# 歌词生成大数据系统案例教学指导手册

**教学适用的课程：**《大数据架构与技术》

**教学适用的专业：**电子信息

**教学目的与用途：**理解实际生活中的大数据应用；掌握数据挖掘、数据预处理相关知识点；掌握大数据分析的相关知识点；掌握深度学习的相关知识点；掌握数据可视化的相关知识点。

**教学内容:**

（1）理论依据：本案例是基于深度学习的歌词生成大数据系统。案例整合了常用的数据挖掘与深度学习方法，进行数据预处理，建立循环神经网络模型以学习歌词风格和生成歌词，并根据用户需求进行歌词数据可视化。

（2）涉及知识点：数据挖掘；数据预处理；循环神经网络；数据可视化。

（3）分析路径：首先分析了歌词对于歌曲的重要性和存在的问题。随后分析了网易云音乐的爬虫设计思路，使用Python数据挖掘，获取了网易云音乐中的歌词数据。再对挖掘到的数据进行预处理，删除了其中包含的无关信息。然后建立了循环神经网络模型来学习歌词风格，并使用训练好的模型来生成用户指定主题和风格的歌词。最后应用可视化相关技术向用户展示模型生成的歌词。

**启发思考题:**

如果循环神经网络模型生成的歌词质量不理想，请问有哪些改进思路呢？（参考答案：可能是输入的歌词数据不够多，模型没有学习到有用的参数；可以试试其他神经网络模型来处理该问题，例如LSTM模型等等。）

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的数据挖掘相关知识、数据预处理相关方法、深度神经网络的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：我想在座的每个同学都或多或少的爱听音乐。对于我们听众来说，歌曲中的歌词是评价歌曲好坏极其重要的一个方面。我们都知道，随着人工智能的不断发展，计算机已经可以做一些以前看来只能人来做的事情。那么，计算机能否自动为我们生成歌词呢？这节课，我们一起来看看吧！
5. 结束总结：这节课，我们从无到有，通过训练一个循环神经网络，建立了一个可以自动为我们生成歌词的系统。在没有学习这节课之前，同学们是不是觉得这是计算机根本不能做到的事呢？其实，计算机能做的事情远不止自动生成歌词这么简单。更多的应用还等待着同学们在学习完成大数据课程后自己发现。只要大家努力学习，我相信同学们一定能做到更多现在看来不可思议的工作！
6. 案例引导建议：在教学过程中，教师可以引导学生发散思维，将本案例的模型用于其他任务。

**参考文献:**

[1] SHVACHKO K, KUANG H, RADIA S, et al. The Hadoop Distributed File System; proceedings of the IEEE Symposium on Mass Storage Systems & Technologies, F, 2010 [C].

[2] DEAN J, GHEMAWAT S. MapReduce: Simplified Data Processing on Large Clusters [J]. USENIX Association, 2004.

[3] ZAHARIA M, CHOWDHURY M, DAS T, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing; proceedings of the Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, F, 2012 [C].

[4] TAYLOR R C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics [J]. Bmc Bioinformatics, 2010, 11 Suppl 12(S12): S1.

[5] 刘小虎, 李生. 决策树的优化算法 [J]. 软件学报, 1998, 9(10): 4.

[6] 格林贝格. Flask Web开发 [M]. Flask Web开发, 2015.

[7] 王子毅, 张春海. 基于ECharts的数据可视化分析组件设计实现 [J]. 微型机与应用, 2016, 35(014): 46-8.

**其他教学支持材料：**

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（深度学习算法库），Flask和ECharts（可视化绘图库）。