Contents

Document Classification

Problem Definition:

Given the document d and the fixed set of classes $C = \{c_1, c_2, c_3, c_4 \dots, c_n\}$ and the training set m labelled documents $\{(d_1, c_1), (d_2, c_1), (d_3, c_2), (d_4, c_2) \dots (c_m, c_n)\}$, define an objective function f to classify the document d to the respective class.

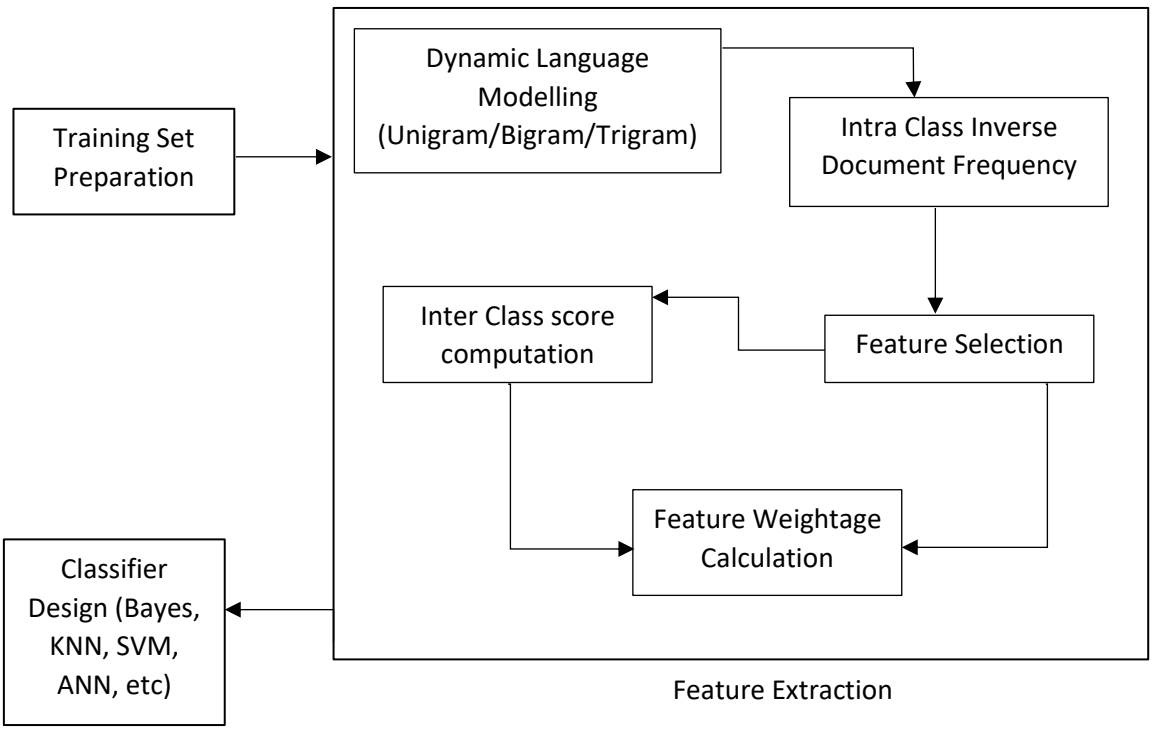
$$f(d) = c \; where \; c \in C$$

Assumption

Features are independent in nature

Solution:

The solution has different steps expressed in the below flow chart.

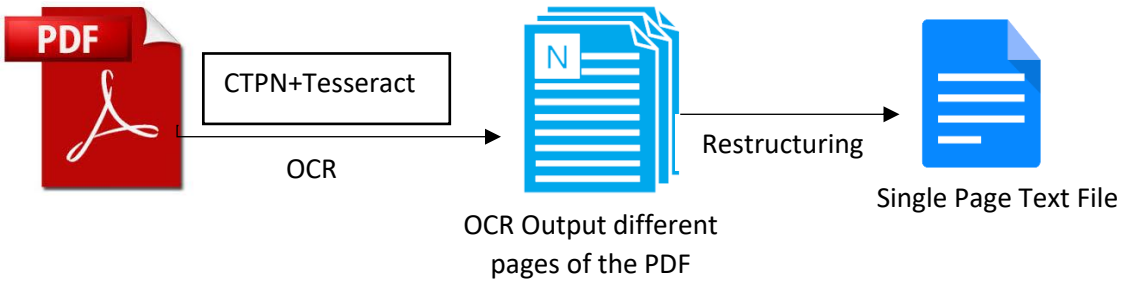


Training Set Preparation

Let's assume that we have three classes $C = \{c_1, c_2, c_3\}$, categories each document manually to these classes. We can follow the folder (directory)data structure for the same.

c_1 c_2 c_3

Once the categorization is done, apply text detection, followed by restructuring module. Once the restructuring is done make a single string representation for each document



Feature Extraction

The input of the feature extraction module is the single page text files for different documents and their labels.

