## 1、环境配置

Gensim 是一个专注于无监督文本语义建模的Python 库，尤其擅长处理大规模文本数据。它的核心功能是将文本转化为向量表示，从而支持语义分析、主题建模、相似性检索等任务。内置Word2Vec、Doc2Vec、FastText、LDA、LSI、等经典算法。API设计直观，例如训练Word2Vec只需几行代码。在本课程专用环境中安装pandas和gensim，与numpy=1.22.3适配的pandas版本是pandas==1.4.3。

## 2、代码部署

gensim Word2Vec的示例代码与数据：<https://github.com/nlpresearchers/nlpbook/tree/master/chapter06/6.2.4> 。其中train.csv是餐厅评价数据，6.2.4.py中包含了后续任务所需全部 API 调用。

如果网络不好无法下载train.csv，可在学习通资料-数据下载。

在本地新建word2vec\_gensim.ipynb文件，将代码仓库中的内容移植至word2vec\_gensim.ipynb文件，注意合理规划单元格，确保每个单元格下方只包含一个结果。

## 3、模型训练

充分理解6.2.4.py中genism Word2Vec的API调用方式，实现以下任务，每一个任务单独一个单元格，上传代码时需展示单元格执行结果（输出置于代码下方）。

1. 使用Skip-Gram训练Word2Vec模型（注意代码示例里使用的是CBOW）。
2. 输出“环境”的词向量及其形状。
3. 输出与“好吃”语义最接近的3个词。
4. 计算“好吃”和“美味”的相似度、“好吃”和“蟑螂”的相似度。
5. 执行向量运算“餐厅+聚会-安静=？”，输出1个最相关结果（类比“王子-男+女=公主”）。

## 5、成果提交

word2vec\_gensim.ipynb单元格保留输出结果，等待与下节课内容一起提交。