# Document Layout Analysis

Chenrui Fan

May 12, 2024

## 1 Introduction

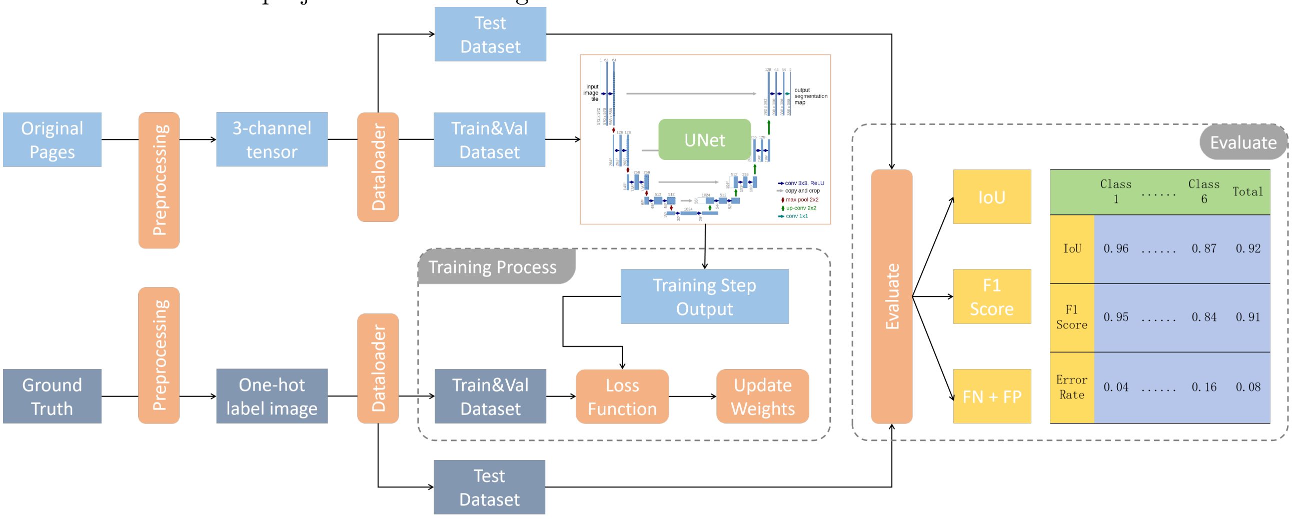
Document Layout Analysis is performed to determine the physical structure of a document, identifying document components. These components may include single connected regions—regions of pixels adjacent to form single areas or groups of text lines. A text line consists of characters, symbols, and words that are adjacent and relatively close to each other, and through which a straight line can usually be drawn, typically with horizontal or vertical orientation.

The dataset used in this project is a typical DLA task. It doesn't have a lot of noise or an overly complex layout, so it's great for novices to practice.

Github Link

## 2 Workflow

The workflow for this project is shown in Figure 1.



| | Class 1 | ...... | Class 6 | Total |
|---|---|---|---|---|
| IoU | 0.96 | ...... | 0.87 | 0.92 |
| F1 Score | 0.95 | ...... | 0.84 | 0.91 |
| Error Rate | 0.04 | ...... | 0.16 | 0.08 |

Figure 1: Workflow

## 3 Dataset Preprocessing

In the data preprocessing stage, I filter out the background part of the original image and remove the yellow (text box) part of the Mask. After that, I will denoise the original image using a Gaussian kernel or other methods. Since the image is too big (960x640), I will resize it to one-half of the original (480x320).

## 4 Dataset and Dataloader

I have divided the dataset into training sets, validation sets, and test sets in the ratio of 8:1:1. Among them, the test set is used to apply evaluation metrics to measure whether the model performs well or not.

For document images in JPG format, I read them as three channels of image data. Note that for Mask images in GIF format, their colors should not be used as labels (e.g., red is labeled [1, 0, 0]), but rather they should be encoded according to how many colors there are in total.

