Feature Engineering:

In order to describe the structures inherent in our data and explain the problem at hand, we update the dependent variable using logarithm transformation and we extract some meaningful features from the text and authors data.

Features Extraction:

To better understand our new features, we define the following elements:

* **Corpus:** is a list of unique tokens cited in all documents.
* **Author\_docs\_occurence:** For each author we count the occurrences of his papers in the whole file *authors\_papers.txt*.
* **Abstract\_ratio:** it is the ratio of tokens set (unique items) of each author’s abstract to its length. (formula)
* **Corpus\_ratio**: It describes the ratio of the tokens set in an abstract to the corpus’s length for each author.

Based on the elements above we create the next variables:

* **Single\_doc:** It is the number of documents that are not co-authored for each author.
* **Docs\_number**: it’s the number of documents written by each author.
* **Max, mean** and **min** of **Author\_docs\_occurence** of each author.
* **Max, mean** and **min** of **Corpus\_ratio** of each author.
* **Max, mean** and **min** of **Abstract\_ratio** of each author.
* **Co\_authored\_sum:** The sum of **Author\_docs\_occurence** that are co-authored (occurrence is different to 1) for each author.

Log transformation:

We use the function y🡪 log(y+1) to guarantee that **h\_index** is strictly positive. This transformation also helps to handle skewed data, reduce the effect of outliers. The 1 is added to avoid infinite values for **h\_index** equals to 0.