Problem statement :-

① client wants to check if customers are satisfied with product or not. If not then what is reason of that (Root cause Analysis).

Data Extraction → ① Rating, Reviews → column
② Sales data

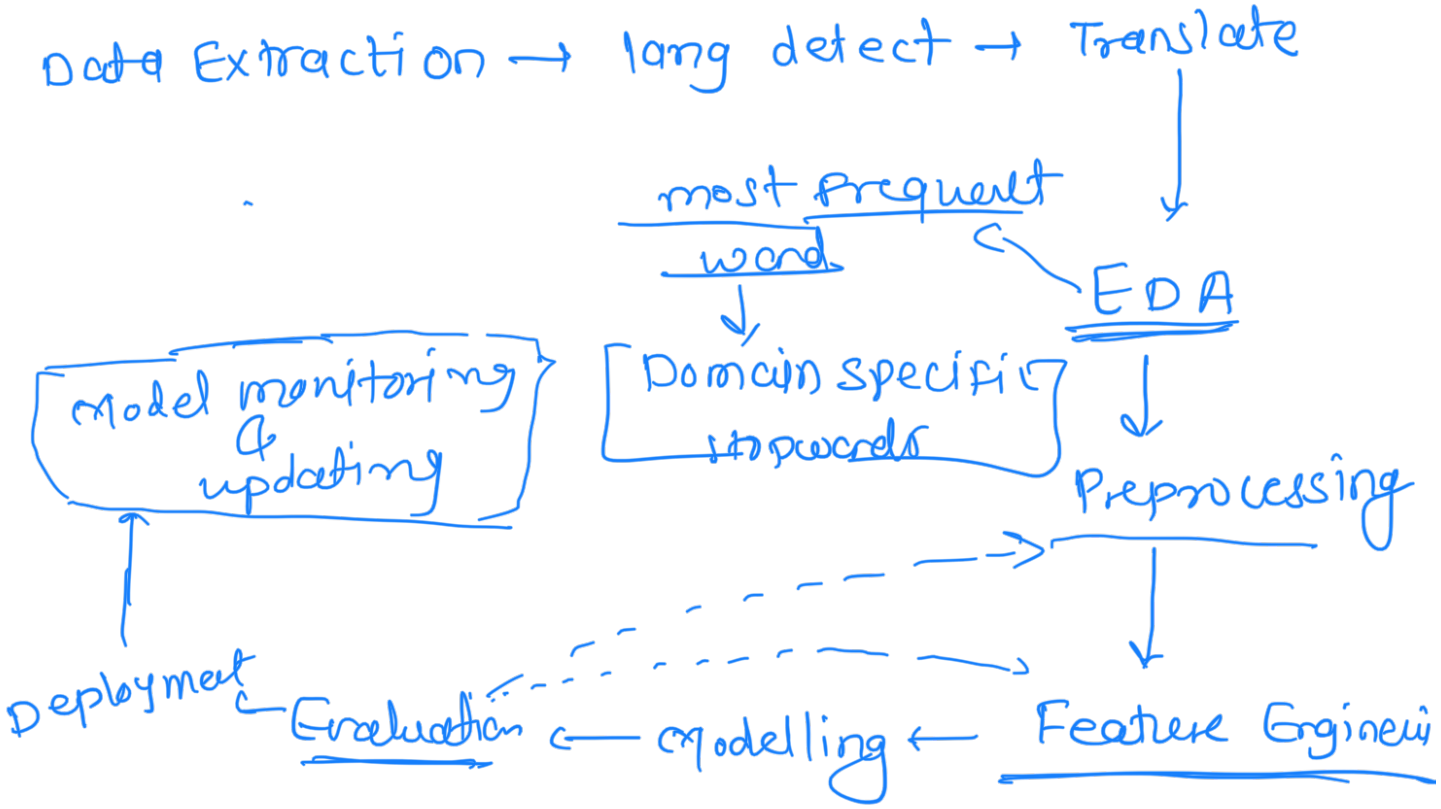
→ ① Feedback forms → column
→ ② Audio files → column

How to merge data ?

