问题：
根据要求，训练和验证的数据分别是 21 万和 2 万句子，包括 SPO 作为训练标签，测试数据分别是 1 万，2 万两数据集。
SPO 标签类型是 object_type, predicate, object, subject_type, subject。‘postag’ 不是特别理解，应该是对词语进行分析之类。

具体内容：
1. 关键词提取
调研的文献：

学位论文
（1） 梁伟明，中文关键词提取技术
（2） 张 丽，文本挖掘中关键词与文本摘要自动提取研究
（3） 许梦馨，基于复杂网络的文本关键词提取分析平台

具体方法包括 PyNLPIR，TD-IDF，TextRank（具体含义可以百度）

（1） Pynlpir（中文提取）
代码链接
https://github.com/tsroten/pynlpir （python 库文件安装包）
https://github.com/NLPIR-team/NLPIR （java）
运行结果

```
import pynlpir
pynlpir.open()

s = '欢迎科研人员、技术工程师、企事业单位与个人参与NLPIR平台的建设工作。'
pynlpir.segment(s)

[('欢迎', 'verb'), ('科研', 'noun'), ('人员', 'noun'), ('、', 'punctuation mark'), ('技术', 'noun'), ('工程师', 'noun'
```

（2） TD-IDF（中文提取）
https://github.com/Jasonnor/tf-idf-python （python）
https://github.com/gaussic/tf-idf-keyword （python）
（3） TextRank（中文提取）
代码链接
https://github.com/letiantian/TextRank4ZH （python）
https://github.com/hankcs/TextRank （Java）
运行结果

```
关键词：
媒体 0.02155864734852778
高圆圆 0.020220281898126486
微 0.01671909730824073
宾客 0.014328439104001788
赵又廷 0.014035488254875914
答谢 0.013759845912857732
谢娜 0.013361244496632448
现身 0.012724133346018603
记者 0.01227742092899235
新人 0.01183128428494362
北京 0.011686712993089671
```

此外，还有

https://github.com/sdunlp/nlp_Chinese


```
"news_test_0.txt": {
    "code": 0,
    "sentiment": -0.5913626456902517,
    "tilte": "河南一精神病院患者用筷子袭击女患者，致三死一重伤",
    "abstract": "事件造成3名女性精神病患者死亡 杨某某被转移到医院过程中 杨某某家属与大众医院联系 并与医院工作人员一同将杨
    "time": "2017-04-01 19:57",
    "keywords": [
        {
            "frequency": 0.02186421173762946,
            "word": "患者"
        },
        {
            "frequency": 0.011507479861910242,
            "word": "某某"
        },
        {
            "frequency": 0.01380897583429229,
            "word": "精神病"
        },
        {
            "frequency": 0.01380897583429229,
            "word": "医院"
        },
        {
            "frequency": 0.014959723820483314,
            "word": "洛宁县"
        }
    ],
    "message": "sucess"
}
```