MACQUARIE University

CLASSIFYING PATENT APPLICATIONS

BY HEMESH TALLURI

Supervised by Diego Moll´a-Aliod

DATE: 9 NOVEMBER 2018

GOAL OF THE PROJECT

- To automate the classification of patent documents into various sections of the primary IPC mark using machine learning algorithms.

- The intention of classification is to enable quick search for patent documents and to track the trends in patent applications.

BACKGROUND AND LITERATURE REVIEW

- Document Classification can be broadly classified into two categories:
  - Supervised
  - Unsupervised

- Factors contributing to classification: Feature Extraction and Topic ambiguity.

- Techniques employed: Expedition maximization, Naïve Bayes classifier, Support Vector Machine, Decision Trees, Neural Network, etc.

- Benzineb K., Guyot J. (2011) Automated Patent Classification. In: Lupu M., Mayer K., Tait J., Trippe A. (eds) Current Challenges in Patent Information Retrieval. The Information Retrieval Series, vol 29. Springer, Berlin, Heidelberg

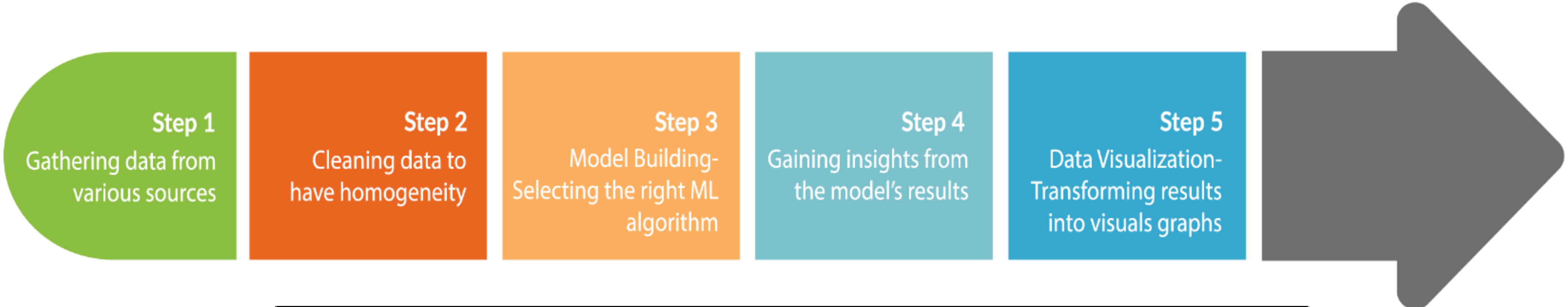
- Seneviratne D., Geva S., Zuccon G., Ferraro G., Chappell T., Meireles M. (2015) A Signature Approach to Patent Classification. In: Zuccon G., Geva S., Joho H., Scholer F., Sun A., Zhang P. (eds) Information Retrieval Technology. AIRS 2015. Lecture Notes in Computer Science, vol 9460. Springer, Cham

RESEARCH QUESTION

- How Exploratory Data Analysis helps to summarise the characteristics of the dataset?
- How to categorise the documents using machine learning techniques?
- Which machine learning algorithm provides better performance and accuracy for text classification?