EDA → M.L → Uni
Bi
multivariate Analysis

NLP → Ngrams → visualize
→ keyphrases →
→ Word cloud →

reviews = multiple language



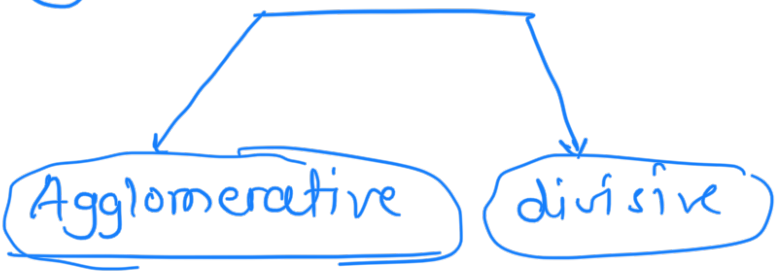
Target column →

1. clustering
2. creating target column with help of other columns in dataset

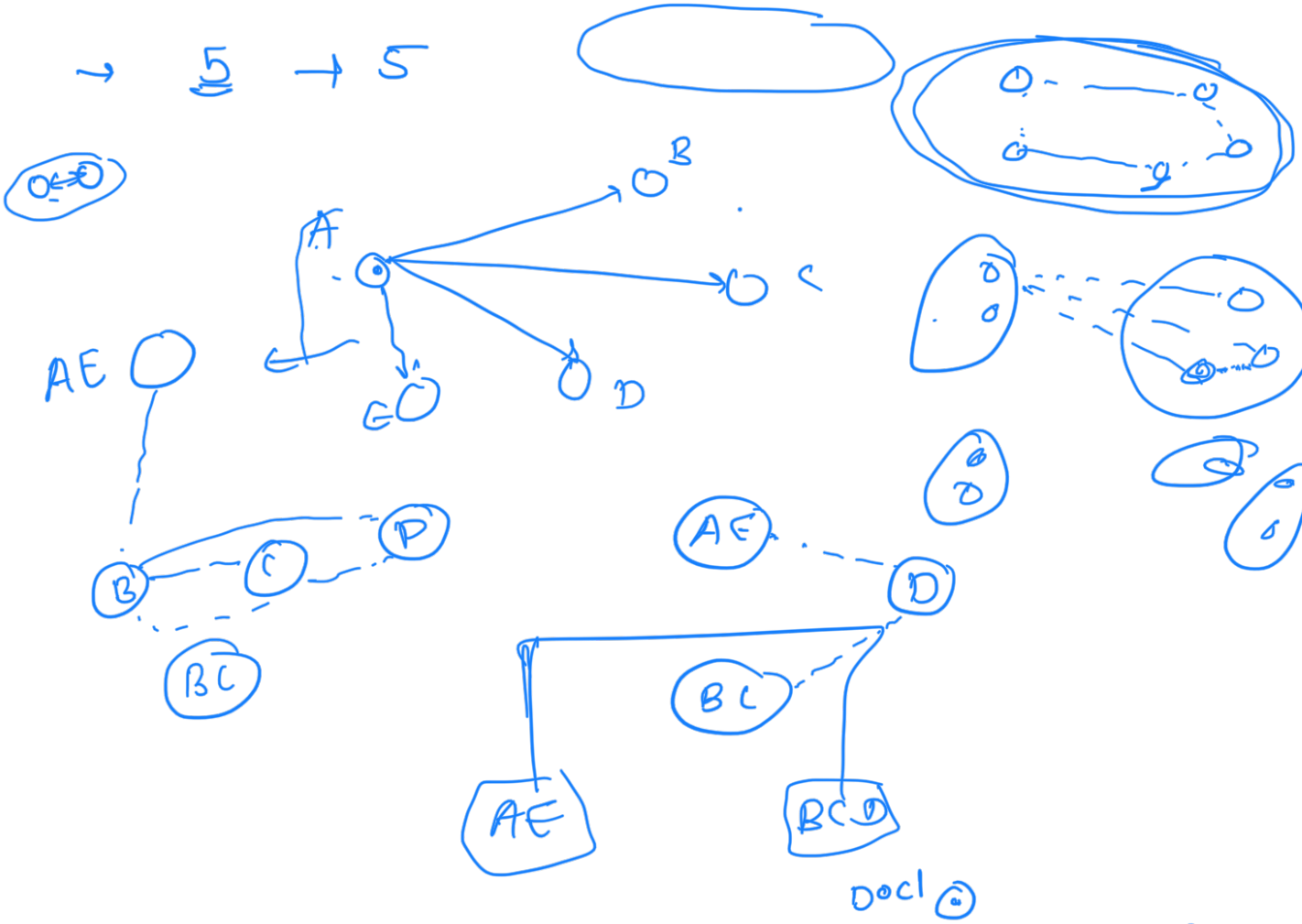
Dataset → Review
Rating 1 - 10.
Target → 1-3 → Neg
→ 3-6 → Neutral
→ 6> → Pos.

