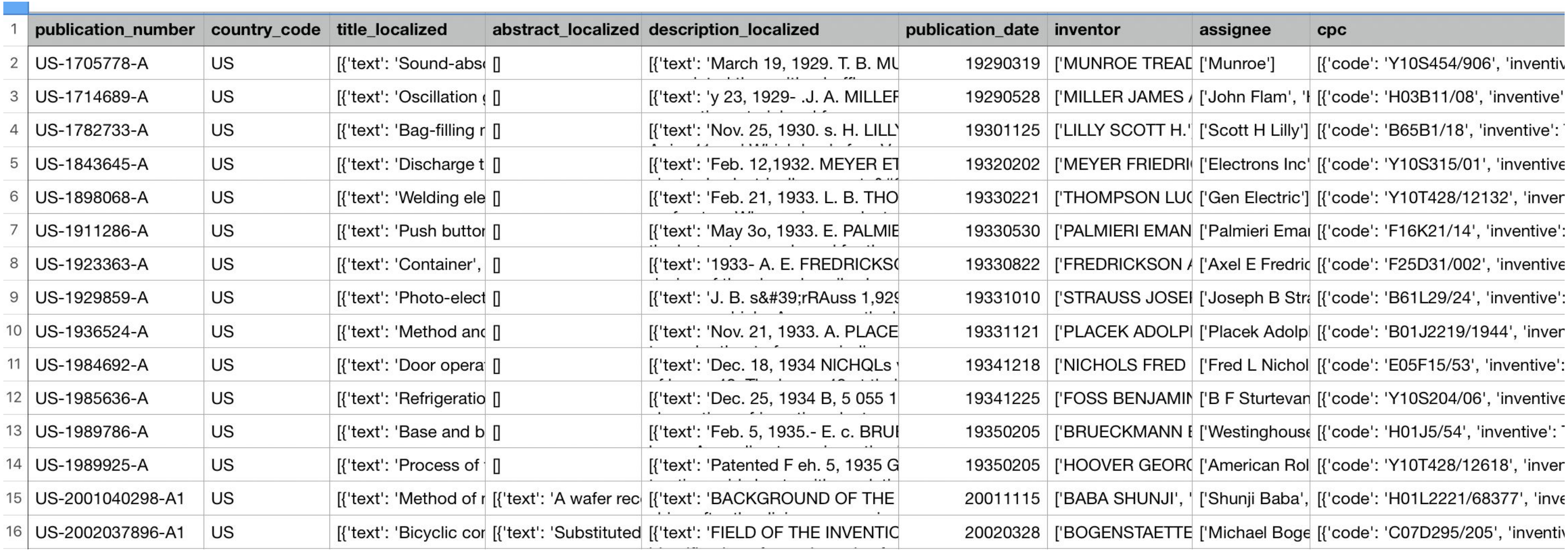
TF-IDF Matrix

Model Training

K-means Clustering
Random Forest Classification

Google Patent Public Dataset (2018)

- Data Collected using BigQuery from Google cloud platform
- 1000 samples and 9 features

