# 软件工程训练营大作业选题要求

软件工程训练营大作业可以从课程项目/企业项目/探索项目中任选一个，要求大作业项目具有一定的工程背景和实用价值，技术上存在一定实现难度。由小组成员合作共同完成该项目。作业必须通过GitHub提交，每一个队提交一个GitHub repo地址，要有commit的信息，这个也是评分项之一。项目完成后撰写结题报告并进行演示答辩，这些也要commit到GitHub repo中，报告和答辩ppt中需要明确写出队员的分工情况。根据每队GitHub上代码的情况，每个成员的贡献程度，以及项目结题报告和答辩的完成情况进行评分。

请使用提供封面完成，并参考提供的目录和内容结构。

其中，课程推荐项目如下：

## 基于YOLO和DeepSort的人体识别与跟踪

**描述：**小组成员自行采集视频或图片数据，结合课上所学内容，基于YOLO-NAS、YOLO-v5、v7等和DeepSort实现人体的识别与跟踪，并完成下面的功能：

1. 基于OpenCV读取摄像头显示画面，利用YOLO实现摄像头人体识别；
2. 在GUI界面中实时显示识别的结果，点击某一个人，可以实现对这个人的跟踪。即点击这个人的bounding box，他的颜色和剩下的将不一样，在界面上可以显示出这个被跟踪的人的bounding box信息。
3. 有一个按钮可以取消跟踪，当被跟踪的人离开画面时，自动取消跟踪；

**选做功能：**

1. 能够根据目标对象的移动调整摄像头/视频画面位置/放大对象所在区域，使得目标对象始终在画面中央。
2. 多目标跟踪，可以在界面上选择多个人就行跟踪。

**参考资料：**

<https://github.com/sujanshresstha/YOLO-NAS_DeepSORT>

<https://medium.com/augmented-startups/deepsort-deep-learning-applied-to-object-tracking-924f59f99104>

https://github.com/deshwalmahesh/yolov7-deepsort-tracking

## 图像分割segmentation软件开发

**描述：**制作一个软件，可以自动的完成图像分割的训练和推理任务，可以使用UNET，或者一些预训练的UNET进行微调，实现下面的功能：

1. 有一个完整可用的GUI界面，可以选择训练的文件夹，先择这个文件夹以后可以配置一些基本的训练的超参数；
2. 点击训练以后自动对文件夹里面的图片和mask进行处理，mask可以只有一个类型，并训练生成模型，保存到模型文件夹，界面上显示模型的性能指标；
3. 选择某一个已经训练好的模型，然后点击推理，选择一个图片文件夹，自动对文件夹中所有图片进行处理，输出对应的mask。

**选做：**

1. 支持多个类型的mask
2. 支持从预训练的模型迁移

**参考资料：**

<https://github.com/milesial/Pytorch-UNet>

<https://github.com/nikhilroxtomar/UNet-Segmentation-in-Keras-TensorFlow>

<https://github.com/zhixuhao/unet>

<https://github.com/kenhding/Coursera/blob/master/Deep_Learning_Specialization/Course_4_Convolutional_Neural_Networks/Week3/DL_C4_week3_2_Image_Segmentation_with_U-Net_v2.ipynb>

<https://github.com/sushantdhumak/Image-Segmentation-with-U-Net>

<https://github.com/NirmalElamon/UNet-model-for-Segmentation>

下面是可以用的数据：

<https://www.kaggle.com/datasets/usharengaraju/unetr-model>

<https://www.kaggle.com/code/hsankesara/unet-image-segmentation>

<https://www.kaggle.com/datasets/quadeer15sh/augmented-forest-segmentation>

<https://www.kaggle.com/datasets/franciscoescobar/satellite-images-of-water-bodies>

<https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation>

<https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database>

<https://www.kaggle.com/datasets/faizalkarim/flood-area-segmentation>

<https://www.kaggle.com/datasets/mahmoudreda55/satellite-image-classification>

## 一下是企业提供的可选项目

## 小组成员根据课程提供的数据（A股股票数据），实现下列任意一个kaggle的算法，并用 qlib 或者 backtrader 对相关方法进行回测。

1. 九坤：项目目标--预测A股股票的未来一天的收益；链接：https://www.kaggle.com/competitions/ubiquant-market-prediction/discussion/338220
2. G-research：项目目标--预测加密货币的短期回报；链接：https://www.kaggle.com/competitions/g-research-crypto-forecasting/discussion/323098
3. optiver：项目目标--预测股票的波动率；链接：https://www.kaggle.com/competitions/optiver-realized-volatility-prediction/discussion/274970
4. jane-street：项目目标--制定股票交易策略；链接：https://www.kaggle.com/competitions/jane-street-market-prediction/discussion/224348
5. Two Sigma：项目目标--预测股票交易价格；链接：https://www.kaggle.com/competitions/two-sigma-financial-modeling/discussion/29793

## 信息检索与推荐系统，小组成员根据课程提供训练数据集实现模型，验证集实现n-fold CV，最终记录在测试集表现（Metric至少需要包含AUC）提交实验流程报告。下列题目二选一：

* 1. 基于蚂蚁金融的语义匹配模型，数据集：<https://tianchi.aliyun.com/dataset/dataDetail?dataId=106411>
  2. 基于淘宝用户行为序列，提交一个用户兴趣模型，对用户针对商品感兴趣程度打分（购买、加购、喜欢都可以视为感兴趣标签，有pv无其他动作视为不感兴趣），11月25日-30日数据训练，12月1日-2日数据验证，12月3日数据测试。时序模型注意不要引入look ahead bias。数据集：https://tianchi.aliyun.com/dataset/dataDetail?dataId=649

## 基于统计方法的中文词汇挖掘实践

（1） 学习了解文本数据挖掘的方法

学习路径参考

https://www.kaggle.com/code/kanncaa1/applying-text-mining

（深度学习-选做）https://www.kaggle.com/code/eliotbarr/text-mining-with-sklearn-keras-mlp-lstm-cnn

（2） 学习经典中文词汇挖掘（新词发现）方法，并使用python实现

（方法及思想）http://www.matrix67.com/blog/archives/5044

（java实现参考）https://github.com/sing1ee/dict\_build

（3）（选做）NLP领域模型发展总结，完成一个深度学习NER任务模型

实现参考（可忽略tpu部分，有条件的同学可学习使用配置nvidia-GPU）

https://www.kaggle.com/code/tanulsingh077/deep-learning-for-nlp-zero-to-transformers-bert