clustering → 1. KMeans

↳ ② Hierarchical

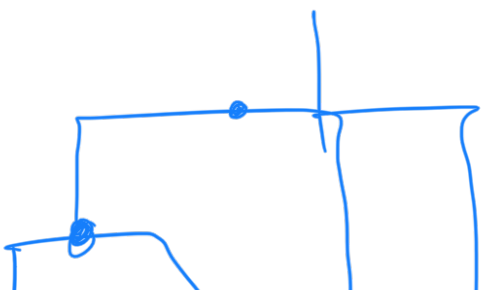


→ 5 → 5

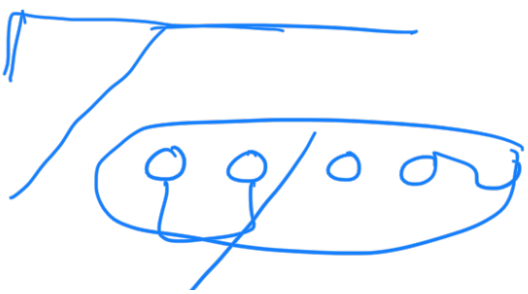


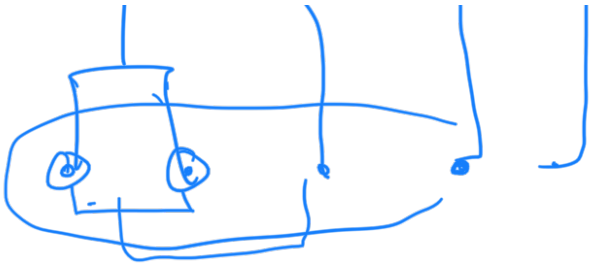
Text → Numerical

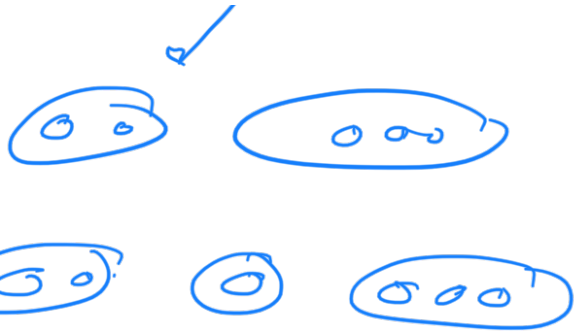
Agglomerative



Divisive







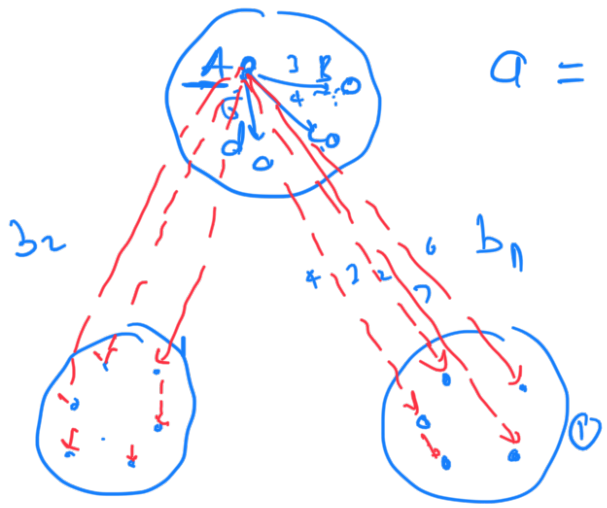
Evaluation metrics for clustering

' Silhoette score' = -1 to +1

(a)