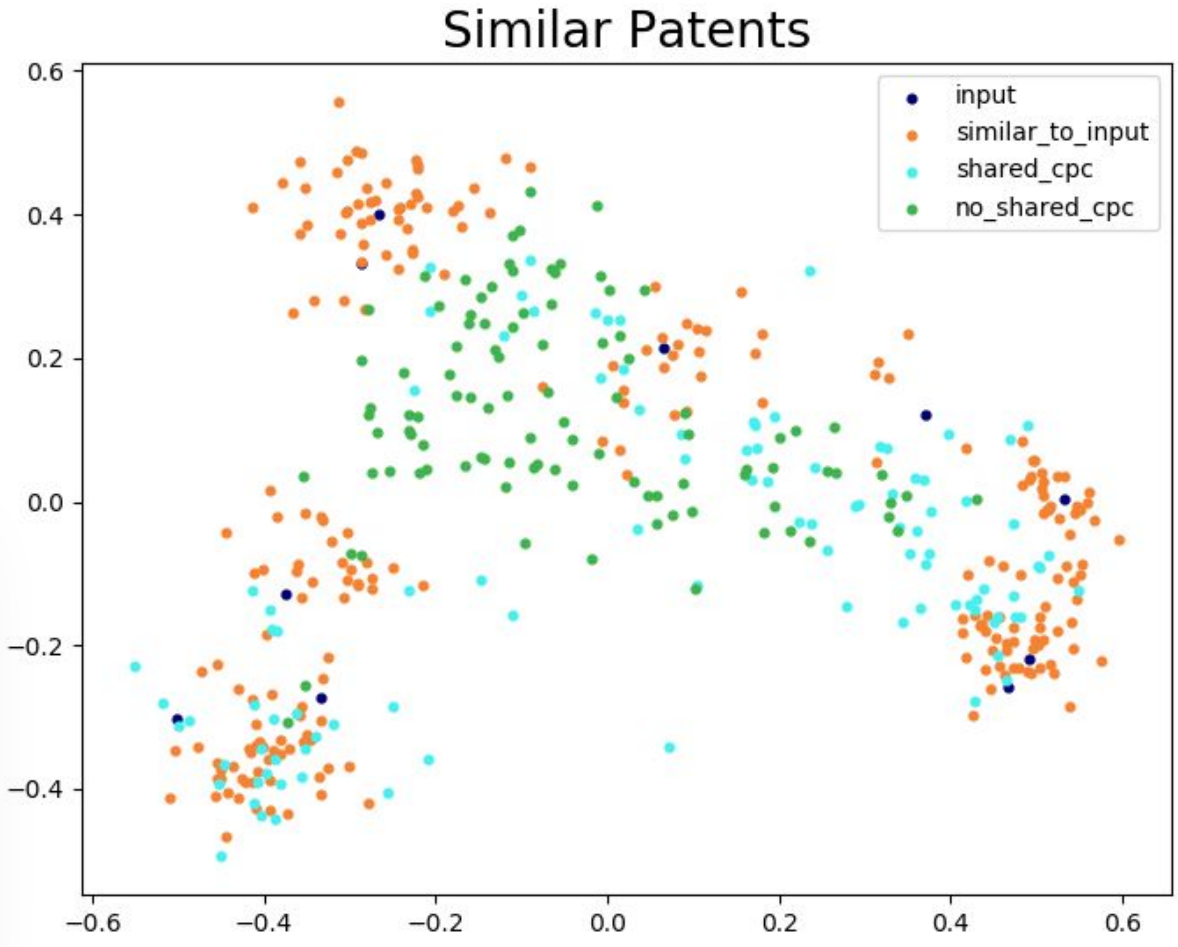


1	publication_number	country_code	title_localized	abstract_localized	description_localized	publication_date	inventor	assignee	cpc
2	US-1705778-A	US	[{'text': 'Sound-abs	[]	[{'text': 'March 19, 1929. T. B. MU	19290319	['MUNROE TREAD	['Munroe']	[{'code': 'Y10S454/906', 'inventiv
3	US-1714689-A	US	[{'text': 'Oscillation	[]	[{'text': 'y 23, 1929- .J. A. MILLEF	19290528	['MILLER JAMES	['John Flam', '	[{'code': 'H03B11/08', 'inventive'
4	US-1782733-A	US	[{'text': 'Bag-filling r	[]	[{'text': 'Nov. 25, 1930. s. H. LILLY	19301125	['LILLY SCOTT H.'	['Scott H Lilly']	[{'code': 'B65B1/18', 'inventive':
5	US-1843645-A	US	[{'text': 'Discharge t	[]	[{'text': 'Feb. 12,1932. MEYER ET	19320202	['MEYER FRIEDRI	['Electrons Inc'	[{'code': 'Y10S315/01', 'inventive
6	US-1898068-A	US	[{'text': 'Welding ele	[]	[{'text': 'Feb. 21, 1933. L. B. THO	19330221	['THOMPSON LU	['Gen Electric']	[{'code': 'Y10T428/12132', 'inver
7	US-1911286-A	US	[{'text': 'Push buttor	[]	[{'text': 'May 3o, 1933. E. PALMIE	19330530	['PALMIERI EMAN	['Palmieri Ema	[{'code': 'F16K21/14', 'inventive':
8	US-1923363-A	US	[{'text': 'Container',	[]	[{'text': '1933- A. E. FREDRICKS	19330822	['FREDRICKSON	['Axel E Fredric	[{'code': 'F25D31/002', 'inventive
9	US-1929859-A	US	[{'text': 'Photo-elect	[]	[{'text': 'J. B. s&#39;rRAuss 1,92	19331010	['STRAUSS JOSE	['Joseph B Str	[{'code': 'B61L29/24', 'inventive':
10	US-1936524-A	US	[{'text': 'Method anc	[]	[{'text': 'Nov. 21, 1933. A. PLACE	19331121	['PLACEK ADOLP	['Placek Adolp	[{'code': 'B01J2219/1944', 'inver
11	US-1984692-A	US	[{'text': 'Door opera	[]	[{'text': 'Dec. 18, 1934 NICHQLs	19341218	['NICHOLS FRED	['Fred L Nichol	[{'code': 'E05F15/53', 'inventive':
12	US-1985636-A	US	[{'text': 'Refrigeratio	[]	[{'text': 'Dec. 25, 1934 B, 5 055 1	19341225	['FOSS BENJAMIN	['B F Sturtevan	[{'code': 'Y10S204/06', 'inventive
13	US-1989786-A	US	[{'text': 'Base and b	[]	[{'text': 'Feb. 5, 1935.- E. c. BRU	19350205	['BRUECKMANN	['Westinghous	[{'code': 'H01J5/54', 'inventive':
14	US-1989925-A	US	[{'text': 'Process of	[]	[{'text': 'Patented F eh. 5, 1935 G	19350205	['HOOVER GEOR	['American Rol	[{'code': 'Y10T428/12618', 'inver
15	US-2001040298-A1	US	[{'text': 'Method of r	[{'text': 'A wafer rec	[{'text': 'BACKGROUND OF THE	20011115	['BABA SHUNJI', '	['Shunji Baba',	[{'code': 'H01L2221/68377', 'inve
16	US-2002037896-A1	US	[{'text': 'Bicyclic cor	[{'text': 'Substituted	[{'text': 'FIELD OF THE INVENTIC	20020328	['BOGENSTAETTE	['Michael Boge	[{'code': 'C07D295/205', 'inventiv