METHODOLOGY

MACHINE LEARNING PROCESS



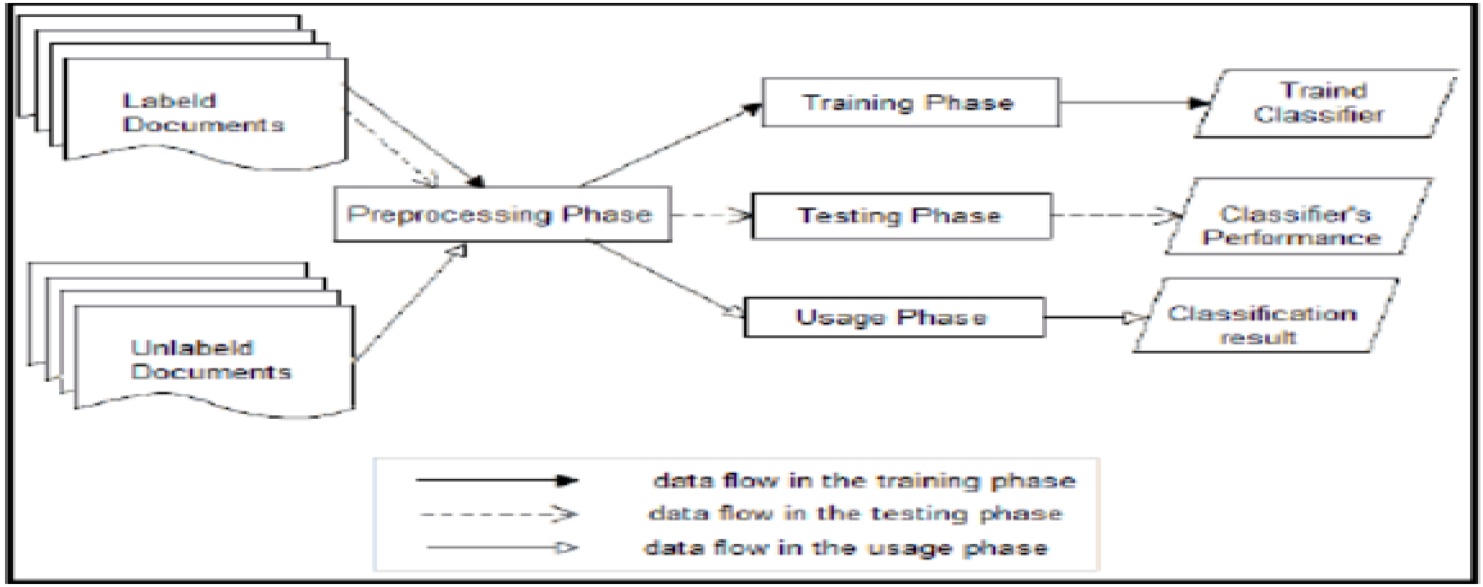
Step 1 Gathering data from various sources

Step 2 Cleaning data to have homogeneity

Step 3 Model Building- Selecting the right ML algorithm

Step 4 Gaining insights from the model's results

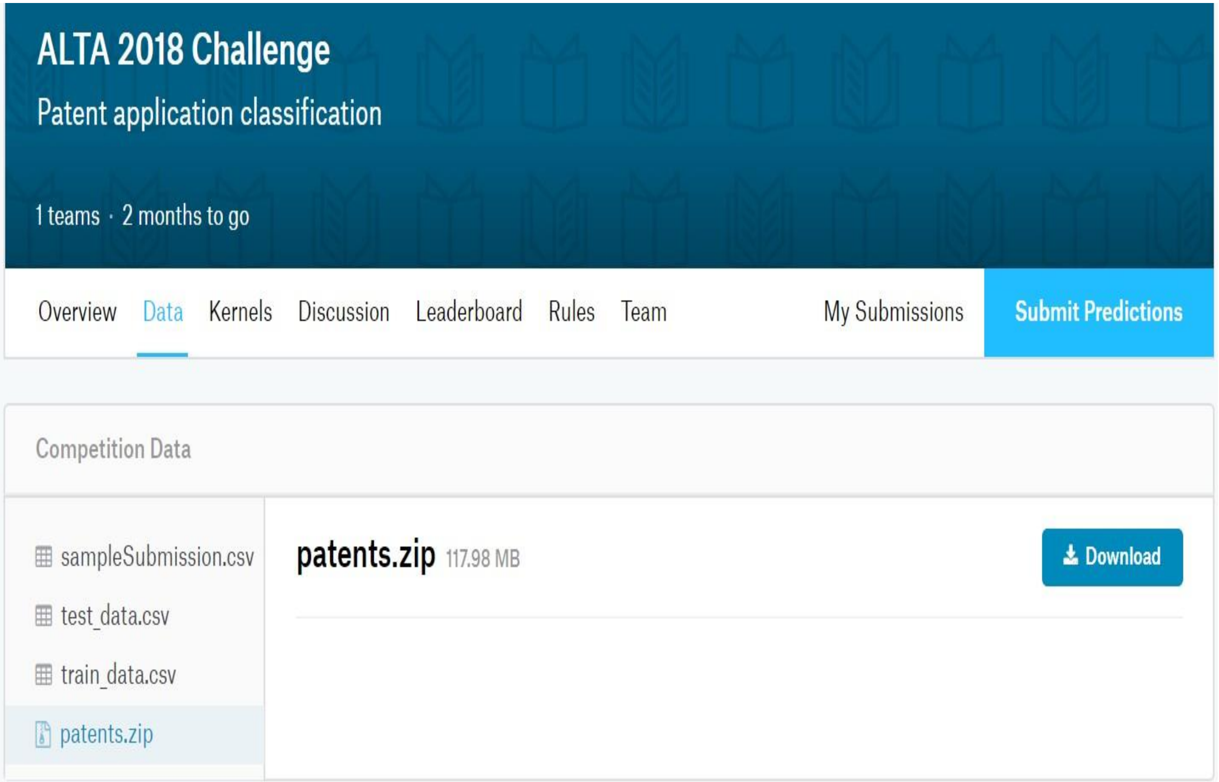
Step 5 Data Visualization- Transforming results into visuals graphs