Dynamic Language Modelling

Generate a corpus of unigram, bigram and trigram from all the campus of the class. At the end we should have bag of words for every class, since we have three classes, we would have three different bag of words corpuses.

```
import nltk
from nltk.util import ngrams
#############################################################################
Text="The first step toward success is taken when you refuse to be a
captive of the environment in which you first find yourself"
tokens = nltk.word_tokenize(Text)
print(tokens)
unigram = list(ngrams(tokens,1)) # Unigram
print(unigram)
bigram = list(ngrams(tokens,2)) # Bigrams
print(bigram)
trigram = list(ngrams(tokens,3)) # Bigrams
print(trigram)
#############################################################################
```

Intra Class Inverse Document Frequency

Let's assume that we have a corpus $W = \{f_1, f_2, f_3, f_4, f_5, \dots, f_p\}$ composed from all the samples of the class c_1. Find the score of each feature sing below expression

$$\alpha_i = \frac{number\ of\ documents\ where\ f_i\ appeared}{total\ number\ of\ document}$$

If we have 3 document and the feature appeared in all the three documents of that class, the score would be 1.

Feature Selection

Select top k features from the corpus whose score is greater than the defined threshold. Also remove the ambiguous unigram and bigram features.

Inter Class Score Computation

Since every class has its own corpus find, we will now check the interclass presence of the feature. It will be computed using the below expression.

$$\beta_i = 1 - \sum_{i=1}^{C} \frac{number\ of\ documents\ where\ f_j\ appeared\ in\ class\ c_i \Big/ total\ number\ of\ documents\ in\ class\ c_i}{total\ number\ of\ classes}$$

We have 3 documents per class, and we have 3 classes, the above expression would be extended as

$$\beta_i = 1 - (\frac{0/3}{3} + \frac{1/3}{3} + \frac{0/3}{3})$$

Here we assume that f_1 (feature-1) appeared only in one document of class 2, hence the score of f_1 is 0.89. similarly, we calculate the inter class score of every feature of every individual class.

Feature Weightage Calculation

Since for every class we have calculated the inter and intra feature score, we will fuse both the score to create a combine score.

$$\omega_i = \alpha_i \times \beta_i$$

When we combine all the features of the classes, we can create a global feature vector. Here W represent the corpus (feature vector) of individual classes.

$$V = \bigcup_{i=1}^{C} W_i$$

At the end of this we would have a global feature vector represented by the two-dimensional array.

$$V = \begin{bmatrix} f_1 & \omega_1 \\ f_2 & \omega_1 \\ f_n & \omega_1 \end{bmatrix}$$

Here $\{f_1, f_2, f_3, f_4, \dots, f_n\}$ are the features represented in the qualitative form [these are the combination of unigram/bigrams/trigrams words] and $\{\omega_1, \omega_2, \omega_3, \omega_4, \dots, \omega_n\}$ are their corresponding weightage.

Classifier Design

Since every document is represented in the high dimensional feature space and the class labels are know to us, we can fit any machine learning classifier.

K- Nearest Neighbour

Input:
Training Population with Class Labels (Here we have documents summarized in vector every vector is represented by the notion $\omega^{i}{}_{j}$ which express that feature j belongs to class i)

$$Training\ Population = \begin{bmatrix} \omega^{1}{}_{1} & \omega^{1}{}_{2} & \omega^{1}{}_{3} & . & \omega^{1}{}_{n} \\ \omega^{1}{}_{1} & \omega^{1}{}_{2} & \omega^{1}{}_{3} & . & \omega^{1}{}_{n} \\ \omega^{2}{}_{1} & \omega^{2}{}_{2} & \omega^{2}{}_{3} & . & \omega^{2}{}_{n} \\ . & . & . & . & . \\ \omega^{C}{}_{1} & \omega^{C}{}_{1} & \omega^{C}{}_{1} & . & \omega^{C}{}_{1} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ . \\ d_m \end{matrix}$$