Descriptive Analysis of Data

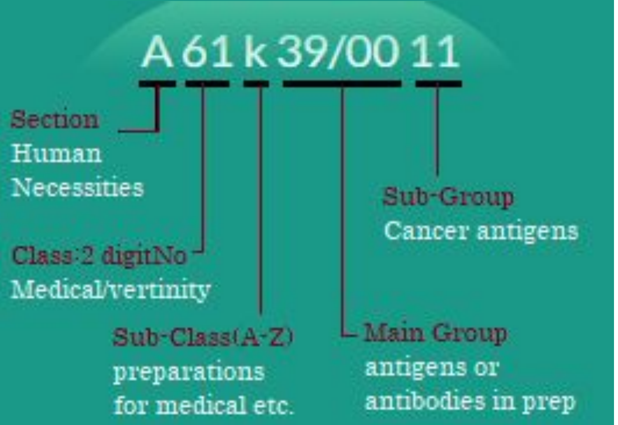
- We found similar patents by comparing their CPC codes using BigQuery.
- Data Visualization using PCA

CPC - Cooperative Patent Classification
It has 9 classes & 250000 sub-classifications:
A- Human Necessities
B- Performing Operations;Transporting
C-Chemistry;Metallurgy
D-Textiles;Paper
E-Fixed Construction
F-Mechanical Engineering
G-Physics
H-Electricity
Y-General tagging of new Technology



Descriptive Analysis of Data

CPC Codes



A 61 k 39/00 11

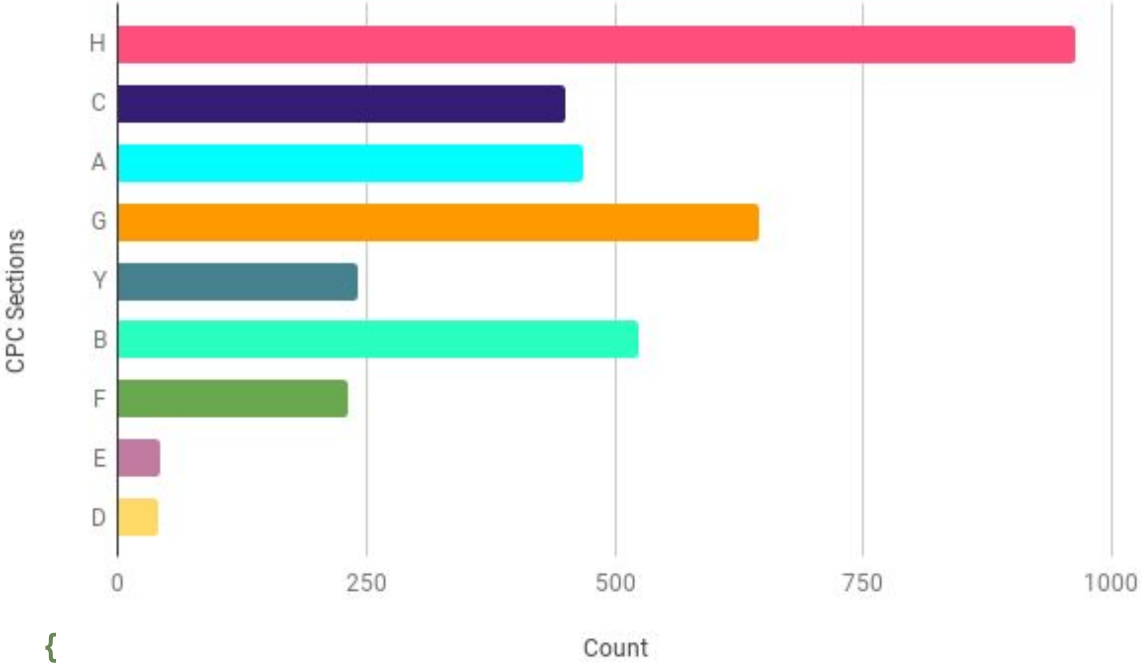
Section
Human
Necessities

Class:2 digitNo
Medical/vertinity

Sub-Class(A-Z)
preparations
for medical etc.

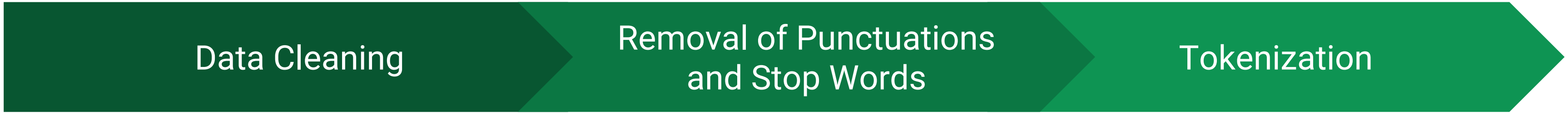
Main Group
antigens or
antibodies in prep

Sub-Group
Cancer antigens



```
{
        "A": "Human Necessities",
        "B": "Operations and Transport",
        "C": "Chemistry and Metallurgy",
        "D": "Textiles",
        "E": "Fixed Constructions",
        "F": "Mechanical Engineering",
        "G": "Physics",
        "H": "Electricity",
        "Y": "Emerging Cross-Sectional Technologies"
}
```

