# CDA LEVEL III

(考试时间120分钟，满分30分)

### 四、案例操作题

根据相同的背景材料和数据实作以下的分类模型，最终须提交对测试数据的预测结果。

1. 题目：从产品的消费者评论中识别此评论的情绪是正评或负评。
2. 说明：韩国专业市调公司Embrain旗下子公司Trendmonitor针对在2017年有去过百货公司或Outlet、19~49岁韩国男女共1,200人，进行「消费者评论」相关调查。结果发现，在消费生活中「消费者评论」发挥着非常大的影响力。 86.9%的消费者都同意「消费者评论」是有其必要的，尤其以女性和20岁代消费者，特别认同「消费者评论」的需求性。基本上，消费者对「消费者评论」的信任度相当高，七成 (69.3%) 表示对一般消费者实际购买/使用商品之后，所分享的使用经验心得或商品评价等，保持信任的态度。相反的，不相信「消费者评论」的比例只有4.3%。由此可知，提早了解消费者评论是属于正评或负评(尤其是负评)，是相当重要的一件事。如何从产品的消费者评论中发现重要的情绪特征，文本挖掘是一个最好用的工具。有鉴于此‚本考题将提供许多产品的消费者评论的相关数据，包含CID(评论序号)、Clothing ID(被评论的产品序号) 、Age(评论人年龄) 、Title(评论标题) 、Review Text(评论内容) 、Positive Feedback Count(认为该评论有用的人数(点赞数)) 、Division Name(产品所属大类) 、Department Name(产品所属中类) 、Class Name(产品所属小类) 、以及目标字段Recommended IND(正负评(1是正评, 0是负评))。考生可利用文本挖掘中的分类技术，建立一个能预测产品消费者评论的情绪是正评或负评的一个的分类预测模型。
3. 进行方式：考生将取得训练数据及测试数据。格式为.csv(以逗号分隔字段名称及数据)的文件格式。训练数据包含18,789笔产品消费者评论的相关数据；每笔数据包含10个字段(1个 评论序号CID、8个产品消费者评论的相关信息及1个目标字段Recommended IND正负评：1是正评, 0是负评。测试数据包含4,697笔产品消费者评论的相关数据；字段个数与训练数据相同，但没有目标字段。特别要注意的是考生可以运用机器学习及深度学习的算法来建立模型，但使用深度学习时，需考虑机器本身的效能及时间的限制。考试是以结果论，没有输出任何的结果供评估时，不会有此部分的分数。同时，若能善用特征工程的技术来建立模型，预测结果通常会比较理想。考生以训练数据为基础，建立一个预测产品消费者评论的情绪是正评或负评的分类预测模型，并输出一个测试结果的档案(考生姓名\_results.csv)。考生姓名\_results.csv中有2个字段，分别是评论序号CID以及预测贴文情绪的结果。results.csv的形式如下：

CID,Predicted\_Results

3,0

4,1

9,0

...

最后需将 考生姓名\_results.csv 连同代码及建模的相关档案 (未传代码及相关档案者以0分计分)，一并压缩成.zip拷贝给监考老师。考生姓名\_results.csv的格式请务必正确(例如，预测结果必须为0或1，不可为y或n)，否则也将酌情扣分。

1. 评分方式：评分方式是以负评的F1来评估预测结果的好坏。参与认证者的成绩以F1的结果排序，F1越大者越好。。

此部分，共30分。

评估的结果在所有考生排名的前5%， 得30分

评估的结果在所有考生排名的前10%，得24分

评估的结果在所有考生排名的前20%，得18分

评估的结果在所有考生排名的前40%，得12分

评估的结果在所有考生排名的前80%，得6分

评估的结果在所有考生排名的后20%，或全猜正评(1)者得0分