实验一：需求获取 实验报告

一、实验目标：软件需求的抽取与分类

二、小组成员：

李祐德 161220073

郭睿杰 171860541

· 蔡恒 171860539

叶宙果 175220002

三、小组分工：

郭睿杰：数据爬虫和聚类分析

蔡恒&叶宙果: 对爬取的数据进行分析，获取需求

李祐德: 整理实验结果和数据，完成实验报告

各自完成的部分所用到的代码，数据等都保存在仓库中的个人文件夹中。

四、实验概要：

我们主要从两个方向获取需求：

1. 从github上vscode项目的issuse中爬取数据，并尝试对标签进行聚类分析

2. 爬取stackoverflow上的3000条提问所属标签，并从中分析需求

Vscode项目需求获取

实验步骤：

1. 确定IDE项目：

选择的是github上的vscode项目，url: <https://github.com/microsoft/vscode>

1. 明确信息源：

将这个项目的open issues和closed issues作为了主要的数据来源。labels作为分类的辅助标准也将起到很大的作用，所以labels也是数据来源。

closed issues：

<https://github.com/microsoft/vscode/issues?page=1&q=is%3Aissue+is%3Aclosed+sort%3Acomments-desc>

open issues：

<https://github.com/microsoft/vscode/issues?page=1&q=is%3Aissue+is%3Aopen+sort%3Acomments-desc>

labels：

<https://github.com/microsoft/vscode/labels?page=i&sort=name-asc>

1. 获取数据：

写了两个爬虫分别获取了issues和labels的数据。open issues和closed issues各获取了200页，都是按评论数降序获取的，因为评论数的多少一定程度可以反映这个issue的重要程度与受关注程度。labels获取了所有的labels。

1. 分析需求：

我们首先尝试了用聚类分析的方式来进行分析，但由于结果不理想，所以后续仍然采用了一些人工处理来分析需求。具体步骤和结果如下：

图片包含 报纸, 文字

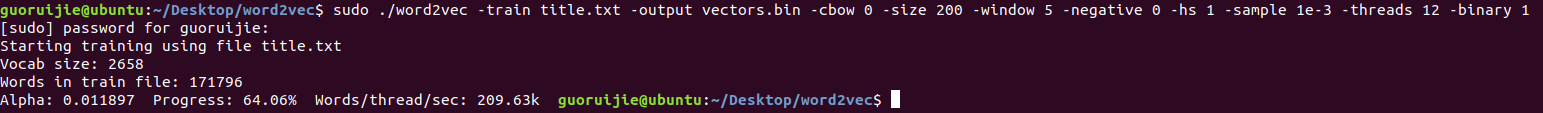
描述已自动生成

利用爬取的数据生成了以上的词云图。通过观察这个词云图，我们去除了一些对分类无用的关键词，并对这些关键词先进行观察性的分类。

我们最终确定对需求的分类准则有以下几条：

(1) 编辑器和语言 (2) 嵌入式 (3) 网络编程 (4) debug (5)工具 (6) 插件 (7) 版本控制 (8)智能（自动）数据库支持

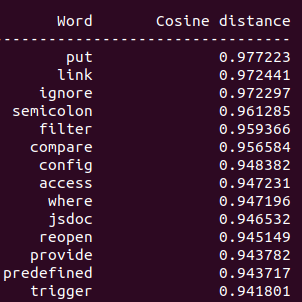
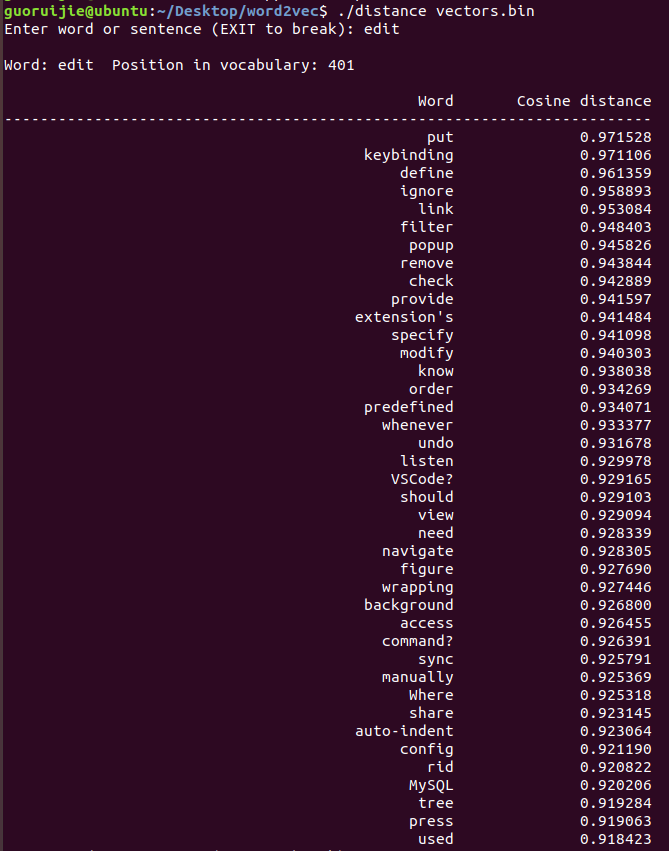
利用word2vec，对文本所有词汇向量化，并用单层神经网络进行训练：