METHODOLOGY

DATA SET

- Around 1000 patent documents are considered to be classified.
- The documents are to be classified into 8 sections which are given symbols from A to H, each representing a category as follows:
  - A: Human necessities
  - B: Performing operations, transporting
  - C: Chemistry, metallurgy
  - D: Textiles, paper
  - E: Fixed constructions
  - F: Mechanical engineering,
  - G: Physics
  - H: Electricity



ALTA 2018 Challenge

Patent application classification

1 teams · 2 months to go

Competition Data

sampleSubmission.csv

test_data.csv

train_data.csv

patents.zip

patents.zip 117.98 MB

METHODOLOGY

DESCRIPTION

- The sample patent documents are collected from Kaggle.
- **Pre-Processing Phase:** Tokenization, Stop words removal, TFIDF Vectorizer.
- **Training Phase:** Trained with an algorithm. Ex: Naïve Bayes classifier, Support Vector Machine, Decision Trees, SGD Classifier, Random Forest.
- **Testing Phase**: Testing the trained classifier and evaluating its capability for the usage.
- **Evaluation:** The performance of a classifier is evaluated by comparing the predicted sections with the actual sections.
- **Usage Phase:** The classifier in this phase is successfully trained, tested and evaluated and ready for classification of new data whose sections are unknown.

METHODOLOGY

TOOLS EMPLOYED

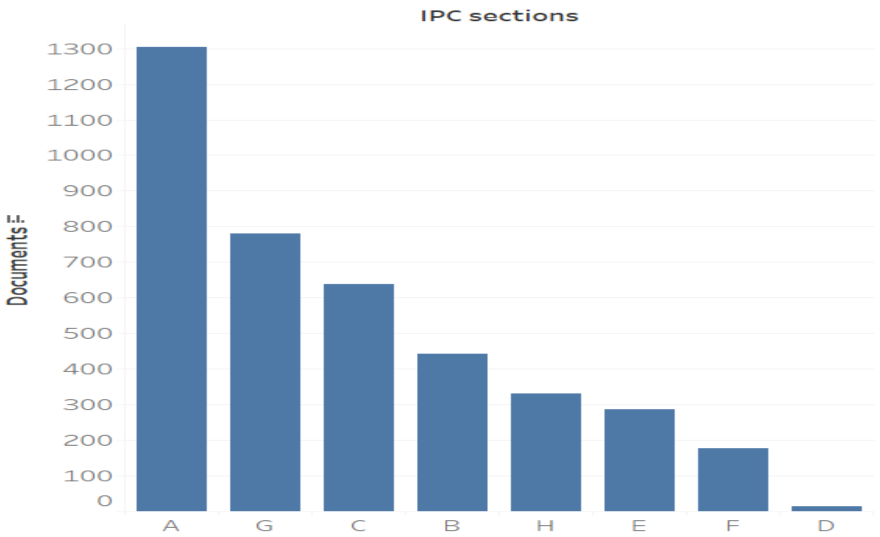
- **PANDAS:**
  - Data analysis tools for the Python programming language
- **NLTK:**
  - Text processing libraries for classification, tokenization, stemming, tagging and parsing.
- **SCIKIT-LEARN:**
  - To build the classification model in this project.
- **MATPLOT:**
  - For data-visualization purpose in this project.

