documents

aochujie

2022 年 06 月 02 日

Contents:

CHAPTER 1

琐事记录

1.1 2022 年 03 月

1.1.1 2022-03-02

transformers bert

今天复现 transformers 的 bert ner 模型。踩了两个坑：

1. 多个短句子拼接起来做 NER，不能乱选拼接符，比如之前用分号，导致准确率一直是 0，其实 bert tokenizer 有自带的一系列 `[unusedX]`，预留给下游任务使用。
2. 一条完整的语料，前面必须加上 `[CLS]`，因为预训练的时候一般都是这么训练出来的，使用的时候也要用这种模式，AB 句之间使用 `[SEP]` 拼接，如果没有 B 句，那也要在最后加一个 `[SEP]`，否则它可能感知不到句子的结束位置。

transformers dataset offline 模式

由于模型需要放到内网来跑。transformers 案例都是从互联网上临时加载数据集，有两种方法可以把数据集下载下来离线使用。

拷贝 cache 目录

外网运行一次数据加载，数据会自动加载到 ~/.cache 目录, 把它拷贝到内网就好了。这个 cache 目录也可以通过 XDG_CACHE_HOME 环境变量配置的。拷贝到内网后，还需要设置 HF_DATASETS_OFFLINE=true 启用离线模式。

手动保存数据

外网下载并导出数据到某个目录

```
import datasets

data = datasets.load_dataset(...)
data.save_to_disk(/YOUR/DATASET/DIR)
```

内网加载数据目录

```
import datasets

data = datasets.load_from_disk(/SAVED/DATA/DIR)
```

1.1.2 2022-03-08

创建一个新的 gradle 项目

安装 gradle

官网下载太慢，可以到腾讯云上去下载。一般来说下载最新版 *-bin* 后缀的二进制版就好了，如果是些 kotlin 的话，可能需要源码编译 gradle 插件，最好下载 -all 的带源码版本，以免某些情况下编译不通过。

执行 gradle init 初始化项目，

配置 gradle

在 settings.gradle.kts 最前面添加仓库配置。

```
pluginManagement {
    repositories {
        maven("https://maven.aliyun.com/repository/gradle-plugin")
    }
}
```

如果 maven 源是 http 而非 https, 对于高版本 gradle(7.x), 需要添加 allowInsecureProtocol 参数:

```
pluginManagement {
    repositories {
        maven {
            url=uri("https://example.com/repository/gradle-plugin")
            isAllowInsecureProtocol=true
        }
    }
}
```

1.1.3 坑

gradle 与新版 spring 不兼容。插件管理的坑，要求 https。mongodb 要求额外引入包: mongodb-driver-sync 自动 boot>=2.3 依赖，spring-data-mongodb 就更新了 3.0 版本，要求 mongodb>=3.6, 而公司一直用的是 3.2 发现虽然报错，但是可以访问，数据也正常生成了。最后发现是 Bean 创建不正确。https://docs.spring.io/spring-boot/docs/2.2.2.RELEASE/reference/html/appendix-dependency-versions.html#dependency-versions

http://gitlab.myhexin.com/10jqka/iwencai/a3/antaeus/antaeus http://gitlab.myhexin.com/10jqka/iwencai/a3/antaeus/antaeus-front

http://gitlab.myhexin.com/10jqka/iwencai/a3/auto-deep/ocr_doc_parser http://gitlab.myhexin.com/10jqka/iwencai/a3/auto-deep/ocr-client http://gitlab.myhexin.com/10jqka/iwencai/a3/auto-deep/auto-deep-back http://gitlab.myhexin.com/10jqka/iwencai/a3/auto-deep/auto_deep_models http://gitlab.myhexin.com/10jqka/iwencai/a3/auto-deep/reading_annotator_back

http://gitlab.myhexin.com/10jqka/iwencai/a3/pdf/kg-pdf-ocr-analyzer

http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/echo_pdf_machine http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/echo_executor http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/echo_integrator http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/echo_datasource http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/kg-dictionary http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/kg_lambda http://gitlab.myhexin.com/10jqka/iwencai/a3/echo/event_rule_front

CHAPTER 2

Indices and tables

- genindex
- modindex
- search