And K (number of neighbourhood)

Output: The relevant class "c" $\epsilon\ C$

Algorithm
Step1: Find the Euclidean distance between the test vector and all the train vectors

$$Distances = \begin{bmatrix} distance(d, d_1) & c_1 \\ distance(d, d_2) & c_1 \\ distance(d, d_3) & c_2 \\ . & . \\ distance(d, d_m) & c_C \end{bmatrix}$$

Step2: sort the distances in ascending order and take the top k element of the matrix
Step3: find the count of the class in top k rows, the class who have majority (maximum count) would be the output class
Step4: Return the output class

Multinomial Bayes Classifier

$$p(H_i|E) = \frac{p(E|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E|H_k) \times p(H_k)}$$

Evidence=Features

$$p(c_1|\mathrm{f}) = \mathrm{p}(\mathrm{f}|c_1) \times p(c_1) \Big/ \mathrm{p}(\mathrm{f}|c_1) \times p(c_1) + \mathrm{p}(\mathrm{f}|c_2) \times p(c_2) \ \ldots (1)$$

$$p(c_2|\mathrm{f}) = \mathrm{p}(\mathrm{f}|c_2) \times p(c_2) \Big/ \mathrm{p}(\mathrm{f}|c_1) \times p(c_1) + \mathrm{p}(\mathrm{f}|c_2) \times p(c_2) \ \ldots (2)$$

From equation 1 & 2

$$\boldsymbol{p(c_1|\mathrm{f}) = \mathrm{p}(\mathrm{f}|c_1) \times p(c_1)}$$
$$\boldsymbol{p(c_2|\mathrm{f}) = \mathrm{p}(\mathrm{f}|c_2) \times p(c_2)}$$

Sometimes it can be evaluated using the probability distribution function

We will estimate $\mathbf{p}(\mathbf{f}|\boldsymbol{c_1})$& $\mathbf{p}(\mathbf{f}|\boldsymbol{c_2})$with the help of Multivariate Normal Distribution.

Multivariate Normal Distribution:

$$f(\underline{x}) = \frac{1}{(\sqrt{2})^M |\sum|^{1/2}} e^{\{-\frac{1}{2}(\underline{x}-\underline{\mu})' \sum^{-1} (\underline{x}-\underline{\mu})\}}$$