RESULTS

TRAINING PHASE

- For training purpose, labelled documents → 3972
- For usage purpose, unlabelled documents → 1000
- After pre-processing phase, various algorithms are used for training the data.
- Table shows the accuracy (F1) obtained for each classifier:

Classifier	F1_Score
Decision Tree	0.4650253883880472
Naïve Bayes	0.512331313307886
Random Forest Classifier	0.5848495679872141
SGD	0.7044207445578385
k-Nearest Neighbor	0.6512946324677934
Gradient Boosting Classifier	0.5770175258881318



INPUT DATA

IPC Sections	Number of Documents
A	1303
G	781
C	637
B	442
H	330
E	287
F	178
D	14

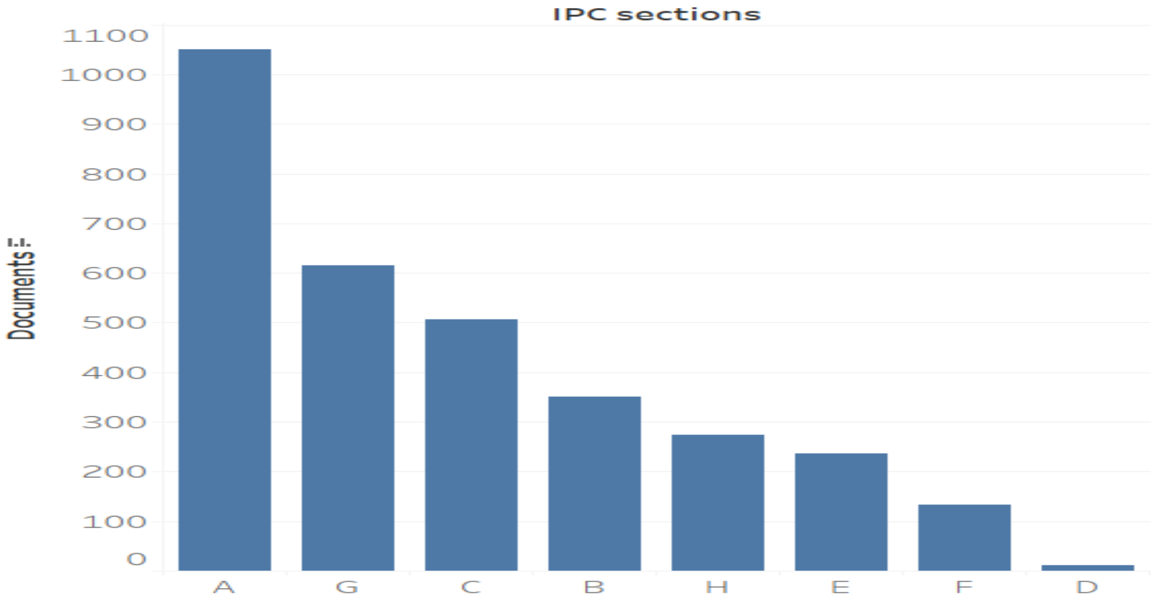
RESULTS

EVALUATION PHASE

- Evaluation of high performer classifier i.e., SGD Classifier -

- Splitting the train data (3972 documents) into test data (20%) and train data (80%)