Here just a single-channel image can be used, where the integer value of each pixel directly represents the category label. For example, 0 for the background, 1 for the first category, 2 for the second category, etc.; or a One-hot encoded Mask can be used: in this case, the label of each pixel is encoded as a vector whose length is equal to the number of possible labels. For example, if there are three categories (background, category 1, category 2), then the label of each pixel will be a three-dimensional vector (in our task, there are six categories, excluding the yellow text box).

## 5 Model

I tested the performance of the UNet model and its two variants, UNet++ and UNetDense, on this task.

Originally I proposed the idea of using cross-layer connections in ResNet to solve the ”leakage” problem. However, after using the one-hot mask, the problem disappeared. However, I still tested the performance of my modified model. The results are shown in the Experiments section.

## 6 Training Process

During the training process, I use both cross-entropy loss at the pixel level and boundary loss at the edge level of the segmented image, but of course, the specifics will depend on the training results.

I accumulate the gradients in the training set, update the model, and judge the results in the validation set, based on the results to determine whether to save the model and then get the best-performing model in the validation set.

## 7 Metrics

This is the most confusing part of this task: how do we know that the model is performing well? In other words, what are the most common metrics used on image segmentation tasks?

**Pixel Accuracy** refers to the percentage of pixels in an image that are correctly categorized. While it looks great, and after all, we do use it as an optimization goal for training, it's not a good metric. Let's say, for example, that the model gives a segmented image that is broadly very similar to Mask, and calculates a Pixel Accuracy of 95% or more. But in reality, some keywords or vignettes need to be segmented that are not segmented at all - a phenomenon known as class imbalance, which suggests that high Pixel Accuracy doesn't always mean superior segmentation. This requires us to look for other metrics.

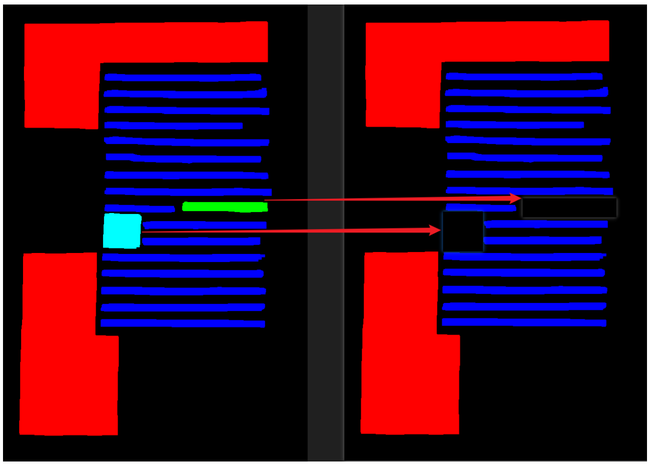


Figure 2: Class Imbalance

.

**Intersection-Over-Union (IoU)**, one of the most commonly used metrics in semantic segmentation, is calculated as the area of overlap between the predicted segmentation and the true situation divided by the area of concatenation between the predicted segmentation and the true situation. For binary (two-class) or multi-class segmentation, the average IoU of an image is computed by averaging the IoUs of each class (as in our task). This will give more attention to image blocks that are not very large but need to be segmented.

**Dice Coefficient (F1 Score)** refers to the 2 * overlap area divided by the total number of pixels in the two images. Dice coefficients are very similar to IoU. They are positively correlated, ranging from 0 to 1, where 1 indicates the greatest similarity between prediction and fact. But they are not the same thing: in general, the IoU metrics tend to quantitatively penalize individual instances of bad classification more than the F score, even if they both agree that the instance is bad. In the same way that L2 penalizes the largest errors more than L1, IoU metrics tend to have a "squaring" effect on errors relative to the F score. As a result, F scores tend to measure something closer to average performance, while IoU scores measure something closer to worst-case performance.