$$= \frac{(b-a)}{\max(a,b)}$$

a =



$$a = \frac{3+4+5}{3} =$$

b =

b_2

b_1

$$b_1 = \frac{4+3+2+6}{4}$$

$$b_2 = \frac{7+2+6+6}{4}$$

$b = \min(b_1, b_2)$ =

$$score = \frac{b - a}{\max(a, b)}$$

a = Cohesion = mi

b = separation

$$= -1 \text{ to } +1$$

Yellowbricks → Silhoetter visualizer

- kMeans → count, TFIDF, wordwec
- Hierarchical → count, TFIDF, Word2Vec

② Document classification

→ Domain

- Health → Supply chain manager
- Insurance
- Banking
- Retail

Insurance → Aadhar, PAN, Doc, report, POC, Identification

Column → Civil score, Proof, RC Book

DOC → Training data

Data Extraction

↳ Image → pytesseract, openCV

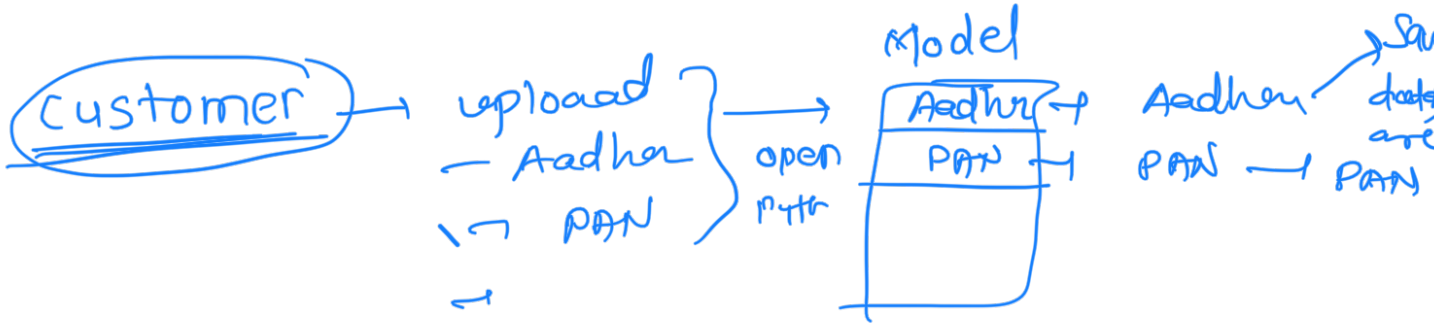
↓

Regular Expression

↓

Text	label
Aadhar text	Aadhar
PAN	PAN

client



Aah → model.predict (Text_data) = PAN card

Health domain → US

→ doc → Insurance

Health recept

→ Identification group

Key phrases = ① Poor customer service,
Faulty product,
② damaged product,
Bad quality of product,
Customer service is not
good,]

Ontology creation

①
{ "Poor Customer Service": [" customer service (keyphrase)
is not good",
10 ___],
" damaged Product": [" Bad quality of
Product",
" Faulty Product"
}

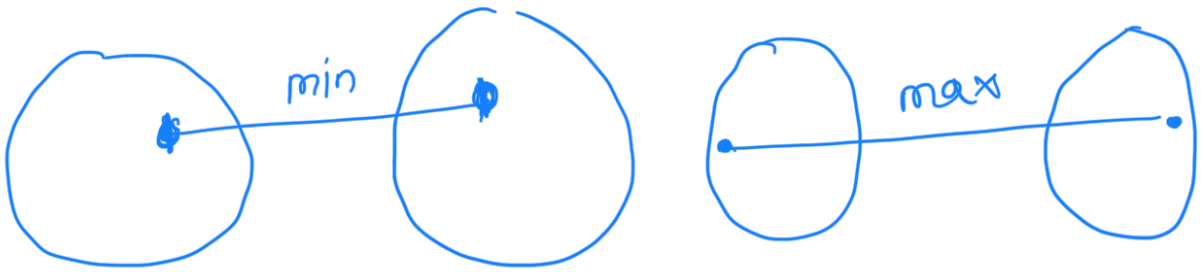
Mapping = Training data.