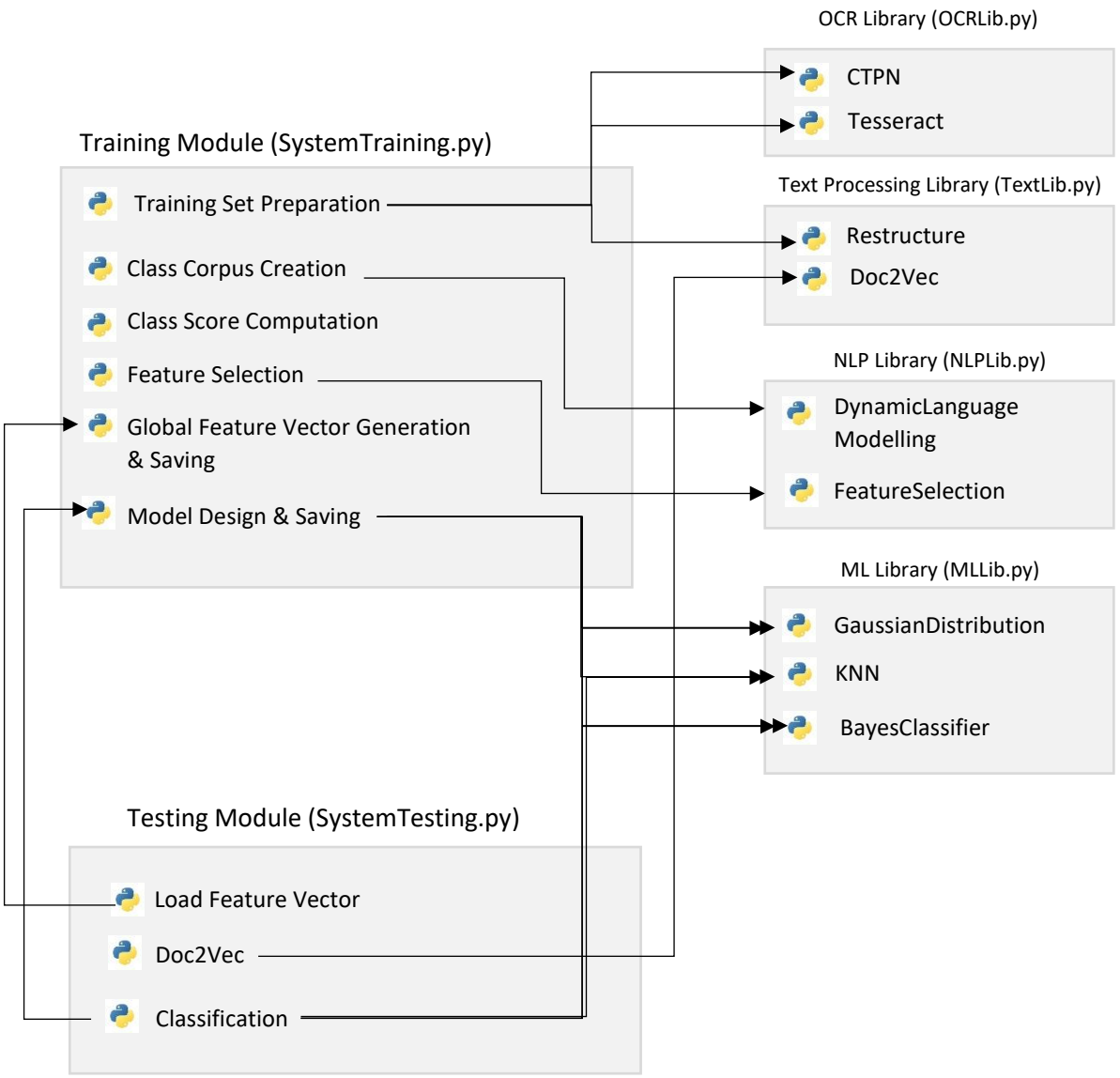
$\underline{x}$: test vector and $\underline{x} \in R^M$

M : Dimension of the data

$\sum$:Co-variance Matrix, $|\sum|$: determinant of $\sum$

$\underline{\mu}$: mean vector

Implementation Details



Documents for Classification

---

1. USA Loan documents:

a. **Closing Disclosure(CD)**: It is a five-page form that provides final details about the mortgage loan you have selected. It includes the loan terms, your projected monthly payments, and how much you will pay in fees and other costs to get your mortgage (closing costs)

b. **Loan Estimate(LE):**
- It is a three-page form that you receive after applying for a mortgage. The Loan Estimate tells you important details about the loan you have requested.
- The lender must provide you with a Loan Estimate within three business days of receiving your application
- The form provides you with important information, including the estimated interest rate, monthly payment, and total closing costs for the loan.

c. **Schedule K-1 (Form 1120S):** It is a source document that is prepared by a corporation as part of the filing of their tax return (Form 1120S)

d. **Schedule K-1 (Form 1065):**
- Partnerships file an information return to report their income, gains, losses, deductions, credits, etc.
- A partnership does not pay tax on its income but "passes through" any profits or losses to its partners. Partners must include partnership items on their tax or information returns.

e. **Schedule K-1 (Form 1041):**
- The income, deductions, gains, losses, etc. of the estate or trust.
- The income that is either accumulated or held for future distribution or distributed currently to the beneficiaries.
- Any income tax liability of the estate or trust.
- Employment taxes on wages paid to household employees

f. **Form 2210:** Use Form 2210 to see if you owe a penalty for underpaying your estimated tax and, if you do, to figure the amount of the penalty.

2. Financial Documents(India):

1. Form 16A
2. Form 16B
3. Form 26AS
4. ITR
5. Bank Statement

3. KYC Documents Classification:

1. Aadhaar Card
2. Passport
3. Voter's ID Card
4. PAN Card
5. Driving License

NOTE: We tried with text classification but didn't achieve a good result so decided to go with deep learning algorithms(resnet50)

4. Educational Documents:

1. Diploma certificate.
2. Degree certificate.
3. HSC certificate.
4. SSLC certificate.
5. Post Graduation

5. Discharge Summary Classification:

1. Cancer
2. Kidney
3. Diabetes
4. Malaria
5. Heart disease
6. Accidents
7. pneumonia
8. Stroke
9. Liver
10. Lung