Things to work on:

1. Define similarity
2. Clustering, silhouette scores
3. Clean data further
   1. objective-> subjective (probably use Venky bhaiya’s repo)
      1. Mcq append
      2. Fill in the blanks
      3. Square bracket
   2. Letter-gap-letter-gap
4. Apply all QQD things to EM
5. SERP snippets for answers
6. bucketing
7. SSMBA:<https://drive.google.com/file/d/1JqzyUugIsCpgF4gXuW0qZ3Tx3qWKmwB7/view?usp=sharing>

Workflows

Workflow 1

Got the answers in a separate column -> need to append them :

- can simply concatenate

- or use venky bhaiyas repo

=> Add a column for SERP answers => apply all the QQD things to EM

Workflow 2

Start silo scores

Workflow 3

See SSMBA

Workflow 4

Combine all of the above to get some result at least

1. Understood the working of obj2subj repo - Link [here](https://github.com/VenkteshV/obj2subj_updated)
2. Played around with SERP results. Probably snippets will be useful for us. Colab [here](https://colab.research.google.com/drive/1eAQ6Q0vmLtMatP13W2ojJ0BwQ4OAMosg?usp=sharing)
3. SSMBA notebook

Things to discuss

1.Update on Obj2subj: The obj2subj repo even though gives correctly phrased questions, they are not exactly the same questions as before. The problem statement for the current project is to find duplicate questions, and if we look closely at the duplicates provided by the academicians we may notice that the final aim is to filter out questions with the exact same meaning. Since the questions generated by obj2subj seem to have different meaning after being processed, I am a bit skeptical towards this approach.

2. Can;t we somehow train a model which separately evaluates questions by their duplicity and then if two questions are similar syntactically in question, we check their answer similarity ? We give output as 1 if both give 1.

3. Discuss with Mohania sir about using T5 model (only the first 2 steps basically) ob o2s ?

1. Assertion-reason questions
2. Buckets ready => maybe we give 1-2 k of duplicate pairs and get them labelled ?
3. Rule based => Embeddings.

MOM:

1. AWS setup - done
2. Discussion on O2S
3. Discussion on SSMBA

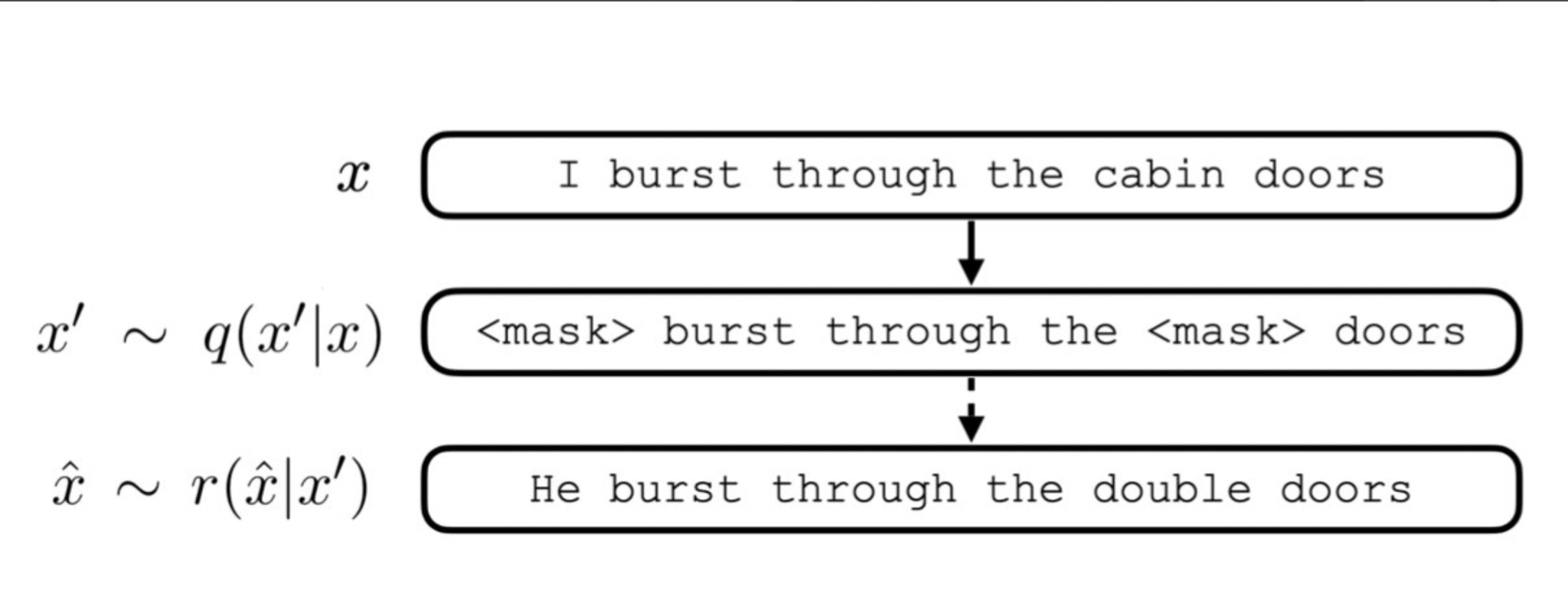
Things to do urgently :

1. Test out the “same answer” model
2. Silo score complete + bucketing + doubtful questions to be sent to acads
3. Use o2s on declarative-sentence questions only. See the result
4. If good result in 3, try filtering out multioptinal questions in EM dataset

**Literature Review**

1. **SSMBA**

The whole idea of SSMBA is detailed in the image below. Corrupt the data with MLM -> reconstruct with BERT -> new examples pseudo-labeled with the original labels as before.

****

1. **O2S paper (Venky bhaiya)**

Questions are of three types:

1. Multioptinal (*choose one of the following*, or *choose the statement*)
2. WhWord based
3. Declarative sentence

This paper deals with declarative-sentence questions. Applies rule-based templates (like sentences ending with “by”, “on” ) i.e. looking at the last 1 or 2 tokens => using PoS rules => PAA => T5 model by concatenating the answer and highlighting it as the answer => rank the questions using a pre-trained ranking model (msmarco-distilroberta-base-v22

Conclusion: works well with declarative sentences. Try applying this model to the declarative questions