' Text "   " Root Cause "
[Faulty Pro.] = damaged Product

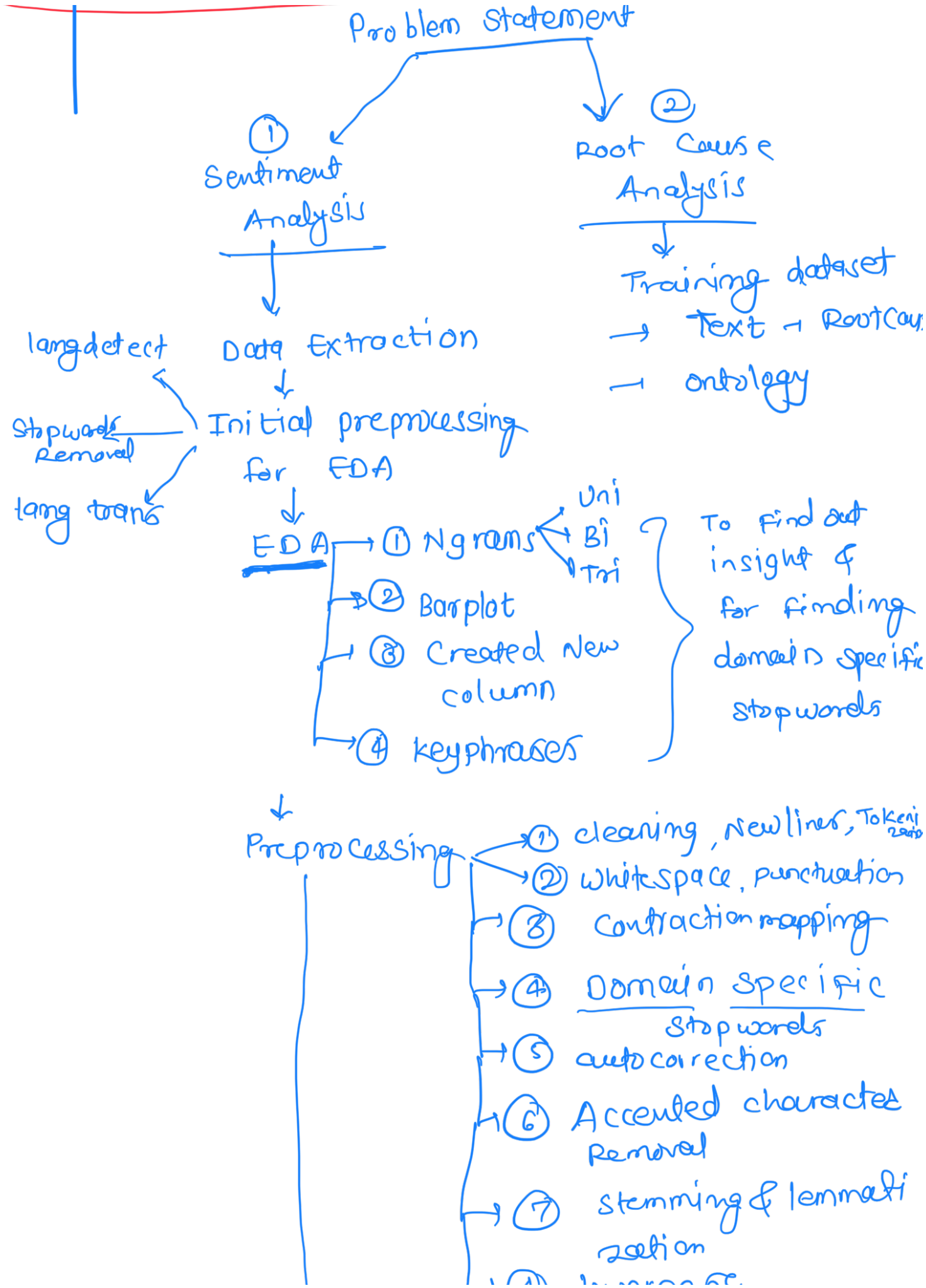
Pro

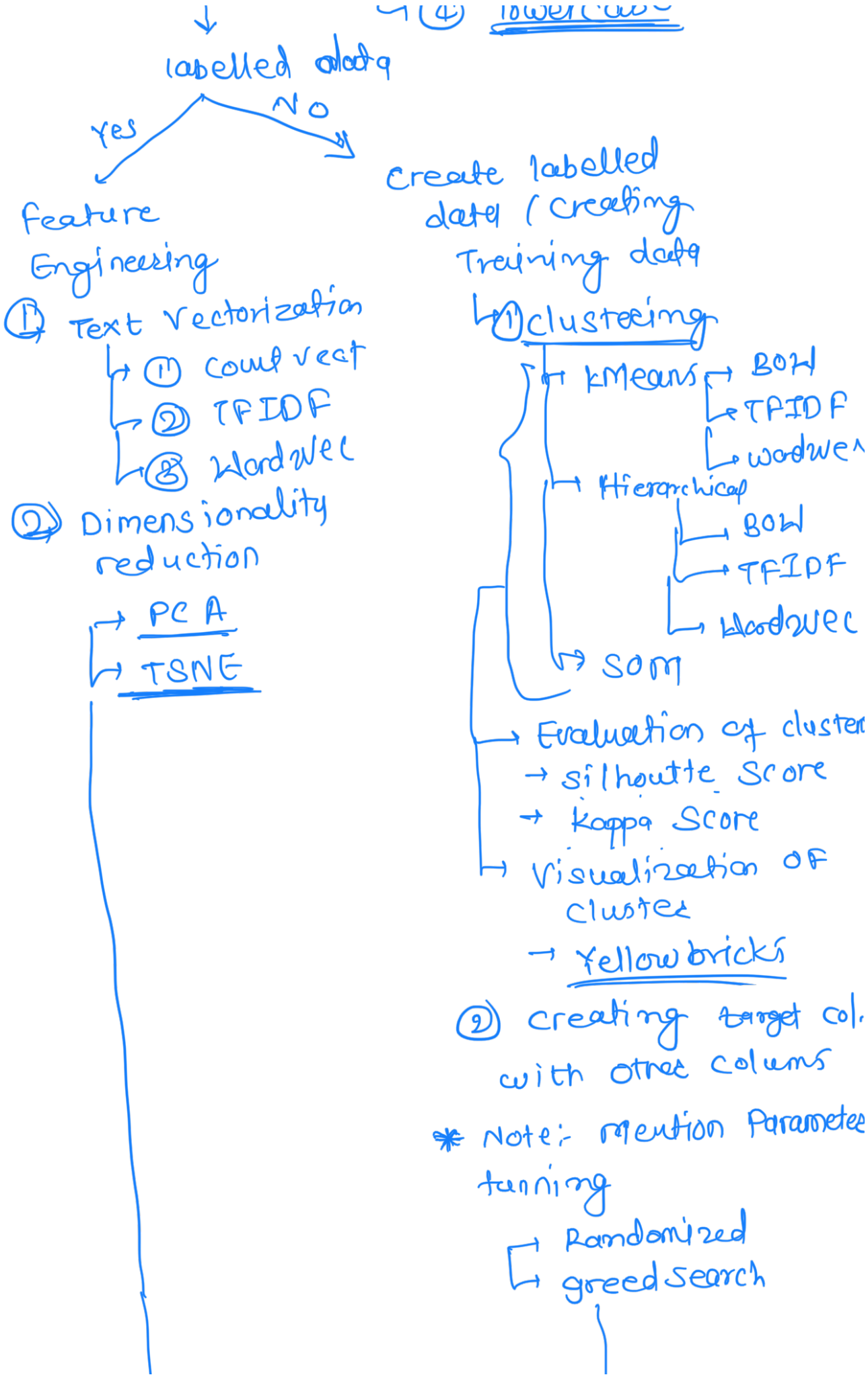
Text	YAKE	RAKE	Page
	—	—	— → Keyph

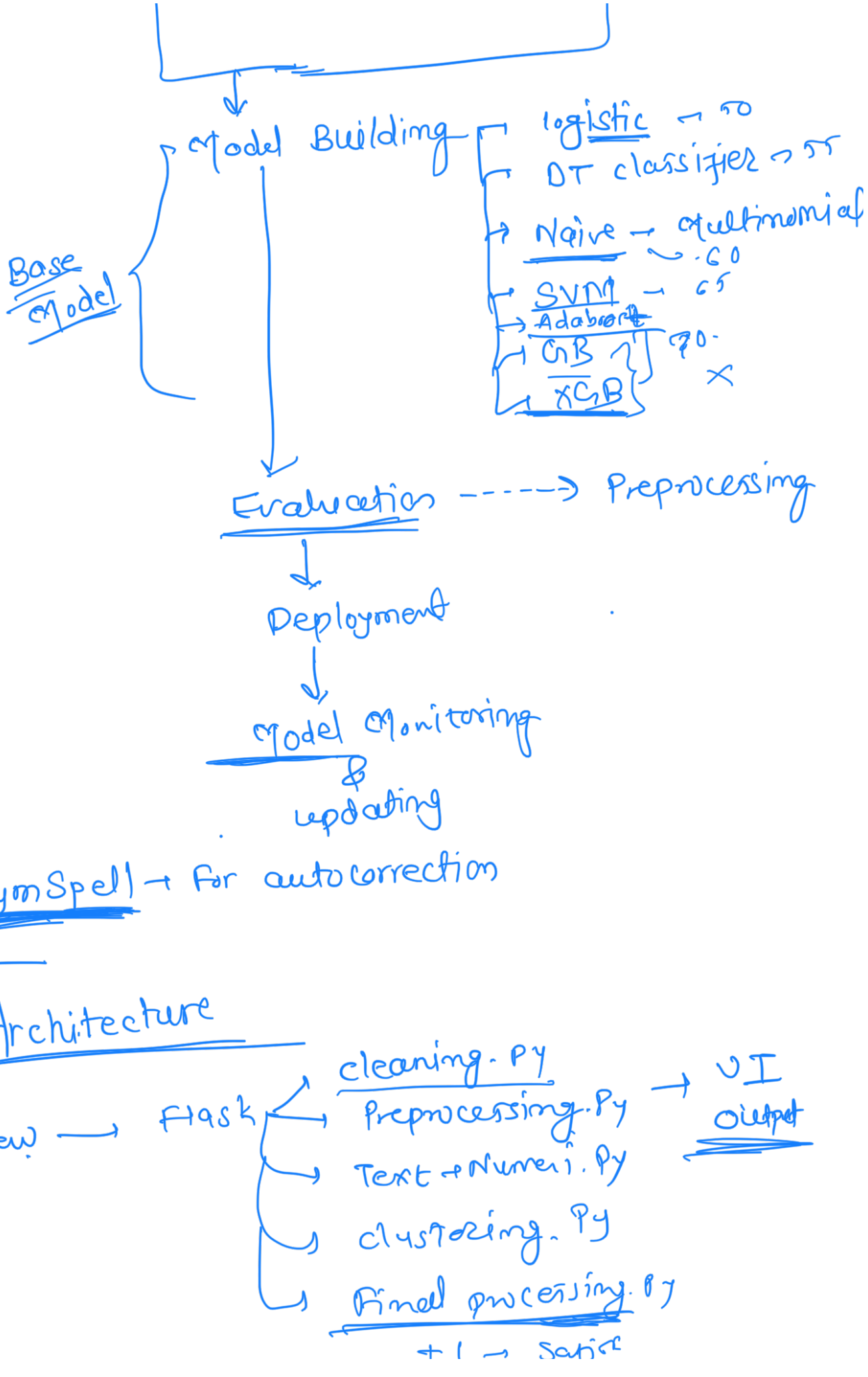
linkage= single , complete → max



Average

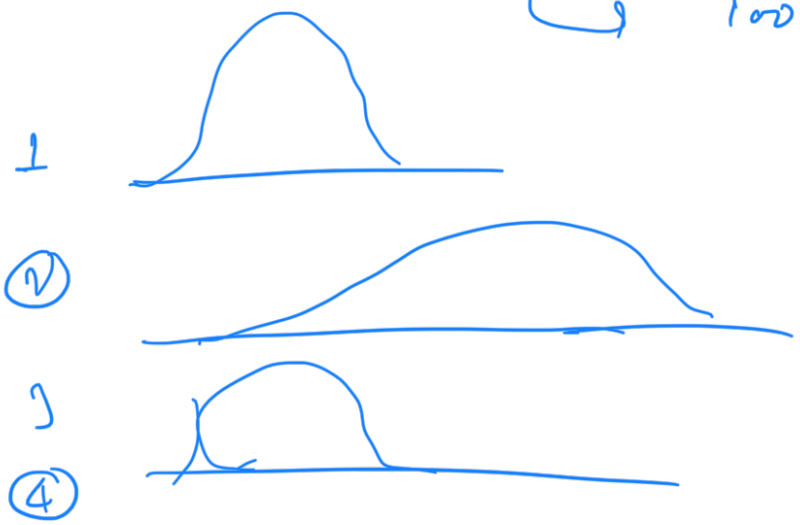

Problem Statement
1
Sentiment Analysis
2
Root Cause Analysis
Training dataset
→ Text → RootCau
→ ontology
Data Extraction
langdetect
Stopwords Removal
lang trans
Initial preprocessing for EDA
EDA
1 Ngrams
Uni
Bi
Tri
2 Barplot
3 Created New column
4 keyphrases
To find out insight & for finding domain specific stopwords
