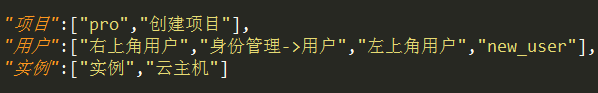
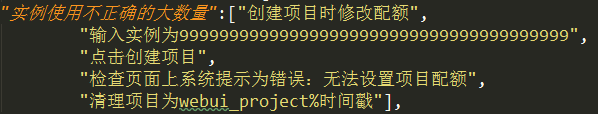
首先，根据case的特征制作特征辞典，如图所示：



其次，根据特征将case处理为一串特征序列。

例如：



将会根据上图特征辞典处理为：



将每一个case都处理为类似的特征序列后，则试图计算不同序列之间的相似程度。

相似程度定义为：



其中：





Pi为第i个特征关键词出现的频率，diffranki为两个序列中秩的差的归一化。特别地，在dice中默认序列中没有重复元素，若有则也算作不同元素。

例如：

a = ['项目', '实例', '项目', '用户', '项目']  
b = ['实例', '项目', '实例', '项目', '项目']

其中：



其秩和分别为：

{'实例': 2, '项目': 9, '用户': 4}

{'实例': 4, '项目': 11}

秩和对应的差为：

{'实例': 2, '用户': 4, '项目': 2}

归一化后为：

{'实例': 0.25, '用户': 0.5, '项目': 0.25}

比例为：

{'实例': 0.3, '项目': 0.6, '用户': 0.1}

最终：

