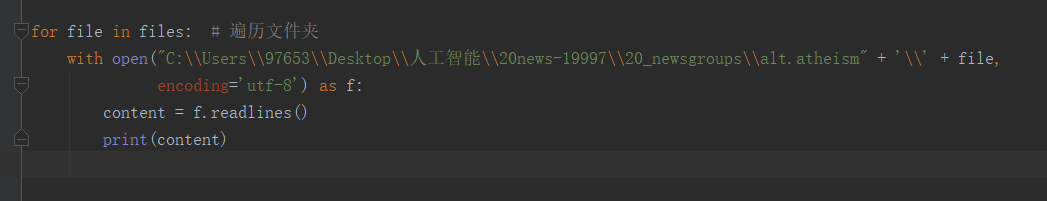
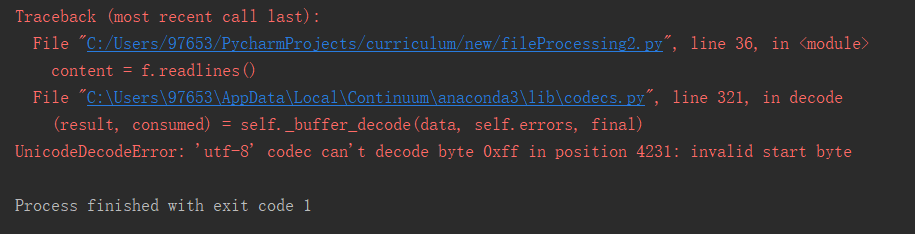
2018/9/22 \*The First Update

A few days ago I downloaded the news dataset from the given website. And the first I thought was to use the python ***os*** module to process the files and read them as a long list made up by the word strings. Afterwards, I could process them and build the word dictionary and derive the statistical information. However, it didn’t work as I expected. When I try to open the files and read them,

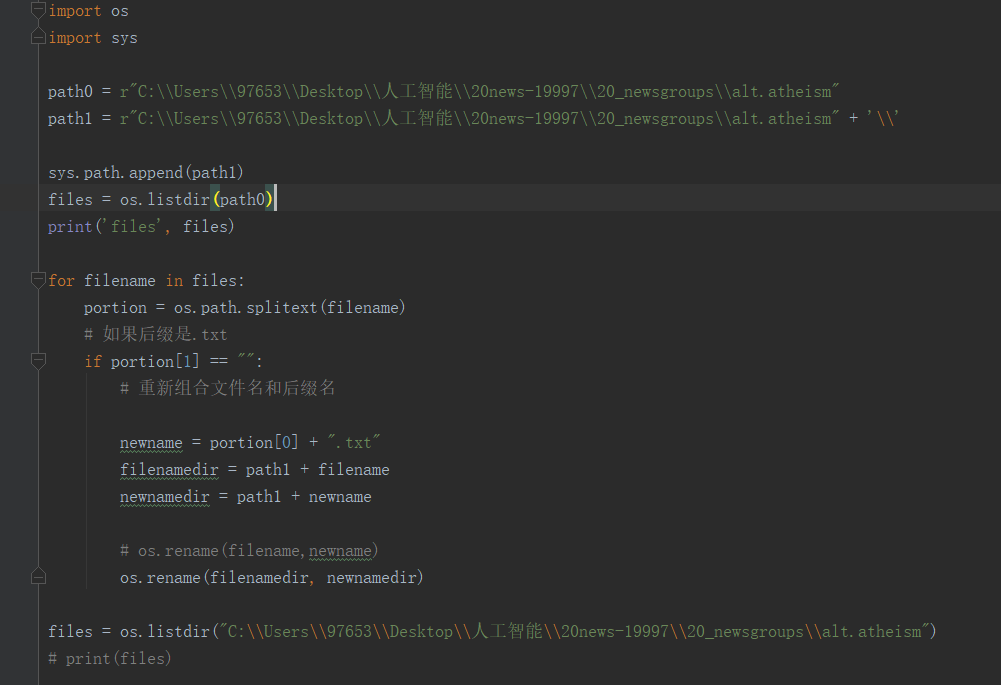


error occurred as the picture below.



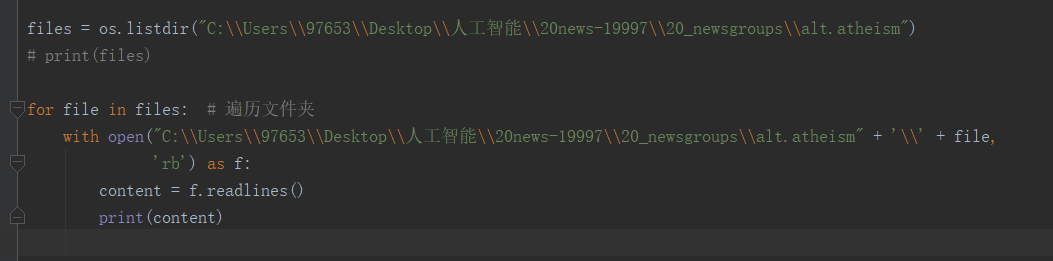
It was weird because I had never been in a situation like this, and I don’t know which part was wrong and basically I couldn’t find the ***position 4321***. I had no idea about why the dataset couldn’t be decoded with utf-8.

Later I found that the files given didn’t have a suffix name, thereby I did an array of operations to modify the suffix name to txt, resulted by my thinking that the error might come from the suffix name.



But unfortunately the error remained and nothing changed, so I did more digging on the error itself. I found an article on stackoverflow and it recommended me to change the method to read the files. I read them with ‘rb’. Without decoding them it became easier, now I am capable of lading them without obstructs.

Here’s the code:



2018/9/22 21：55

Fuck！

The code format is ISO-8859-1 and ascii…

I should’ve known that.

Oh, some formats are None.

I give up

I would ignore them.

Some more important errands are waiting for me.

2018/9/23 14：57

Shit

I got no idea about how to get rid of the useless information

Some rubbish are annoying

2018/9/26 \*The last update in the vsm stage

I gave up to clear the useless url information, because I thought it wouldn’t affect the result greatly. It’s because that every text possesses the similar information and in the bayes analysis it wouldn’t make a big difference.

Therefore I did every procedure and the model is built now.

But I have something unfinished.

I didn’t match the small-scaled dictionary with the overall dictionary, because I think there’s no need for this. It is a sparse matrix, I can fetch the data whenever I want I assume.