Preprocessing
1 cleaning, newliner, Tokeni zatio
2 whitespace, punctuation
3 contraction mapping
4 Domain Specific Stopwords
5 autocorrection
6 Accented characte Removal
7 stemming & lemmati zation


④ lowercase
labelled data
Yes
No
Feature Engineering
① Text Vectorization
① count vect
② TFIDF
③ Word2Vec
② Dimensionality reduction
PCA
TSNE
Create labelled data (creating Training data
① clustering
KMeans
BOW
TFIDF
word2vec
Hierarchical
BOW
TFIDF
Word2Vec
SOM
Evaluation of cluster
→ silhoutte Score
→ Kappa Score
Visualization of clusters
→ Yellow bricks
② creating target col. with other columns
* Note: mention Parameter tuning
Randomized
greedsearch


Model Building
Base Model
logistic → 70
DT classifier → 77
Naive → Multinomial
.60
SVM → 65
Adaboost
GB
XGB
70.
X
Evaluation ----> Preprocessing
Deployment
Model Monitoring & updating
* SymSpell → for autocorrection
Architecture
UI Review
Flask
cleaning.py
Preprocessing.py
Text → Numeri.py
clustering.py
Final processing.py
→ UI output

* Monitoring & Model updating :-

data drift → Column = Normally
↳ Skewed
Column = 30-40
100-150
} Model Impact



→ Model Switching → Adaboost → 85
SVM → 80
Naive → 75

Past
↓
Dr. Data drift → Adaboost → 50
SVM → 85
Naive = 70

script → max (SVM, Naive, Adal)