## 8 Experiment

I use a 4090 on my personal computer for training. The batch size is 1 (for easy visualization during training) and num of epochs is 300. A full training session takes about half an hour. The results are shown in Table 1:

Table 1: IoU and F1 score for different models across various classes

| **Model** | | | | IoU | | | |
|---|---|---|---|---|---|---|---|
| | Cyan | Magenta | Green | Red | Blue | Black | Total |
| UNet | 0.962 | 0.973 | 0.953 | 0.992 | **0.967** | **0.988** | 0.972 |
| UNet++ | **0.970** | 0.972 | 0.952 | **0.993** | 0.964 | **0.988** | **0.973** |
| UNetDense | 0.631 | 0.959 | 0.929 | 0.989 | 0.961 | 0.985 | 0.909 |
| UNetRes | 0.400 | **0.976** | **0.957** | 0.993 | **0.967** | 0.986 | 0.880 |
| | | | | F1 Score | | | |
| | Cyan | Magenta | Green | Red | Blue | Black | Total |
| UNet | 0.980 | 0.986 | 0.976 | **0.996** | **0.983** | **0.994** | 0.9858 |
| UNet++ | **0.985** | 0.986 | 0.975 | 0.996 | 0.982 | **0.994** | **0.9864** |
| UNetDense | 0.574 | 0.979 | 0.963 | 0.995 | 0.980 | 0.993 | 0.9138 |
| UNetRes | 0.400 | **0.988** | **0.978** | 0.996 | **0.983** | 0.993 | 0.8897 |

You can see that UNet and UNet++ perform similarly, UNetDense doesn't perform that well, and UNetRes, which I designed myself, performs poorly in predicting the first label but performs equal to or better than UNet in predicting the other labels. This suggests that the introduction of residual concatenation helped the segmentation task to some extent.

The loss, IoU, and F1 score during training UNet are shown in Fig 3:

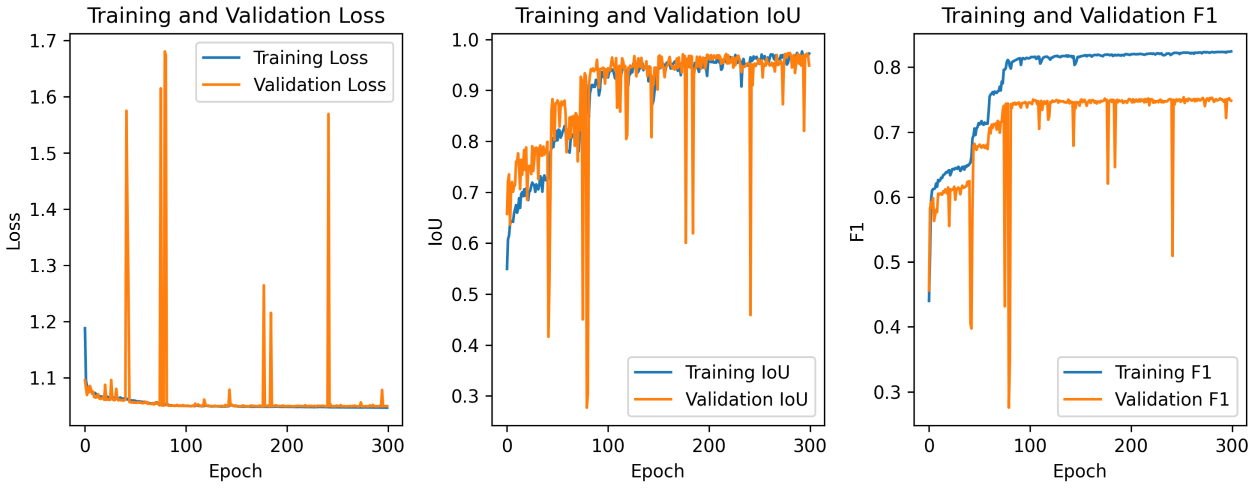


Figure 3: Training step

# 9 Result

I think the results are great and the visualization comes out as follows:

