Week2 优秀作业

Thinking部分见文件：名企BI04\_Lesson02\_Thingking.txt Action部分见文件：名企BI04lesson02\_Action1\_购物篮可视化.ipynb GitHub仓库链接：https://github.com/Meiyun-Tan/kkbAssignment/tree/master/BI04%E5%90%8D%E4%BC%81%E7%8F%AD Action的总结： 1. 词云可视化中： WordCloud(max\_words).generate(text) ,会显示text字符串中，前max\_words个词； WordCloud()方法中参数collocations最好设为False，否则text中相同的词可能会被当做重复词而不计算，导致单词排名不准； WordCloud().generate\_from\_frequencies(freq) 可以根据 {商品:频数}的字典生成词云图； WordCloud().process\_text(text) 可以从一整串字符串中统计出每个空格隔开的单词出现的频数，并以{商品:频数}的字典形式返回； generate\_from\_frequencies()、process\_text()连用所得到的结果，与generate()的结果相同，不同的地方在于，我们可以在process\_text()的结果中得到它们的频数。 2. seaborn 绘制直方图默认会显示密度分布曲线，比Matplotlib更加直观 经过5中检查频数不一致的分析，发现： collections.Counter()也是个批量统计字符串中每个单词词频的好方法； 需要注意的是，wc.processtext()在统计词频时，它会将字符串中的&符号作为分隔符，如果不想被拆分，需要提前将该符号替换为其他词，如:'and'或 ''。 3. Series.str.get\_dummies()进行one-hot编码时，行内重复元素只记录一次。

https://github.com/MuyuCesare/AI/tree/master/BI/Homework%E9%A1%B9%E7%9B%AE%E5%AE%9E%E8%AE%AD%201-1