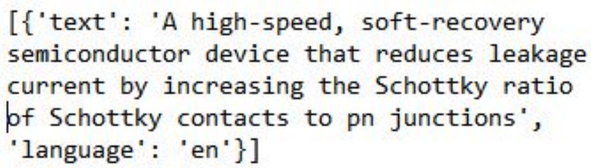
Data Preparation



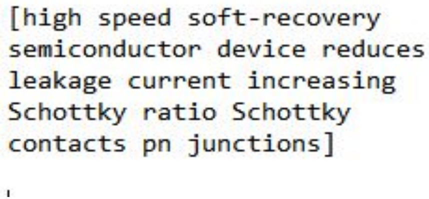
Data Cleaning

Removal of Punctuations and Stop Words

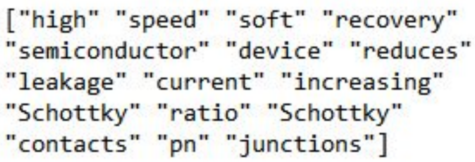
Tokenization



```
[{'text': 'A high-speed, soft-recovery
semiconductor device that reduces leakage
current by increasing the Schottky ratio
of Schottky contacts to pn junctions',
'language': 'en'}]
```



```
[high speed soft-recovery
semiconductor device reduces
leakage current increasing
Schottky ratio Schottky
contacts pn junctions]
```



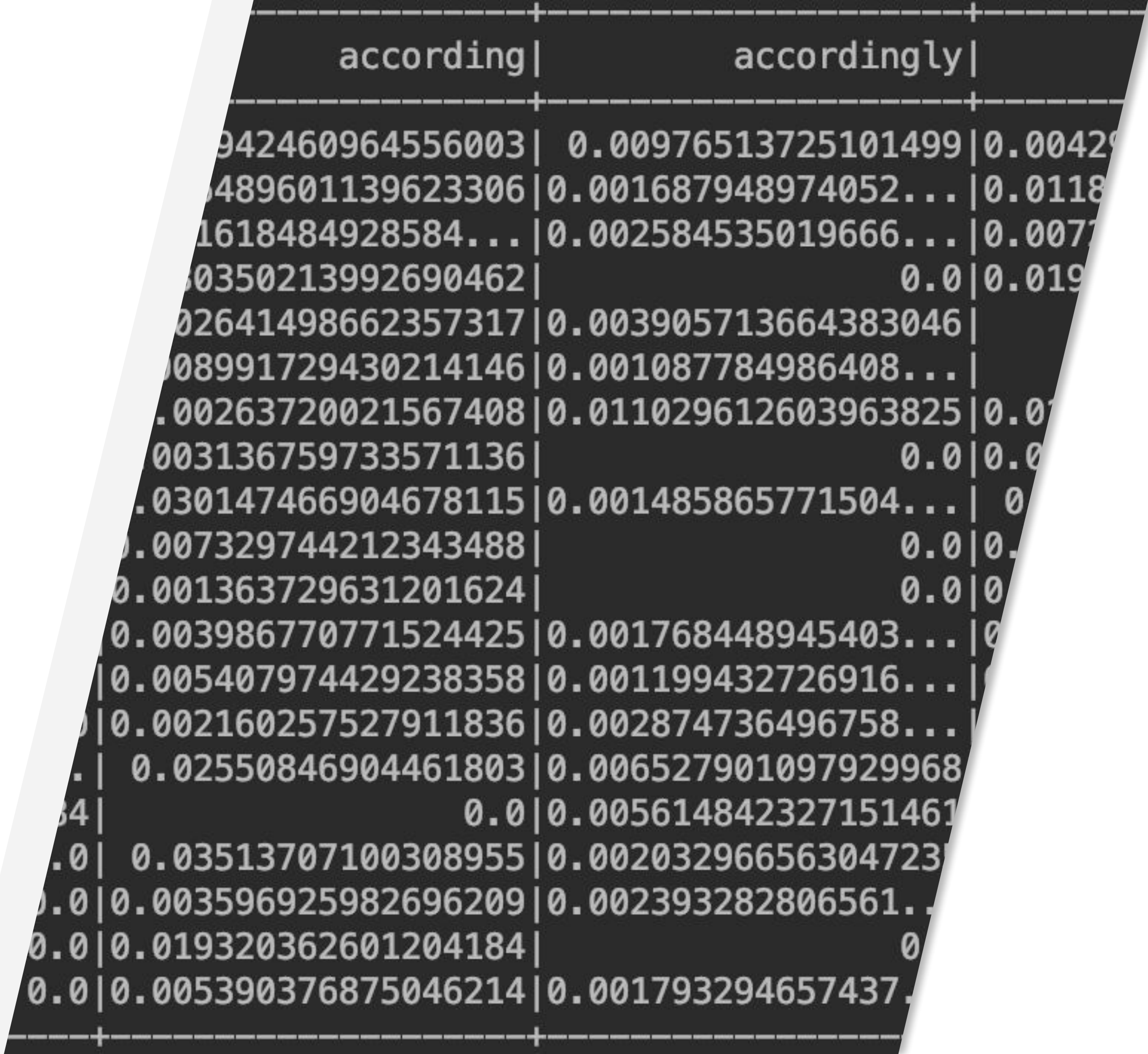
```
["high" "speed" "soft" "recovery"
"semiconductor" "device" "reduces"
"leakage" "current" "increasing"
"Schottky" "ratio" "Schottky"
"contacts" "pn" "junctions"]
```

