Domain and Context:

[Quora](https://www.quora.com/) is a platform that empowers people to learn from each other. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers. A key challenge is to label questions to relevant topics in order to improve search results as well as user feeds.

Problem Statement:

When a user adds a question to Quora, automatically suggest topics which might be relevant to the question. For example, a question about Dijkstra's algorithm would probably fit well under the topics "Algorithms" and "Graph Theory".

While most topic modelling applications deal with a large document text and a small topic ontology, we will work with a short question text and [more than a million potential topics](https://data.quora.com/The-Quora-Topic-Network-1) to tag on the question, which makes it a much more challenging problem to solve.

The following insights are expected to be drawn after performing labelling:

* Distinguish between high-quality and low-quality questions
* Improve related questions suggestions to users
* Find out the labels which are disparaging or inflammatory with respect to any particular religion, region or ethnicity.

Dataset: [Source🡪 https://www.kaggle.com/c/quora-insincere-questions-classification/data](https://www.kaggle.com/c/quora-insincere-questions-classification/data)

|  |  |
| --- | --- |
| **File descriptions** | **Data Fields** |
| * **train.csv** - the training set * **test.csv** - the test set * **sample\_submission.csv** - A sample submission in the correct format * **enbeddings/** - (see below) * GoogleNews-vectors-negative300 - <https://code.google.com/archive/p/word2vec/> * glove.840B.300d - <https://nlp.stanford.edu/projects/glove/> * paragram\_300\_sl999 - <https://cogcomp.org/page/resource_view/106> * wiki-news-300d-1M - <https://fasttext.cc/docs/en/english-vectors.html> | * **qid** - unique question identifier * **question\_text** - Quora question text |