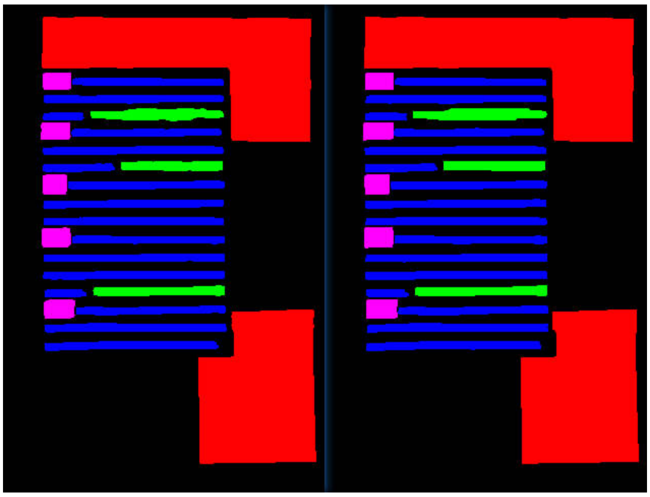


Figure 4: Prediction and groundtruth mask

# 10 Conclusion

This is the first time I've written all the code for a segment task step by step. It mainly includes reproducing UNet, improving UNet, reproducing UNet++, writing Dataset classes for the dataset, writing training and testing code, and so on.

The most important part is how to label the Mask image. I've used RGB as a label before, and the result is a "leakage" problem, as shown in the figure:

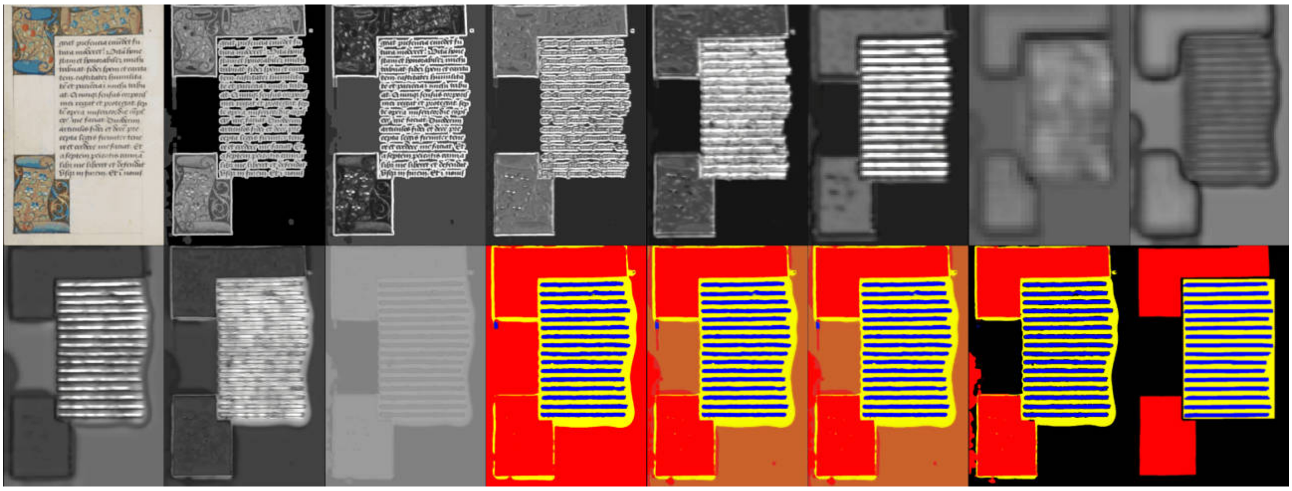
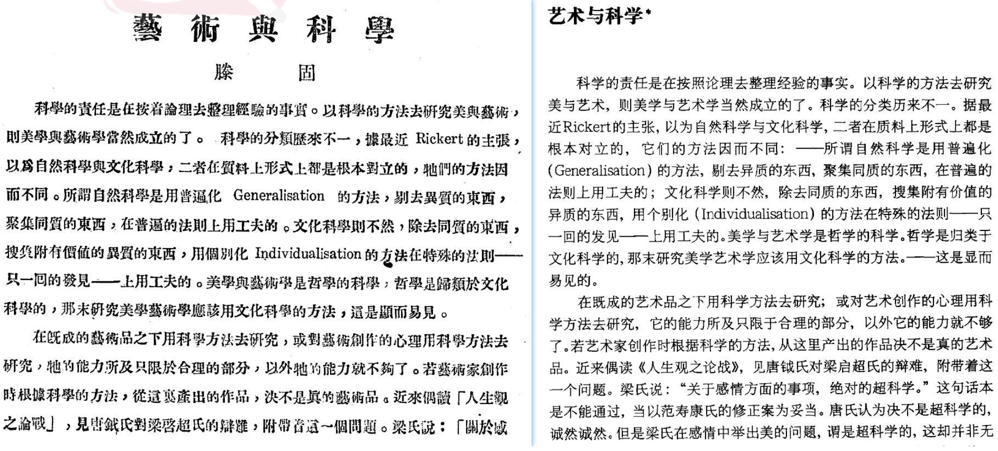