Feature Extraction

Tf-IDf matrix

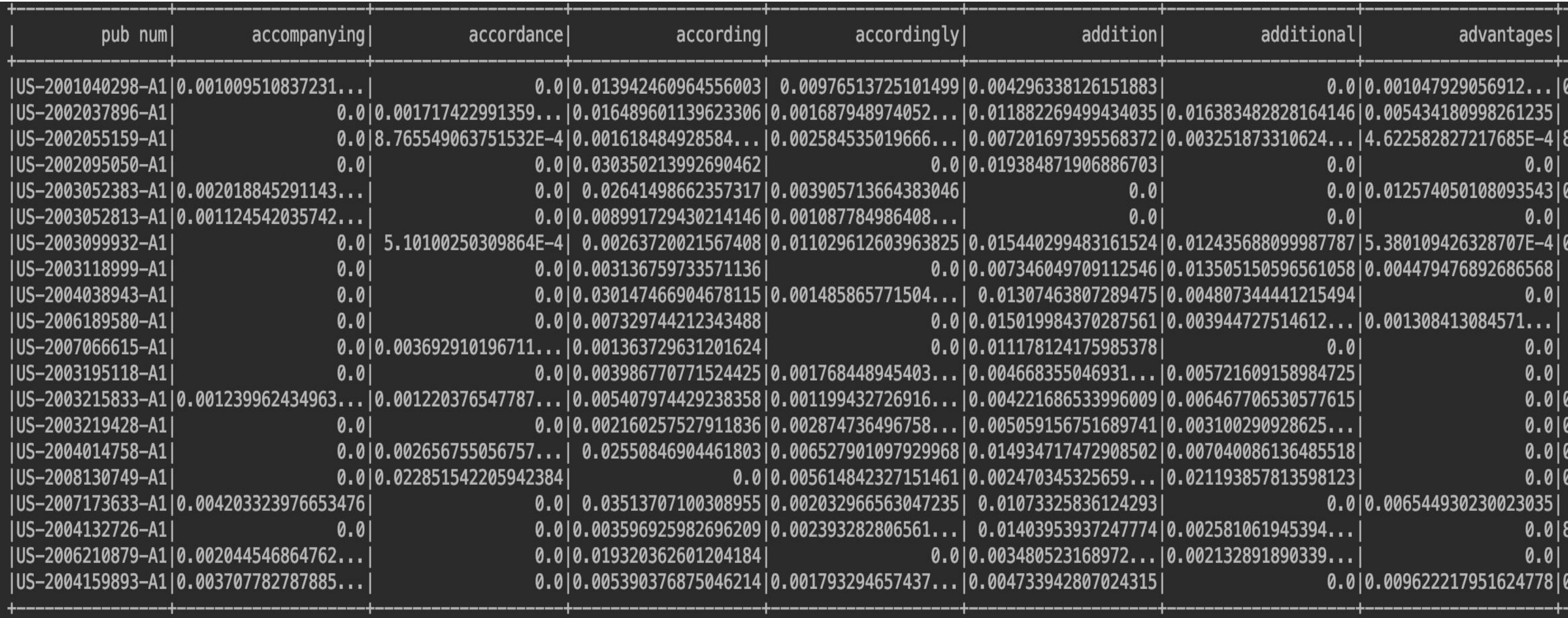
Parameters used:

- min_df=0.4
- smooth_idf=True
- lowercase=True
- analyzer='word'
- use_idf=True



```
-------------------+--------------------+--------
          according|         accordingly|
-------------------+--------------------+--------
    942460964556003| 0.00976513725101499|0.0042
    489601139623306|0.001687948974052...|0.0118
   1618484928584...|0.002584535019666...|0.007
   0350213992690462|                 0.0|0.019
  02641498662357317|0.003905713664383046|
  08991729430214146|0.001087784986408...|
 .0026372002156740 8|0.011029612603963825|0.0
 0031367597335711 36|                 0.0|0.0
 .030147466904678115|0.001485865771504...| 0
 .007329744212343488|                 0.0|0.
0.001363729631201624|                 0.0|0
0.003986770771524425|0.001768448945403...|0
|0.005407974429238358|0.001199432726916...|
|0.002160257527911836|0.002874736496758...|
.| 0.02550846904461803|0.006527901097929968
34|                 0.0|0.005614842327151461
.0| 0.03513707100308955|0.00203296656304723
.0|0.003596925982696209|0.002393282806561..
0.0|0.019320362601204184|                  0
0.0|0.005390376875046214|0.001793294657437.
----+--------------------+-----------------
```

