数据部分：

所需文件：

1. captioning_dataset_path: Path to GoodNews captioning dataset json file

有

2. fake_articles: Path to generated articles

有

3. image_representations_dir: Directory which contains the object representations of images

有已经resize好的图片 需要自己跑faster-rcnn

4. real_articles_dir: Directory which contains the preprocessed Torch text files for real articles

5. fake_articles_dir: Directory which contains the preprocessed Torch text files for generated articles

6. real_captions_dir: Directory which contains the preprocessed Torch text files for real captions

原始数据有 456都需要放入presumm进行处理

7. ner_dir: Directory which contains a dictionary of named entities for each article and caption

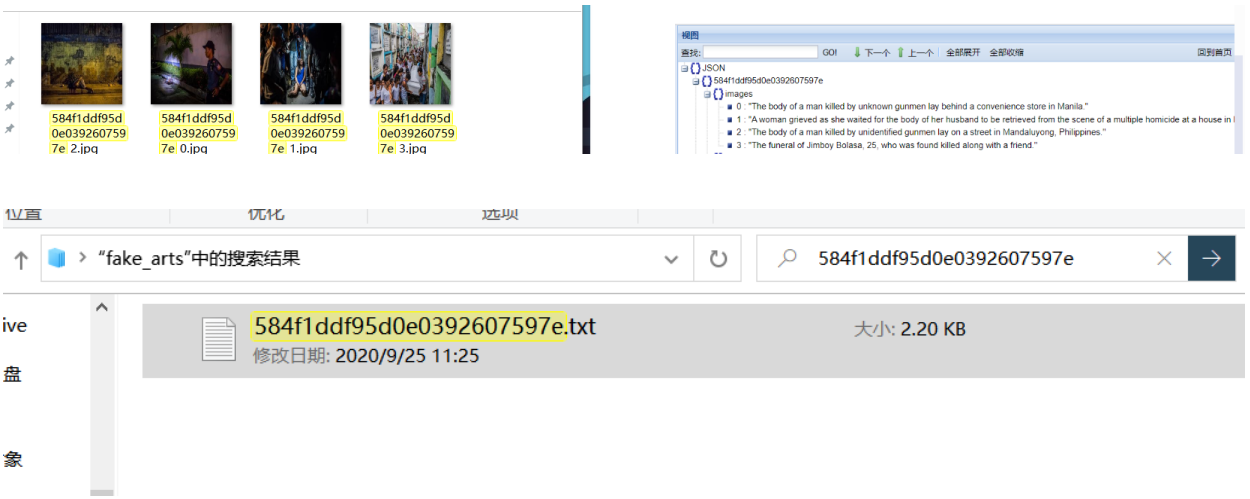
有

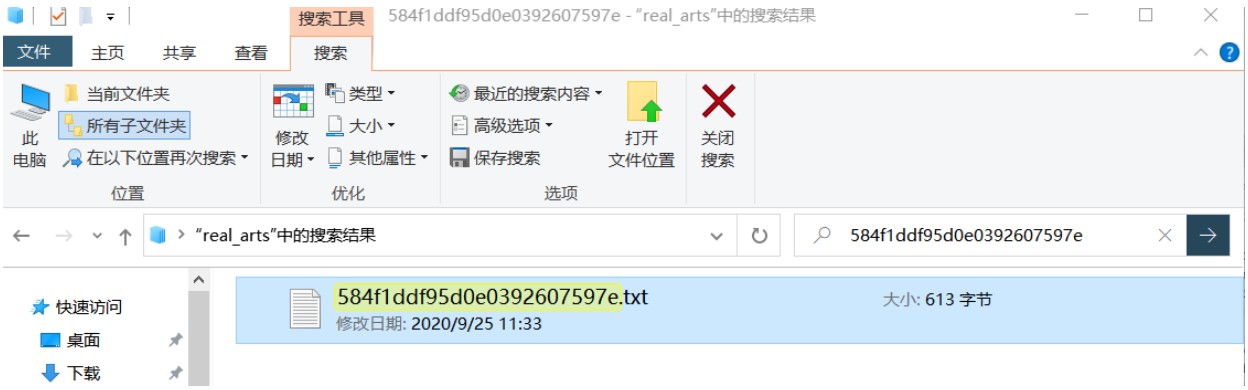
json文件样式：

584f1ddf95d0e0392607597e 为文章ID

584f1ddf95d0e0392607597e _? 为caption及其对应图片id

真假文章的id 相同





一篇文章对应多张图片 一张图片只有一个caption

真实的caption存于json文件中

fake_arts_fake_caps 与 real_arts_fake_caps 文件名相同的文件中的caption不一样