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

A] Possibly from the Kaggle - **Similar Sentences Clustered Data**

The data is mostly a collection of the research articles and Us documents thus it might not be able to categorize other generalized topics

Or perhaps this one:  
**Textual Similarity -from Kaggle**

It has 2 paragraphs marked 0 or 1 based on their similarity also the paragraphs are random topics… might be more useful in the case of Quora .but these are paragraphs and not sentences ..

Can we do web scrapping for the newspaper articles covering the same topics .. and perhaps compare them but Ig it might be difficult to sort them out..

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

A] tokenization spell check,

* 1. What text preprocessing step would you apply?
  2. Is advanced text preprocessing required?

POS ,parsing

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?
   3. What extrinsic evaluation metrics would you use?
3. 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?