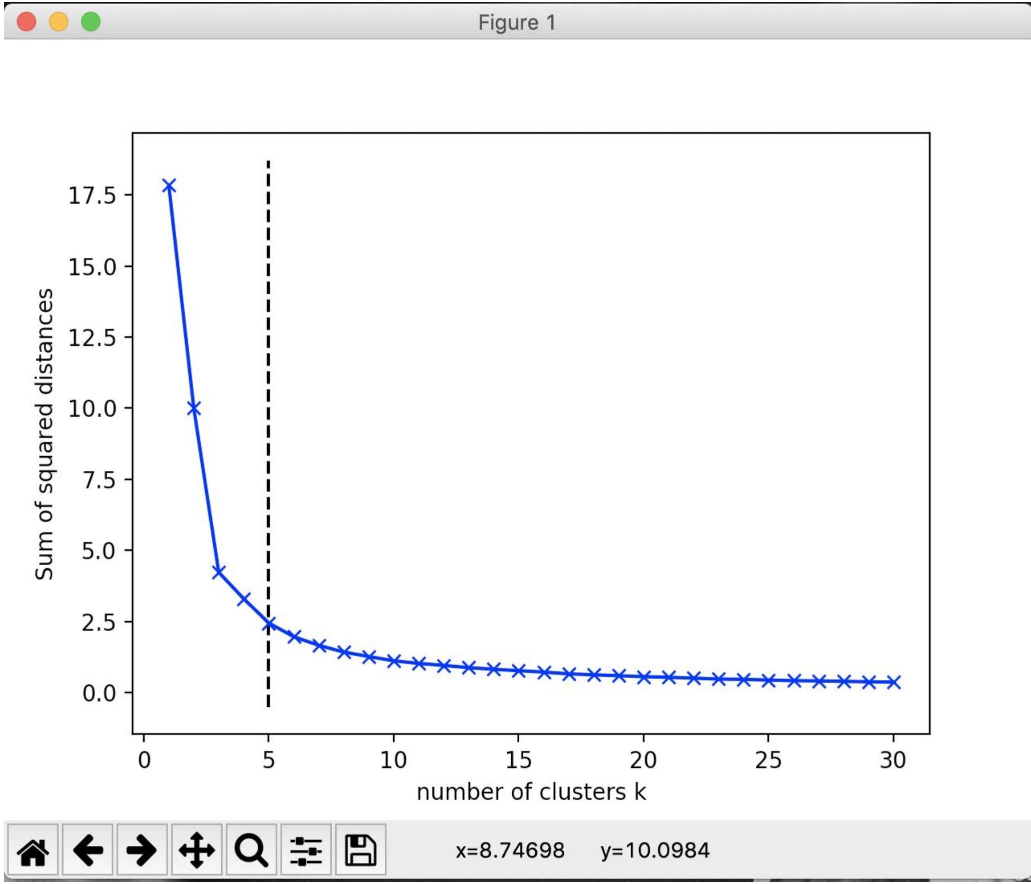
Term Frequency -Inverse Document Frequency (TF-IDF)



```
+----------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-
|         pub num|        accompanying|          accordance|           according|         accordingly|            addition|          additional|          advantages|
+----------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-
|US-2001040298-A1|0.001009510837231...|                 0.0|0.013942460964556003| 0.00976513725101499|0.004296338126151883|                 0.0|0.001047929056912...|0
|US-2002037896-A1|                 0.0|0.001717422991359...|0.016489601139623306|0.001687948974052...|0.011882269499434035|0.016383482828164146|0.005434180998261235|
|US-2002055159-A1|                 0.0|8.765549063751532E-4|0.001618484928584...|0.002584535019666...|0.007201697395568372|0.003251873310624...|4.622582827217685E-4|8
|US-2002095050-A1|                 0.0|                 0.0|0.030350213992690462|                 0.0|0.019384871906886703|                 0.0|                 0.0|
|US-2003052383-A1|0.002018845291143...|                 0.0| 0.02641498662357317|0.003905713664383046|                 0.0|                 0.0|0.012574050108093543|0
|US-2003052813-A1|0.001124542035742...|                 0.0|0.008991729430214146|0.001087784986408...|                 0.0|                 0.0|                 0.0|
|US-2003099932-A1|                 0.0| 5.10100250309864E-4| 0.00263720021567408|0.011029612603963825|0.015440299483161524|0.012435688099987787|5.380109426328707E-4|0
|US-2003118999-A1|                 0.0|                 0.0|0.003136759733571136|                 0.0|0.007346049709112546|0.013505150596561058|0.004479476892686568|
|US-2004038943-A1|                 0.0|                 0.0|0.030147466904678115|0.001485865771504...| 0.01307463807289475|0.004807344441215494|                 0.0|
|US-2006189580-A1|                 0.0|                 0.0|0.007329744212343488|                 0.0|0.015019984370287561|0.003944727514612...|0.001308413084571...|
|US-2007066615-A1|                 0.0|0.003692910196711...|0.001363729631201624|                 0.0|0.011178124175985378|                 0.0|                 0.0|
|US-2003195118-A1|                 0.0|                 0.0|0.003986770771524425|0.001768448945403...|0.004668355046931...|0.005721609158984725|                 0.0|
|US-2003215833-A1|0.001239962434963...|0.001220376547787...|0.005407974429238358|0.001199432726916...|0.004221686533996009|0.006467706530577615|                 0.0|0
|US-2003219428-A1|                 0.0|                 0.0|0.002160257527911836|0.002874736496758...|0.005059156751689741|0.003100290928625...|                 0.0|0
|US-2004014758-A1|                 0.0|0.002656755056757...| 0.02550846904461803|0.006527901097929968|0.014934717472908502|0.007040086136485518|                 0.0|0
|US-2008130749-A1|                 0.0|0.022851542205942384|                 0.0|0.005614842327151461|0.002470345325659...|0.021193857813598123|                 0.0|0
|US-2007173633-A1|0.004203323976653476|                 0.0| 0.03513707100308955|0.002032966563047235| 0.01073325836124293|                 0.0|0.006544930230023035|
|US-2004132726-A1|                 0.0|                 0.0|0.003596925982696209|0.002393282806561...| 0.01403953937247774|0.002581061945394...|                 0.0|8
|US-2006210879-A1|0.002044546864762...|                 0.0|0.019320362601204184|                 0.0|0.003480523168972...|0.002132891890339...|                 0.0|
|US-2004159893-A1|0.003707782787885...|                 0.0|0.005390376875046214|0.001793294657437...|0.004733942807024315|                 0.0|0.009622217951624778|0
+----------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-
```