- Accuracy obtained in SGD classifier is 0.6966019090315566



INPUT DATA (80%)FOR SGD CLASSIFIER

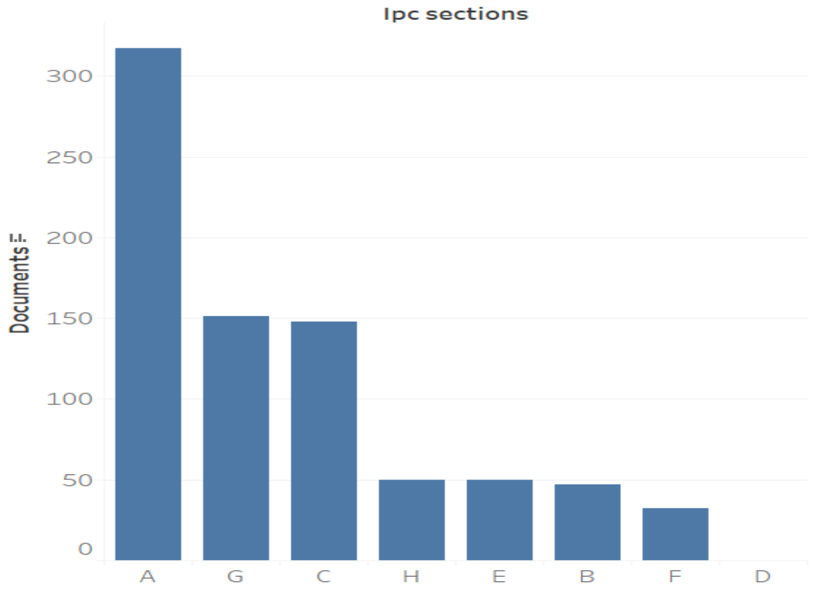
IPC Sections	Number of Documents
A	1050
G	615
C	506
B	351
H	273
E	237
F	133
D	12