Figure 5: "Leakage" problem

Later on, I read other papers and realized that we can convert masks into one-hot-coded images. So I experimented, but the result is still very poor. After that, I restored the result and Mask to the original picture and found that Mask could not be restored normally, after checking for two days, I found that it was necessary to convert the GIF picture into an RGB picture after reading it, to generate the one-hot label. After the modification and training, the IoU quickly increased from 28% to 97%.

All in all, this project made me realize the importance of metrics and labels. Just by using the right one-hot label, you can get the most basic UNet to have an IoU and F1 score of 97% or more. Choosing good metrics is especially important - it lets you know how well the model is working, on which labels the predictions are not so good, what to improve next, etc.

On the other hand, this class gave me a comprehensive understanding of the tasks on Document Image Analysis. It also allowed me to work on the same problems quickly. Yesterday my sister was discussing a problem her professor encountered while doing document proofreading with me. Document proofreading involves manually comparing two versions of an article for minor differences, such as one or two-word changes, different punctuation, and so on. She provided me with the following document:



藝 術 與 科 學

滕 固

科學的責任是在按着論理去整理經驗的事實。以科學的方法去研究美與藝術，則美學與藝術學當然成立的了。 科學的分類歷來不一，據最近 Rickert 的主張，以爲自然科學與文化科學；二者在質料上形式上都是根本對立的，牠們的方法因而不同。所謂自然科學是用普遍化 Generalisation 的方法，剔去異質的東西，聚集同質的東西，在普遍的法則上用工夫的。文化科學則不然，除去同質的東西，搜集附有價值的異質的東西，用個別化 Individualisation 的方法在特殊的法則——只一囘的發見——上用工夫的。美學與藝術學是哲學的科學；哲學是歸類於文化科學的，那末研究美學藝術學應該用文化科學的方法，這是顯而易見。

在既成的藝術品之下用科學方法去研究，或對藝術創作的心理用科學方法去研究，牠的能力所及只限於合理的部分，以外牠的能力就不夠了。若藝術家創作時根據科學的方法，從這裏產出的作品，決不是眞的藝術品。近來偶讀「人生觀之論戰」，見唐鉞氏對梁啓超氏的辯難，附帶着這一個問題。梁氏說：「關於感

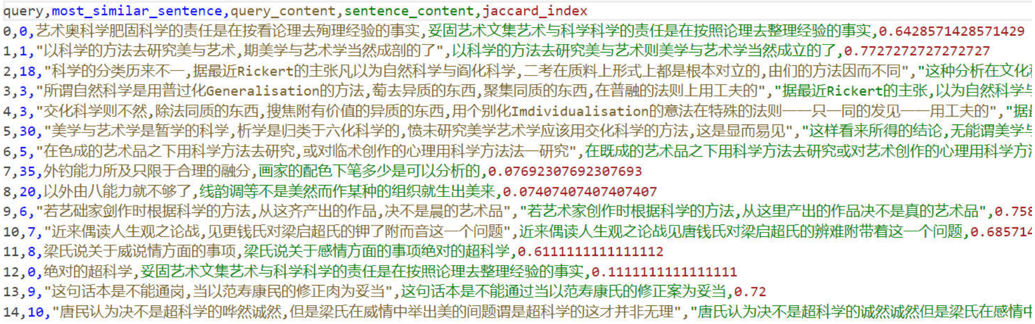
艺术与科学*

科学的责任是在按照论理去整理经验的事实。以科学的方法去研究美与艺术，则美学与艺术学当然成立的了。科学的分类历来不一。据最近Rickert的主张，以为自然科学与文化科学，二者在质料上形式上都是根本对立的，它们的方法因而不同：——所谓自然科学是用普遍化(Generalisation)的方法，剔去异质的东西，聚集同质的东西，在普遍的法则上用工夫的；文化科学则不然，除去同质的东西，搜集附有价值的异质的东西，用个别化(Individualisation)的方法在特殊的法则——只一回的发见——上用工夫的。美学与艺术学是哲学的科学。哲学是归类于文化科学的，那末研究美学艺术学应该用文化科学的方法。——这是显而易见的。

