

1. Data Collection

Source: KCC Dataset (data.gov.in) & KVK
(Punjab & UP > 40K rows, CSV)
Features: Query Type & Query Text

2. Initial Preprocessing

Identify Duplicates, Incomplete
sentences, Empty column

3. Data Filtering

Drop Weather Queries

Drop Duplicates

Drop Government Scheme
queries

Drop contact info queries

Drop Farming Fair queries

Drop Animal & Flower queries

4. Expert Review & Annotation

Consult with Domain Experts define
slot values

5. Output

Clean Annotated
Data

6. Integration

Feed into RASA NLU
module for training