RESULTS

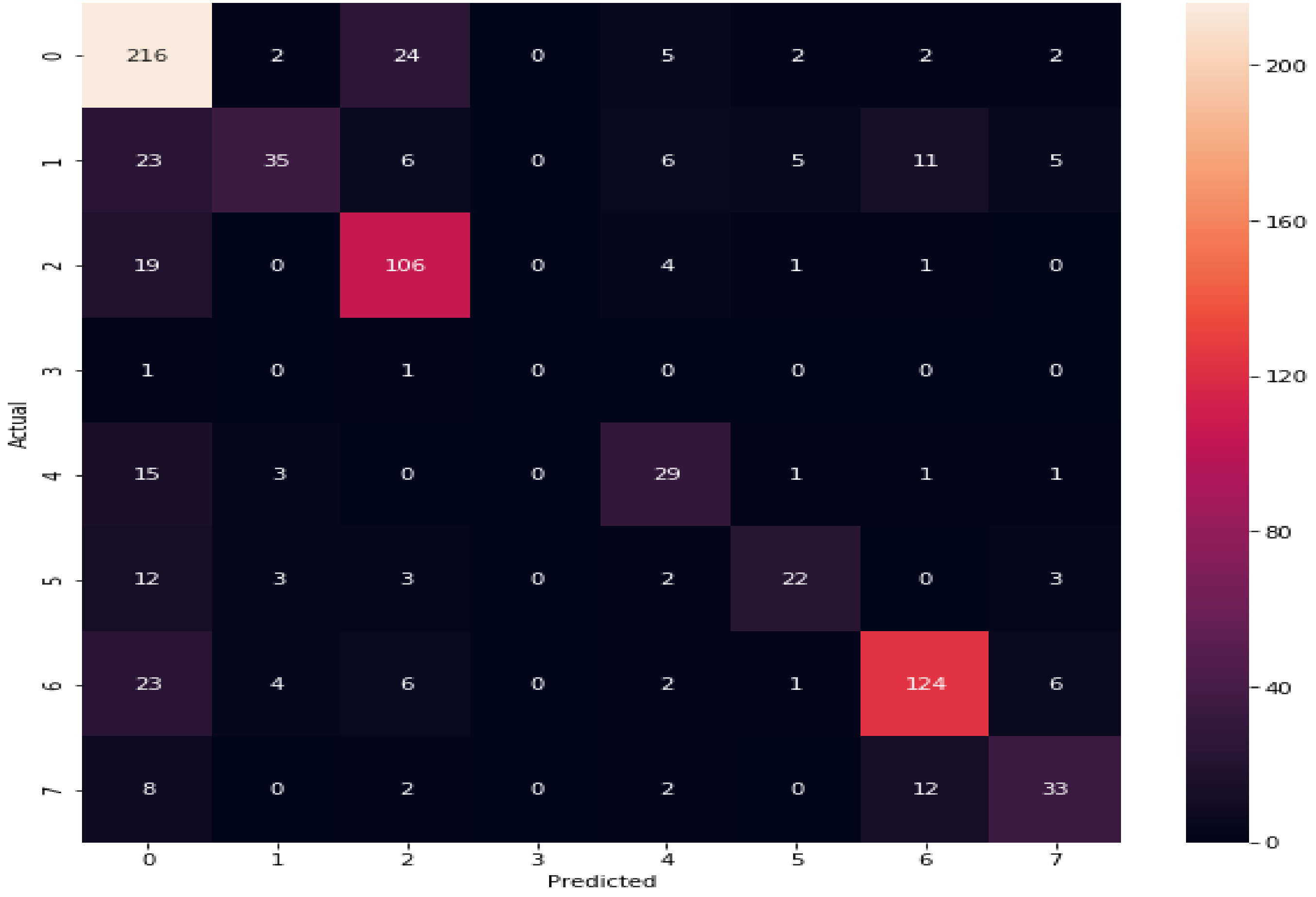
EVALUATION PHASE

- Now, predicting the test data (20%):

- Building a confusion matrix based on actual data and predicted data.



PREDICTED DATA

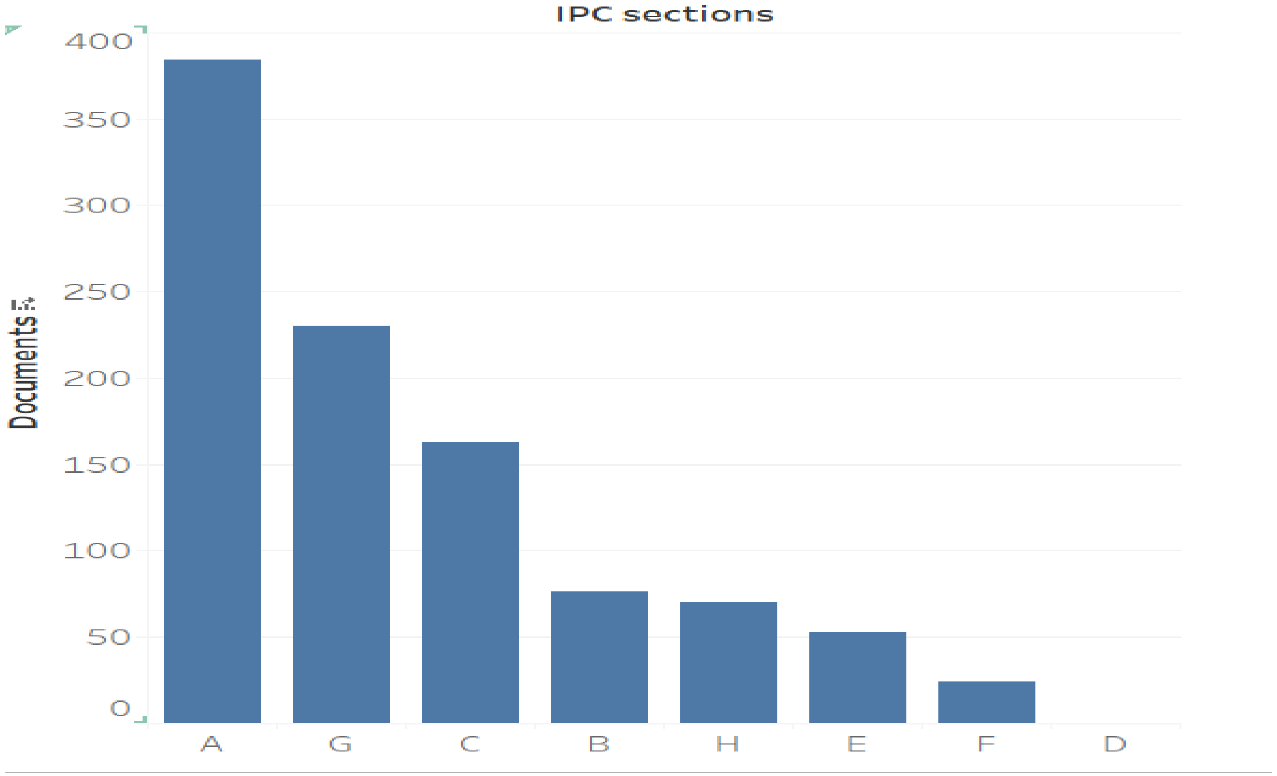


	0	1	2	3	4	5	6	7
0	216	2	24	0	5	2	2	2
1	23	35	6	0	6	5	11	5
2	19	0	106	0	4	1	1	0
3	1	0	1	0	0	0	0	0
4	15	3	0	0	29	1	1	1
5	12	3	3	0	2	22	0	3
6	23	4	6	0	2	1	124	6
7	8	0	2	0	2	0	12	33

RESULTS

USAGE PHASE

- Using SGD Classifier, the unlabelled documents are classified.