在既成的艺术品之下用科学方法去研究；或对艺术创作的心理用科学方法去研究，它的能力所及只限于合理的部分，以外它的能力就不够了。若艺术家创作时根据科学的方法，从这里产出的作品决不是真的艺术品。近来偶读《人生观之论战》，见唐钺氏对梁启超氏的辩难，附带着这一个问题。梁氏说：“关于感情方面的事项，绝对的超科学。”这句话本是不能通过，当以范寿康氏的修正案为妥当。唐氏认为决不是超科学的，诚然诚然。但是梁氏在感情中举出美的问题，谓是超科学的，这却并非无

Figure 6: Different versions of the same article, one in Traditional Chinese and one in Simplified Chinese.

I designed the following workflow (Github Link):
- Using i2OCR to get text files.
- Translate traditional Chinese characters into simple ones.
- Remove line breaks in text to get two whole strings.
- Remove spaces and use any punctuation mark as a separator to get an array of strings.
- Use Bert to embed the sentence into the feature vector.
- Use Faiss to do retrieval.
The result is a one-to-one correspondence of sentences between the two documents:

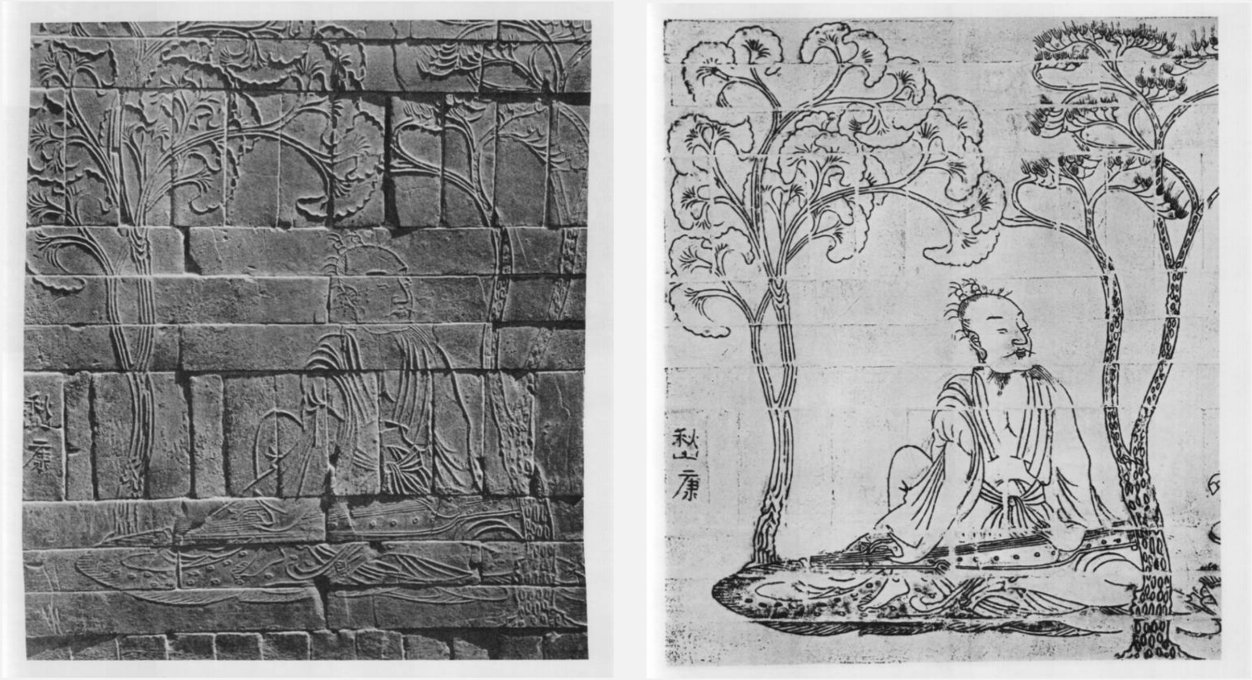


```
query,most_similar_sentence,query_content,sentence_content,jaccard_index
0,0,艺术奥科学肥固科学的责任是在按看论理去殉理经验的事实,妥固艺术文集艺术与科学科学的责任是在按照论理去整理经验的事实,0.6428571428571429
1,1,"以科学的方法去研究美与艺术,期美学与艺术学当然成剖的了",以科学的方法去研究美与艺术则美学与艺术学当然成立的了,0.7727272727272727
2,18,"科学的分类历来不一,据最近Rickert的主张凡以为自然科学与阎化科学,二考在质料上形式上都是根本对立的,由们的方法因而不同","这种分析在文化
3,3,"所谓自然科学是用普过化Generalisation的方法,蓟去异质的东西,聚集同质的东西,在普融的法则上用工夫的","据最近Rickert的主张,以为自然科学
4,3,"交化科学则不然,除法同质的东西,搜焦附有价值的异质的东西,用个别化Imdividualisation的意法在特殊的法则——只一同的发见——用工夫的","据
5,30,"美学与艺术学是暂学的科学,析学是归类于六化科学的,愤未研究美学艺术学应该用交化科学的方法,这是显而易见","这样看来所得的结论,无能谓美学
6,5,"在色成的艺术品之下用科学方法去研究,或对临术创作的心理用科学方法法一研究",在既成的艺术品之下用科学方法去研究或对艺术创作的心理用科学方
7,35,外钧能力所及只限于合理的融分,画家的配色下笔多少是可以分析的,0.07692307692307693
8,20,以外由八能力就不够了,线韵调等不是美然而作某种的组织就生出美来,0.07407407407407407
9,6,"若艺础家剑作时根据科学的方法,从这齐产出的作品,决不是晨的艺术品","若艺术家创作时根据科学的方法,从这里产出的作品决不是真的艺术品",0.75
