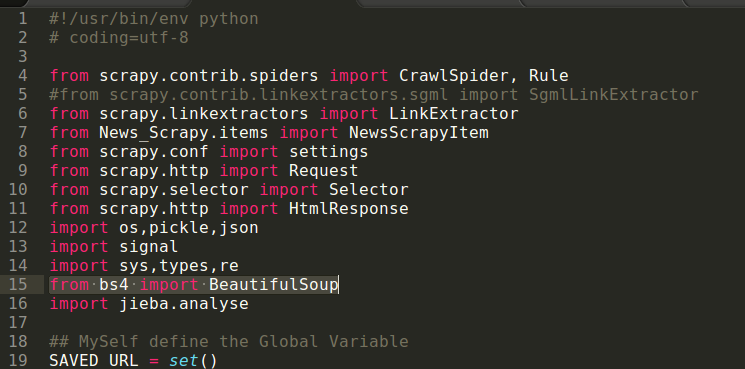
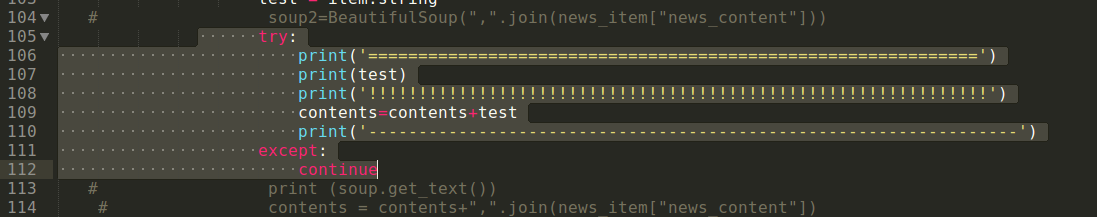
本爬虫设计是使用了scrapy爬虫框架爬取数据，再用sqlalchemy模块操作mysql来存放爬到的数据。目标网站为网易新闻。

1. 爬虫部分设计

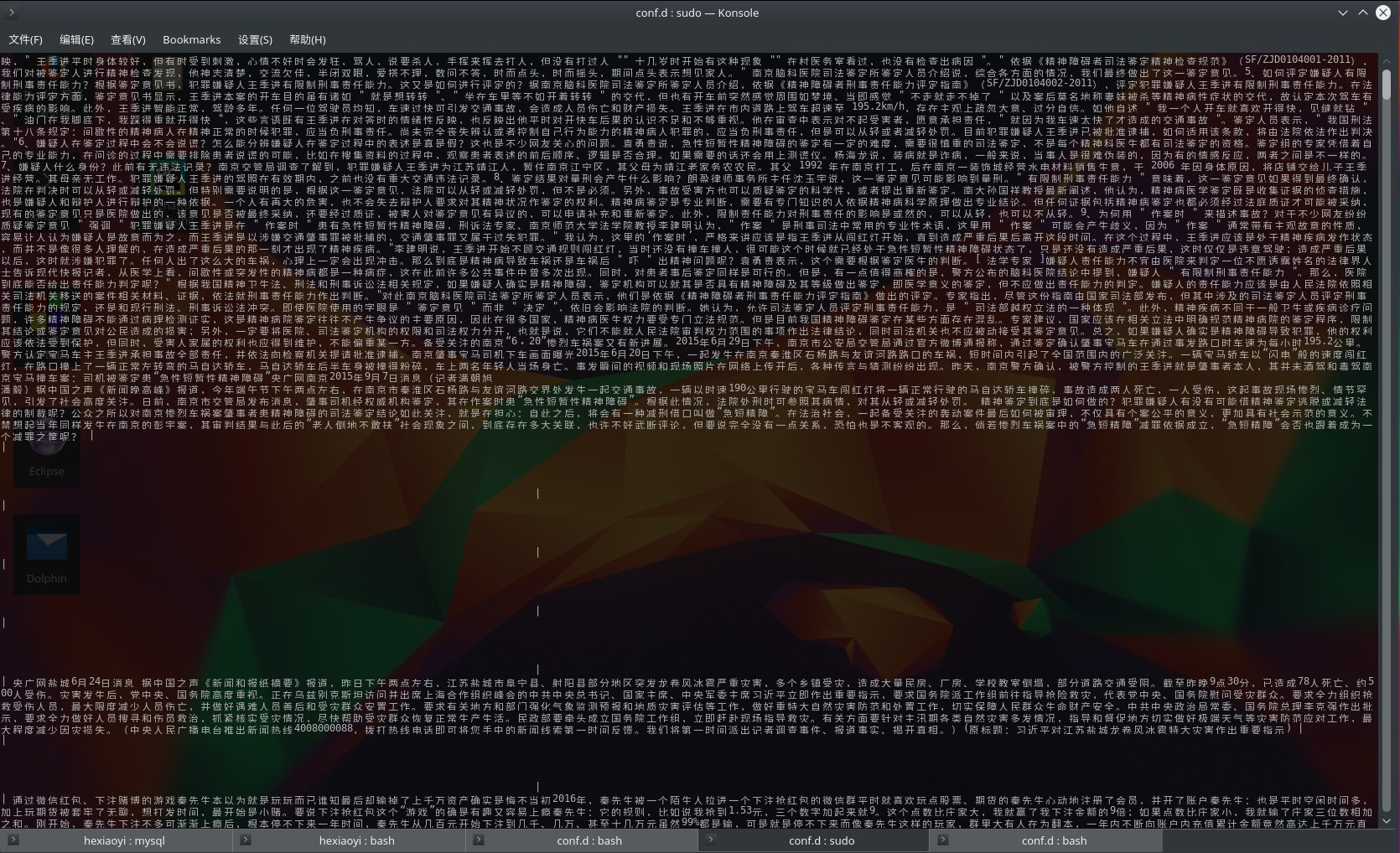
本爬虫采用了scrapy框架，并使用了xpath来判定网页元素。在抓取到网页文本后发现有很多标签残留，所以此处使用了beautifulsoup选择器来去除标签，只留有效的新闻文字。

