dataset.csv: The dataset has 4 columns, please use pandas to load it.

1. class\_1 (int): the first class level of level-4 articles. Its values include people, arts, everyday life, history, etc. Totally 12 values. All of them are listed in the first line of classes.txt

2. class\_2 (int): the second class level. Its values include entertainers, visual artists, etc. Totally less than 250 values. All of them (not finished right now) are listed in the second line of classes.txt

3. title(str): the title of article. I add this column just let you can easily find the article in wiki if you need.

4. content: the content of article in sentences, the whole articles or paragraphs.

classes.txt: It has 3 lines

1. list all values of first class level (seperated by '|', all 12 values are included)

2. list all (not all right now) of second class level (seperated by '|', all 109 values are included)

3. content format:

'a': the whole article

'p': in paragraphs

's': in sentences

all of three versions can be found in "Group Dataset/### articles" folder. I will keep updating it.

Note: the value of class\_1, class\_2 is the index of corresponding class-level lists

eg. class\_1 = 0, class\_2 = 1 means this article's category is "arts/arts general"

Content: The following parts of articles are not included

title and subtitles

notations under images

reference information

The part "See Also"

There should still be many useless information in dataset.