代码结构说明

code文件夹中包含src文件。

代码中的各种路径（原始数据路径、临时文件路径等）是在代码文件中配置的，但不建议修改。代码运行的路径要求在复现过程中说明。

Src

用于import的文件：

- `taacalgo/util.py` : 用于路径配置
- `taacalgo/dao.py` : 用于run中数据导入
- `taacalgo/w2v.py` : 用于配置word2vec以及训练过程
- `taacalgo/model.py` : 用于配置模型以及训练过程

用于运行的文件：

- `run_w2v.py` : 用于训练word2vec，生成序列与词向量矩阵。
- `run_model.py` : 用于训练模型。

运行环境

- linux平台，主流发行版应该都行
- 第三方包，conda有的话就用conda装，没有的话用pip装
    - python=3.6.5
    - numpy=1.18.5
    - pandas=1.0.5
    - scikit_learn=0.23.1
    - pytorch=1.4.0
    - gensim=3.8.3
    - tqdm=4.46.1
    - cos-python-sdk-v5

复现过程

1. 创建一个和code文件夹同一级的文件夹rawdata，里面放原始数据，文件夹的结构如下：
    - rawdata
        - train
            - preliminary
                - *.csv
            - semi_final
                - *.csv

    (解压train.zip即可)
2. 复现直接运行 `run.sh` 文件即可，此文件会在和code文件夹同一级的文件夹tmp_data/output中，存储每次模型跑出来的结果，以及log文件夹中，打印详细的日志。
3. 整合结果产生最后的表以及图参见agg.ipynb

H2 联系方式

宋宁宇邮箱 nngyusong@gmail.com