1. Data Acquisition
   1. From where would you acquire the data?

From web scraping

1. Text Preparation
   1. What kind of cleaning steps would you perform?

Html tags

* 1. What text preprocessing step would you apply?

lemetization/stemming ,

* 1. Is advanced text preprocessing required?
  2. Maybe pos tagging would be useful

1. Feature Engineering
   1. What kind of features would you create?
2. Modelling
   1. What algorithm would you use to solve the problem at hand?
   2. What intrinsic evaluation metrics would you use?

All kpi confusion matrix perplexity

* 1. What extrinsic evaluation metrics would you use?

How many of users interacted with both types of post

1. Deployment
   1. How would you deploy your solution into the entire product?
   2. How and what things will you monitor?
   3. What would be your model update strategy?

| **Stage** | **Summary** |
| --- | --- |
| Data Acquisition | Quora Question Pairs dataset |
| Text Preparation | Clean, tokenize, lemmatize, remove stopwords, use NER, POS if needed |
| Feature Engineering | Overlap metrics, TF-IDF, embeddings (e.g., SBERT) |
| Modelling | Siamese networks / BERT-based sentence similarity models |
| Evaluation | Intrinsic (F1, ROC-AUC), Extrinsic (CTR, user satisfaction) |
| Deployment | API-based microservice, monitor drift/errors, update model periodically |