1) Approach to the problem

- high level description of solution

- pre-processing

- feature extraction

- models

- alternative approaches

2) Evaluation of results

- split 80/20

- logloss

- succeeded/failed tests

- other metrics

Pre-processing

- the text should be cleaned beforehand

- be careful with overcleaning

*The main thing to remember when you are cleaning data is to remove as much 'noise' from it as possible. In this case, noise is anything that doesn't help to prove if pairs of questions are similar or not: punctuation, over common words (the, a, an, of...), limiting the variation of words (What, what, He, he).*

- normalize as much as possible!

- be aware of the risks/drawbacks…

- find a trade-off

Feature extraction