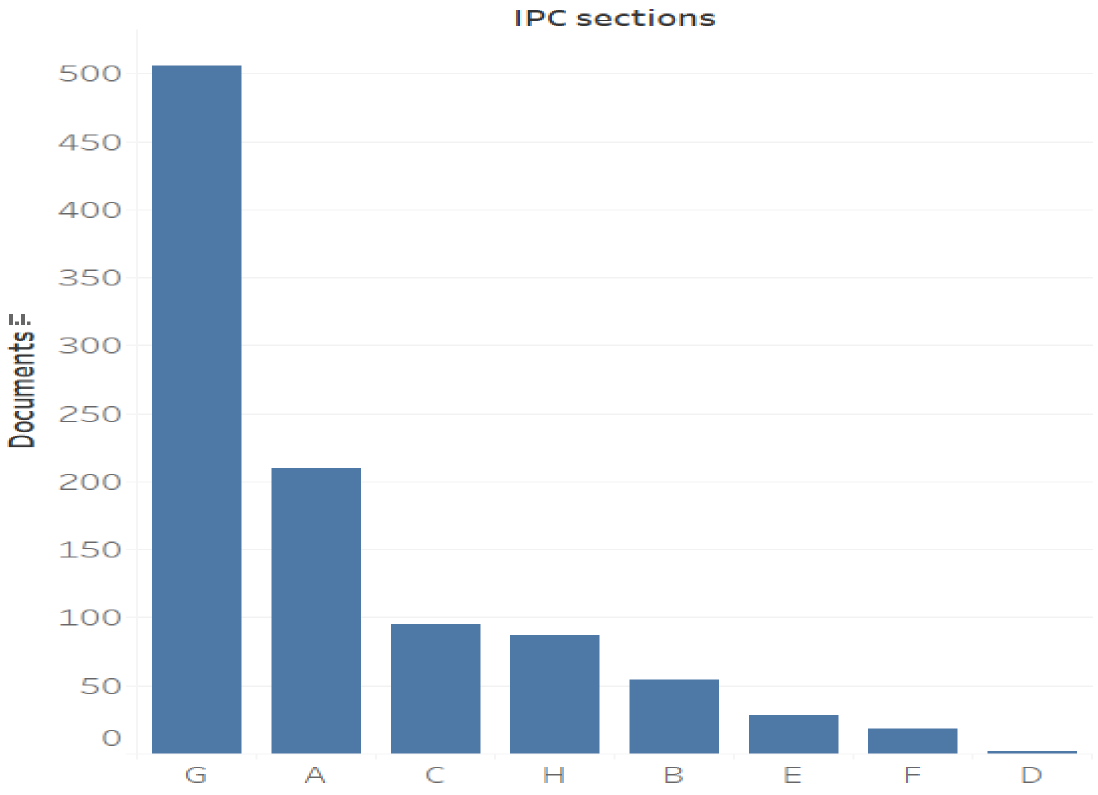
IPC Section	Number of Documents
A	384
G	230
C	163
B	76
H	70
E	53
F	24
D	0

PREDICTED OUTPUT FOR UNLABELLED DOCUMENTS

RESULTS

TESTING THE TOOL

- 1000 documents are tested.



IPC Section	Number of Documents
G	506
A	210
C	95
H	87
B	54
E	28
F	18
D	2

PREDICTED OUTPUT FOR UNLABELLED DOCUMENTS

CONCLUSION

- Various algorithms are tested on trained data for accuracy.
- SGD Classifier is evaluated by confusion matrix.
- The unlabelled patent documents are automatically classified into the first character of the primary IPC mark i.e., the section symbol (A to H) IPC classification.
- The working of the tool is tested on different set of documents.

Thank you
ANY
QUERIES