K means Clustering

- Unsupervised Learning algorithm.
- Elbow point found at K = 5 (represented by dashed line)



Principal Component Analysis (PCA)

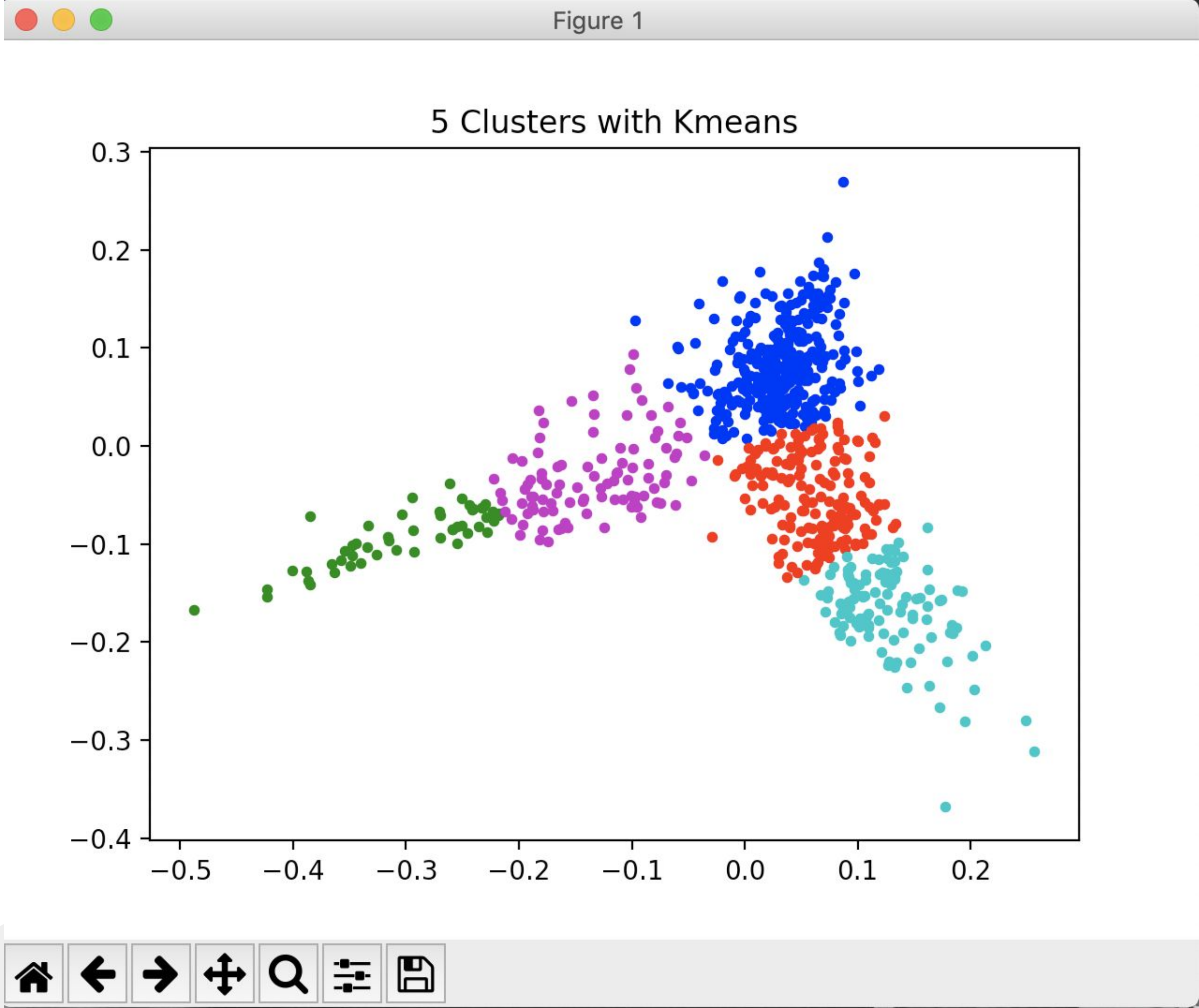
The main idea of **principal component analysis**(**PCA**) is to reduce the dimensionality of a **data** set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.

IN BRIEF, PCA IS USED TO :

- **Reduce number of dimensions in data**
- **Find patterns in high-dimensional data**
- **Visualize data of high dimensionality**