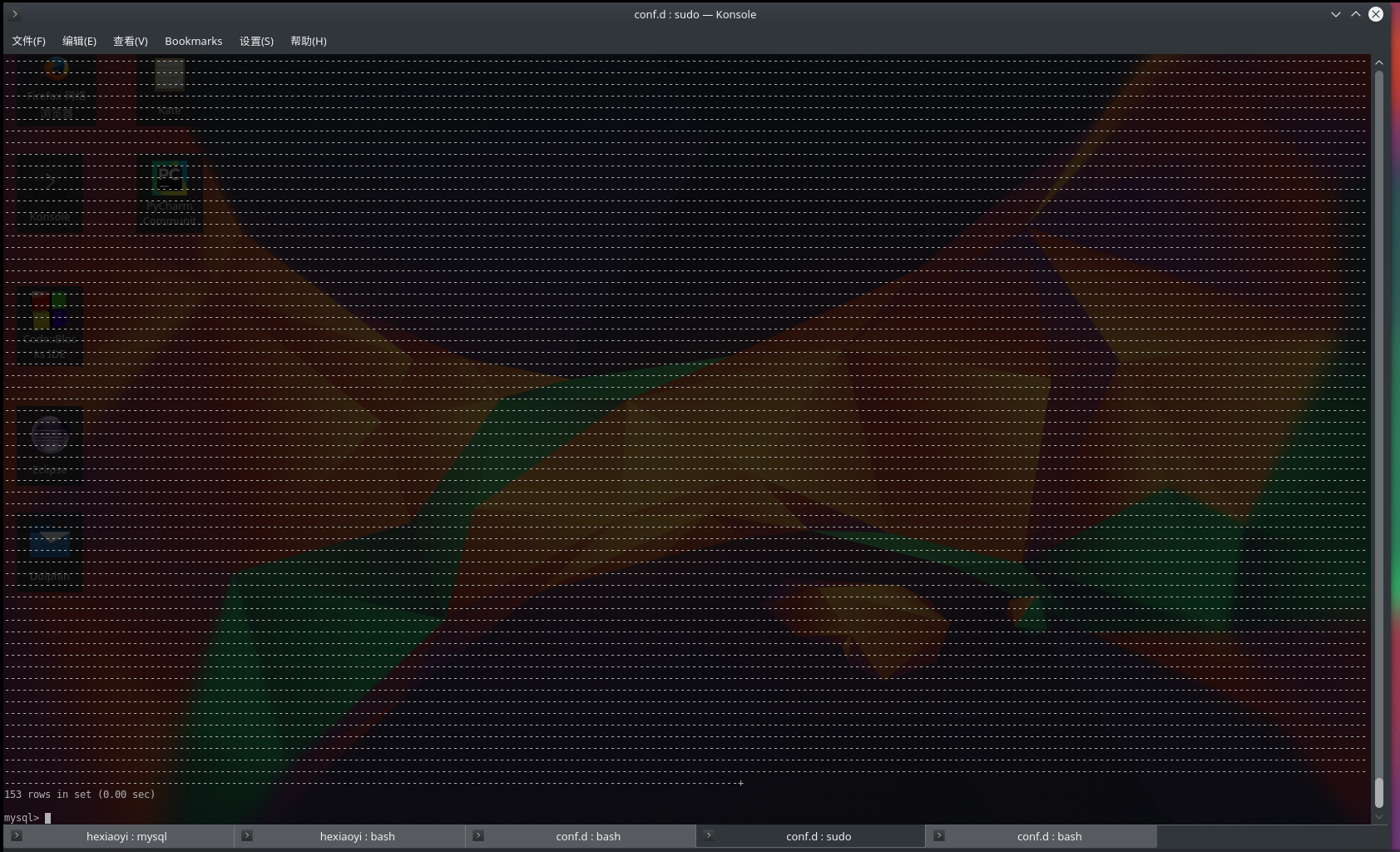


因为网易新闻里有两种不同版式的网页（普通新闻和组图新闻），组图新闻网页抓到的news\_content是None，如果不做处理会导致报错，然后整个网页的内容都不会被抓取到。所以，使用try: except来处理news\_content，如果没抓到news\_content，就给这篇新闻数据的news\_content字段设一个空字符串。



实际运行效果如图：





（因为控制台显示的限制，全部的news\_content没法完整的在控制台显示出来。但加where条件调取之后发现是可以正常调阅的，且153个新闻都已经存储进去了。）

