第十届中国大学生服务外包创新创业大赛企业命题类赛题相关问题解答 F&Q（10.0 补充）

（截止至 3 月 6 日）

一、 【A01】2018 网络零售平台商品分类【浪潮】

1.关于数据编码问题，能否将 test.tsv 和 train.tsv 的文件编码格式改为 utf-8，换行符改为\n？

答：可以。

2. 基于训练样本的各类别不均衡,是否可以做数据增强,进行数据扩充？

答：可以通过自己采集网上商品数据进行补充，但是自己要确保分类准确。

3. 训练集有严重的不均衡性，数据中小于 100 个 case 的类有 771 个，占了绝大多数比例。其他数据的分布与训练集一样吗？

答：这个属于自己需要优化的地方，数据都是随机从企业的采集商品库中取出来。

4. 请问可视化里的批量输入是指，导入文件。还是网页中一次输入多条记录？

答：批量就是，将所有的数据分类可以展示出来。

5. 请问如果给 50W 训练集打标签，是在原文件 type 列旁打还是去掉原来的文件有的标签重新打？

答：新添加标签，不要去掉原来的列，提交的时候注意说明一下。