***Dataset***

The dataset is composed of over 800,000 questions and answers extracted from the [*StackExchange*](https://protect.checkpoint.com/v2/r01/___https://data.stackexchange.com/___.YzJ1OmZhbXU6YzpvOjc1MDYzZWIyYjYxYmFjZWY4YmZhYTBjOGRmODM4YTYwOjc6MGMxYzo4OWQ4NDA3NDViYmRlMWFkMzg4ZWMyNzYxMGQ5YTY4ZTA0NDM3MzQ3NGRlMTMzYzRmNDkzNmEyZTBhNGEzYWIwOnA6VDpU) website for the [*Cross Validated*](https://protect.checkpoint.com/v2/r01/___https://stats.stackexchange.com/___.YzJ1OmZhbXU6YzpvOjc1MDYzZWIyYjYxYmFjZWY4YmZhYTBjOGRmODM4YTYwOjc6YzY2NDpiYjJlOTliMGQ4YmYxYzI5OGJjMmEyZWIzOTE1MzEyZDFkMDA4Njc2ZjA0ZWU0OThkNjA3MTkyYmExNGRlNzJkOnA6VDpU) site. The Cross Validated site is defined as “Q&A for people interested in statistics, machine learning, data analysis, data mining, and data visualization and mostly revolves around statistics and machine learning.”

The file has five variables:

* post\_id: A post can be a question or an answer to the original question. Original questions have a title. Answers do not have a title.
* parent\_id: Posts with a parent\_id are answers to the original post.
* comment\_id: Comments
* text: A title, post, or comment
* category:
  + **title**: Titles are usually one sentence long and may be similar to the following examples:
    - Question on how to construct a confidence interval
    - Choosing model from Walk-Forward CV for Time Series
  + **post**: HTML-formatted paragraphs that contain the following:
    - HTML tags
    - URLs
    - math equations in LaTeX
    - sometimes large tables of numbers
  + **comment**: Comments are shorter than posts. They are simple text paragraphs that can also contain URLs, math equations, and numbers but no HTML tags.

The following aspects make this dataset particularly relevant to building a language model:

* It is domain focused; the domain chosen here ranges from statistics to machine learning.
* With over 800,000 pieces of text, it is also large enough to build a decent language model.
* It contains text of different natures and formats. These different natures make it possible for us to use titles as a separate testing subset and use questions and answers as the main training dataset. We can also use titles for query completion since they are short and to the point.
* The noisy nature of the questions (HTML, LaTeX, URLs, and so on) creates the need for a preprocessing phase for the text. A preprocessing step can be iteratively refined to improve the performance of our language models.

**Note**: This dataset is large and representative of real-world datasets. As such, it provides a lot of flexibility to exploit it. In particular, you will probably have to drop some of the data to be able to implement a quick trial workflow. Reducing the dataset will allow your code to execute within memory constraints in a matter of minutes instead of hours. Finding the right balance between data reduction and script execution is an important part of this project.

The dataset is available [here](https://protect.checkpoint.com/v2/r01/___https://lp-prod-resources.s3.amazonaws.com/66*~*dtymjwdxyfhpjChmfslj_b67p.hxA.lE___.YzJ1OmZhbXU6YzpvOjc1MDYzZWIyYjYxYmFjZWY4YmZhYTBjOGRmODM4YTYwOjc6NzVlYjoyZWJlNDAyYTA3ZDFkYTZiN2ZiZjVmNThmZjAxOWZlNzg0ZjMyZWQ3OTE4ZTg1YWYyMGNhMDcyM2Q1YjdkMDU5OnA6VDpU).