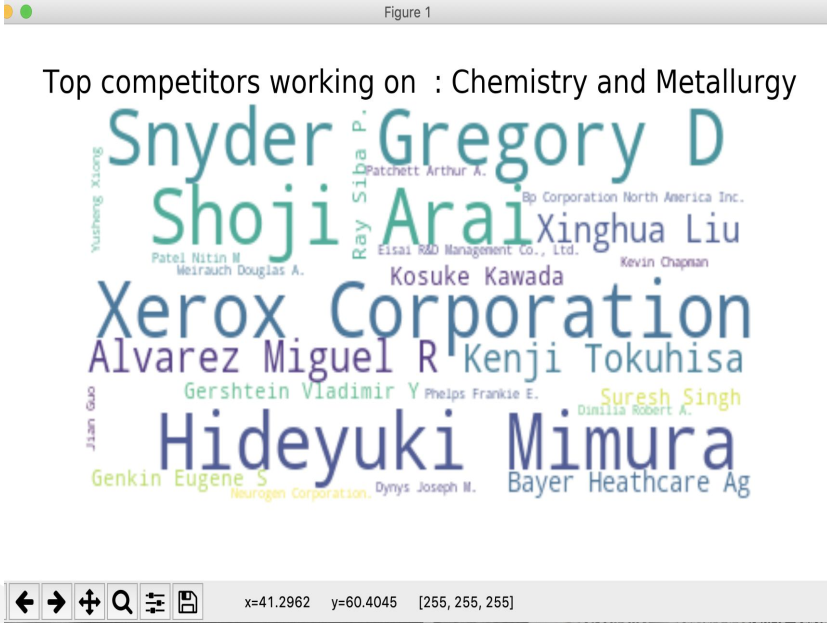
High dimensional data was converted to 2-D for better visualization of clusters using PCA.

- **Silhouette**, with squared euclidean distance = 0.054

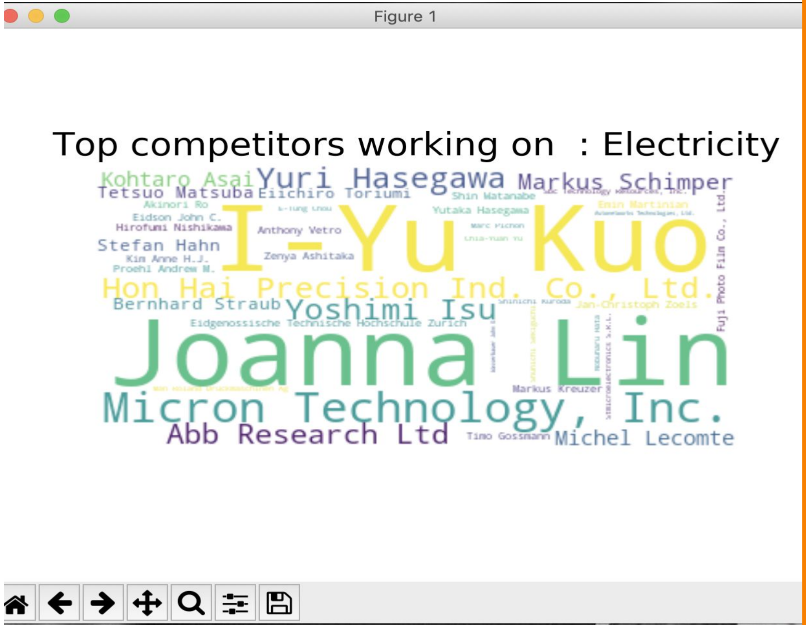


Analysis

- Word clouds generated for two random clusters representing top potential competitors clustered together.
- Technologies in trend was found to be Electricity, Chemistry and Metallurgy which maximum count n the clusters.



Top competitors working on : Chemistry and Metallurgy



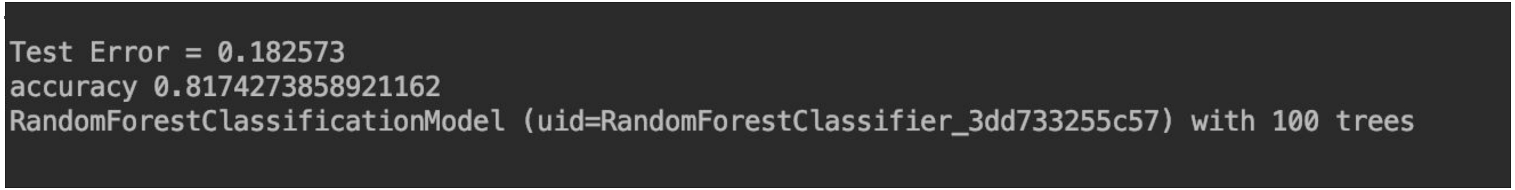
Top competitors working on : Electricity

Random Forest Classifier

70% Training Data

30% Testing Data

- Supervised Learning Algorithm
- Training Data - K clusters of patents were taken and labelled.
- 10 fold cross validation
- Accuracy obtained is 81%



```
Test Error = 0.182573
accuracy 0.8174273858921162
RandomForestClassificationModel (uid=RandomForestClassifier_3dd733255c57) with 100 trees
```

Conclusion

- We are able to find the potential competitors, which can benefit a company to knowing its competitor.
- By analysing similar cpc codes (matching the first character of the cpc code) in the cluster, we can predict in which cluster does a new patent belongs to using classification.

Challenges

- **Cleaning the data:**

  Some of the data had inconsistent type for a column . For Example , in some rows the assignee names were in quotes and in the some of the rows the assignee data was in string representation of a dictionary

- **BigQuery quota exceeded error** :

  BigQuery gave quota exceeded error while writing custom queries.

Future work and Improvements

1. Research is needed to extract better meaningful words from the patent description, title and abstract and neglect non pertinent words.
2. More features can be extracted from the data, dates and countries.
3. The results are restricted to 1000 patents. We would like to do the analysis on complete patent dataset using multiple nodes which would yield more accurate results.

Thank you!