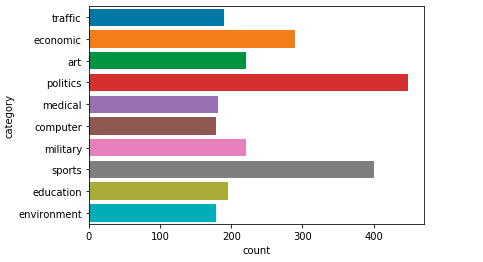
**Chinese Document Classification (News)**

# Data Cleaning and EDA

* Read the data from the source files
* Clean the data by removing the new line characters, stop words and punctuations
* Perform EDA of the cleaned data
  + Median, max and min sentences in a doc
  + Number of Document vs Sentence length plot
  + Word cloud to see the most important words in a class
  + Words frequency in the documents

## **EDA Results**

**Class distribution**

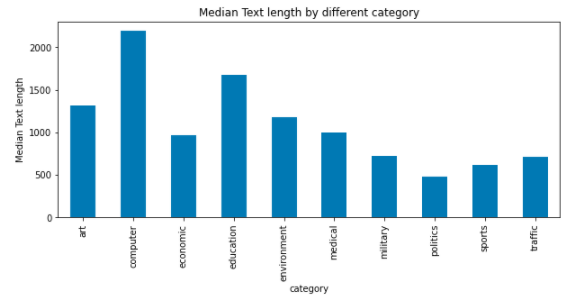


Politics and sports news are greater in number compared to others

**Number of characters & words in each type of news**

- The next step is the analysis of title and text columns in a group of news.

- Here, we wanted to check the character & word count distribution among different categories.



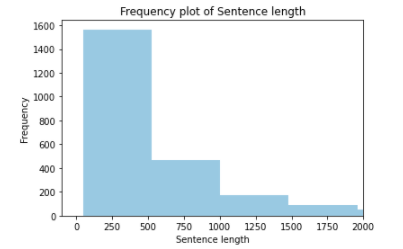
* Median text length of computer and education is quite high.
* Politics, Sports and traffic are on average comparatively smaller compared to others.

**EDA on sentence length after cleaning**

1. Median Sentence length: 396
2. Max Sentence length: 23946
3. Min Sentence length: 47
4. % of document with length greater than median words count : 49.92 %
5. % of document with length greater than 512 words count : 38.32 %

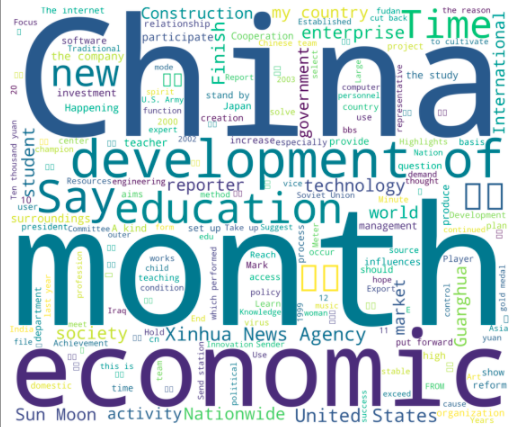
\*(These numbers are after removing punctuations and other spatial characters)

1. Number of Documents vs Sentence length plot



\*( Very large words removed for better visuals)

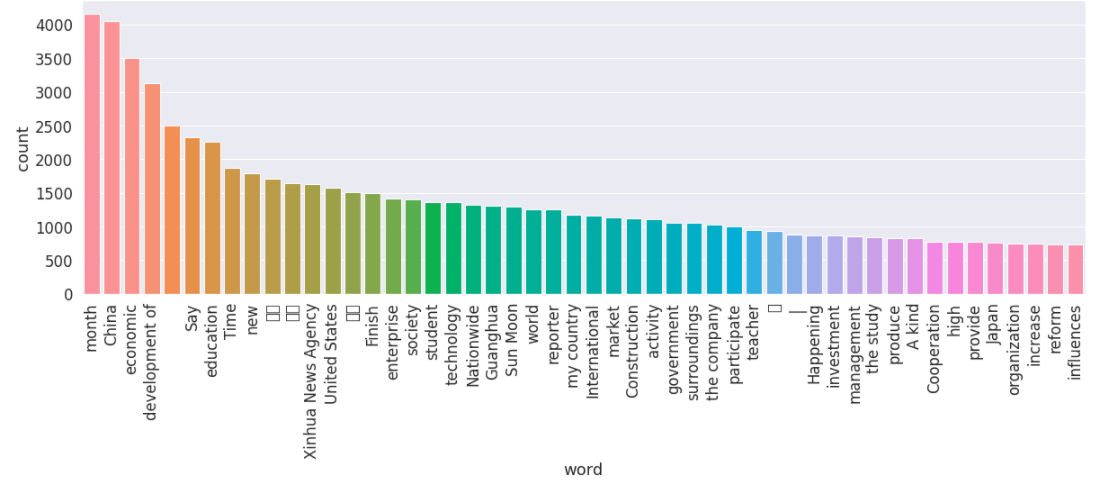
**Word cloud to see the most important words**



Chinese, Month, Economic and Development are most common words.

**Note: Separate Word-cloud for 10 categories are attached in excel file.**

**Words frequency in all documents**



Chinese, Month, Economic and Development are most common words, whereas words like society, student, technology are average occurring words.