**数据获取**

我们从stack overflow中搜索关键词”ide”后得到许多网页，其中包括多种多样的问题和回答，但是鉴于本次实验的时间和机器内存等有限，且过于靠后的网页的问题和回答往往质量和数量不如排名靠前的网页高，所以采取爬取前50页的问答内容。

将每个问答作为一个需求，用kmeans算法对文本进行聚类并获得每个聚类中心，也即可代表属于同一类问答的一条问答。随后获取聚类中心问答中潜在的需求。

最终选取12条最具代表性的需求如下：

1. 自动补全：变量名、类型名在编辑时，有自动补全的可选项
2. 缩进处理：单行或多行的缩进和取消缩进
3. 模块化调试：可单独调试某一函数
4. 错误定位：报错行以及warnings行的定位功能，可在每条报错和warning间切换
5. 工具库：包含更完整的工具库，并能正确定位到所需的工具库
6. 完善编译器的语法解析功能，如减少无缘报错的情况发生
7. 统一缓存的实现，使得同样的代码在不同IDE中有相同的运行结果
8. 提供可视化工具，如可视化项目用到的文件的联系、类之间的关联等
9. 提供更简洁的代码风格配置功能
10. 可在跨系统方面更加包容和完善，如在Windows系统上使用bash
11. 可支持打开多种格式文件，也可支持将文件导出为多种格式
12. 快捷键：如在错误定位中引入快捷键，也可自行定义