10,7,"近来偶读人生观之论战,见更钱氏对梁启超氏的钾了附而音这一个问题",近来偶读人生观之论战见唐钱氏对梁启超氏的辨难附带着这一个问题,0.68571
11,8,梁氏说关于威说情方面的事项,梁氏说关于感情方面的事项绝对的超科学,0.6111111111111112
12,0,绝对的超科学,妥固艺术文集艺术与科学科学的责任是在按照论理去整理经验的事实,0.1111111111111111
13,9,"这句话本是不能通岗,当以范寿康民的修正肉为妥当",这句话本是不能通过当以范寿康氏的修正案为妥当,0.72
14,10,"唐民认为决不是超科学的哗然诚然,但是梁氏在威情中举出美的间题谓是超科学的这才并非无理","唐氏认为决不是超科学的诚然诚然但是梁氏在感情
```

Figure 7: One-to-one correspondence of sentences between the two documents.

However, the result is not very good, because the pdf itself is already very noisy, but the calibration requires that the OCR recognition result is almost completely correct. So the performance of OCR has become a bottleneck. On the other hand, I think the smallest unit of retrieval should be characters rather than sentences so that a more fine-grained comparison can be done.

In addition to this task, she also introduced me to another task about image segmentation, presumably copying from a stone tablet to get a painting, as shown in the figure:



Ji Kang/The picture of the brick mural

Ji Kang/The rubbing of the brick mural

Figure 8: Get a painting from a stone tablet

It can be said that the noise of the original image is just too big. The large number of shadows and cracks from the uneven stone blocks made the picture full of noise, and it was impossible to tell what was painted even with our eyes.

I will try to do these tasks that haven't been tackled yet sometime over the summer.

I had no idea or interest in these tasks before, but now I have the enthusiasm to think about how to solve them. I think that's the most important thing that this class has brought me.