Jump

Jupyter = Ada → Pickle, joblib

Flask

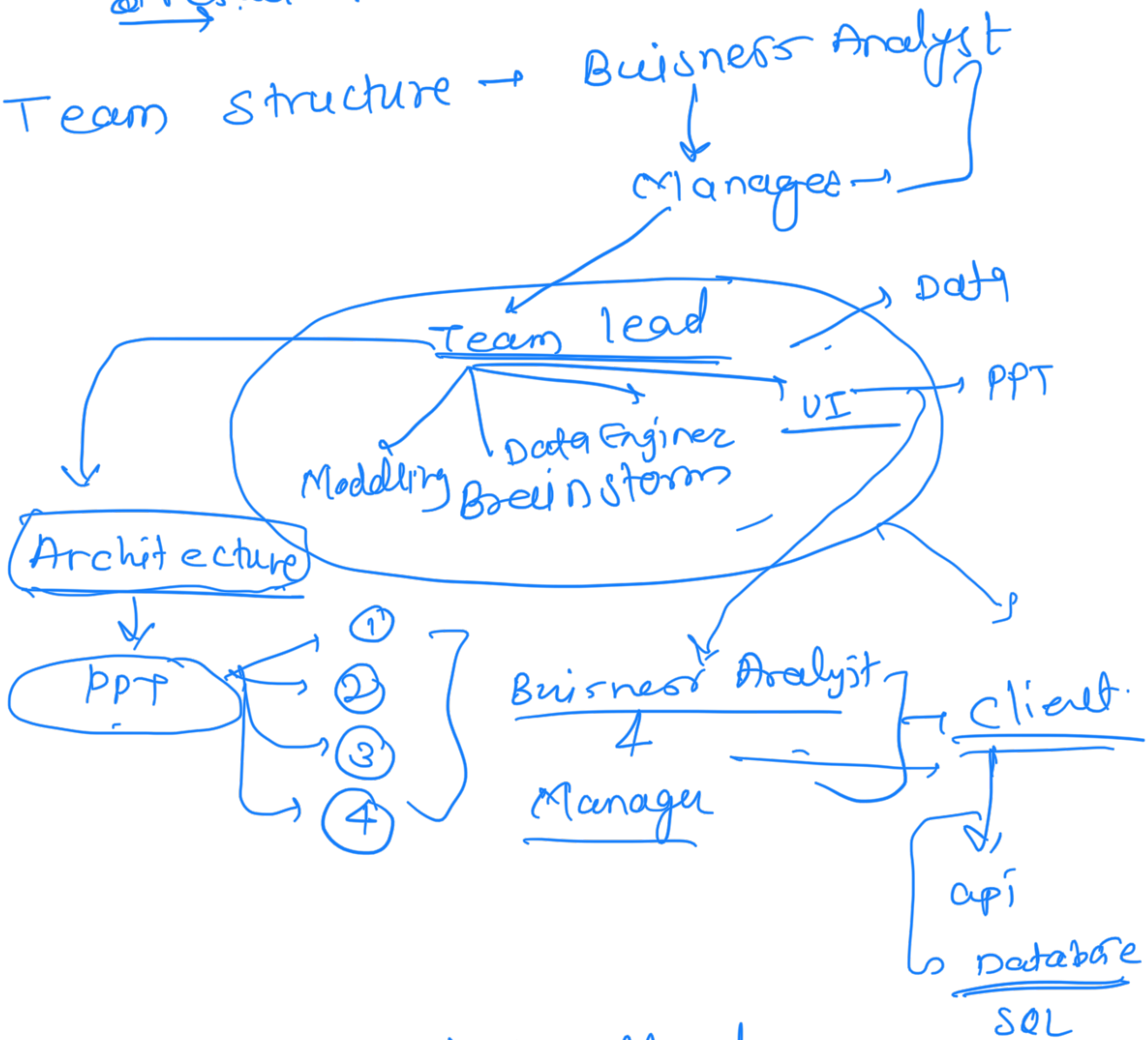
@ app

import cleaning, clustering -- etc

_ cleaning(data)

cln

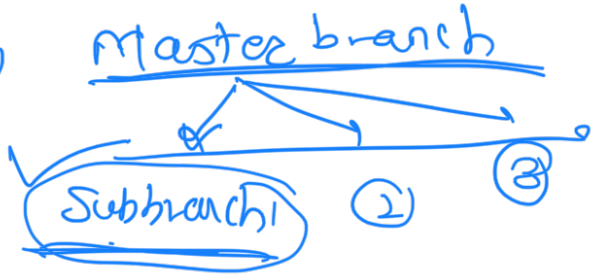
@ result =



Data → Cycles → Hourly
→ Quarterly →
Monthly
Yearly

Scheduler → api. cloud → data saved

git →



Manager / Teamlead
Code Review
or
Peer Review