训练完成，进行聚类分析。聚类分析的相关词汇输出结果中并非所有词汇都令人满意，因此

我对结果进行了筛选，筛选结果如下：

1. 编辑器和语言。首先是编辑器(editor, input, edit)



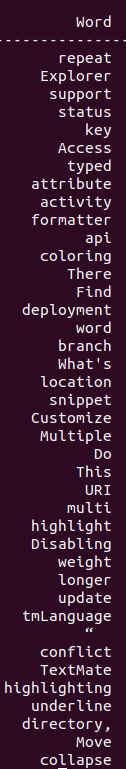
总结关于编辑器的功能如下：

* Auto-indent，要求自动缩进
* Background，要求能够设置背景颜色
* Semicolon，需要对缺少的分号进行提示
* Tree，需要能够展示当前项目目录树
* Navigate，需要有使用导航（帮助）
* Undo，支持撤回，重做（附带的还有一系列的文本编辑功能，复制粘贴等等）

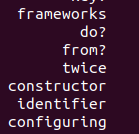
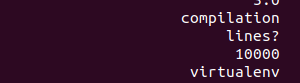
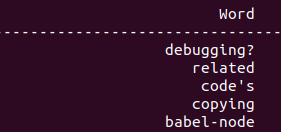
关于语言(language)：

从右图可以看出，相关性比较有价值的词有textmate highlight underline Find color等。因此还是属于编辑类范畴，要求编辑器能够支持：

高亮语法，编辑工具，下划线，文本查找，文字涂色



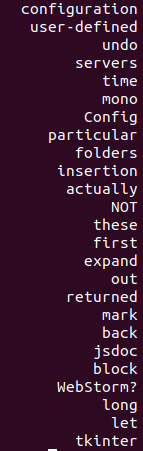
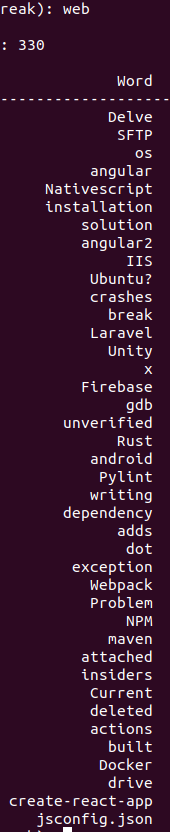
1. 嵌入式开发(embed)



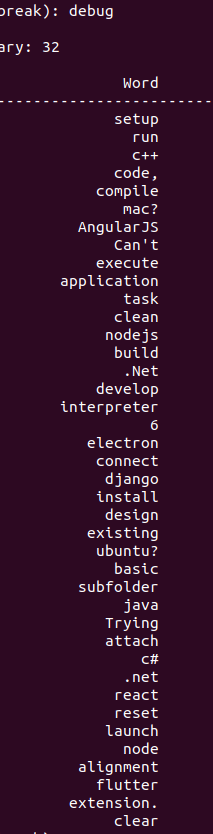
总结：

* 嵌入式开发需要支持：
* Debug工具(debugging)
* 嵌入式开发编译器(compilation)
* 虚拟开发环境(virtual environment)
* 开发框架(frameworks)
* 必要的配置工具(configuring)

1. 网络编程（web,net,.net,html）



总结：支持js框架，支持调试器(如gdb)。还有支持不同平台(Android,ios)

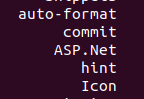


1. Debug

总结：

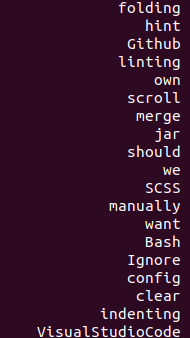
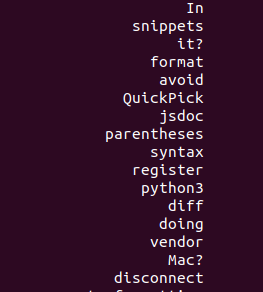
* 对多种语言的支持（c++，.net,c#,nodejs）
* 有良好的扩展插件支持(extension)
* 针对框架的debug能力(Django)
* 跨平台(ubuntu)

1. 插件(extension)（聚类分析结果不够理想）
2. 版本控制(version,git,github)



* 自动格式化

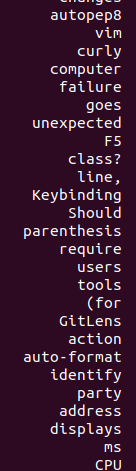
1. 智能(smart,automatic)



总结：

* auto-formatting（自动格式化），indenting（自动补全）.
* 其中Snippet是一个能够提高编码效率的插件，拥有极端强大的自动补全功能，可以考虑在插件列表中添加其获得相应的支持

1. 必要的工具

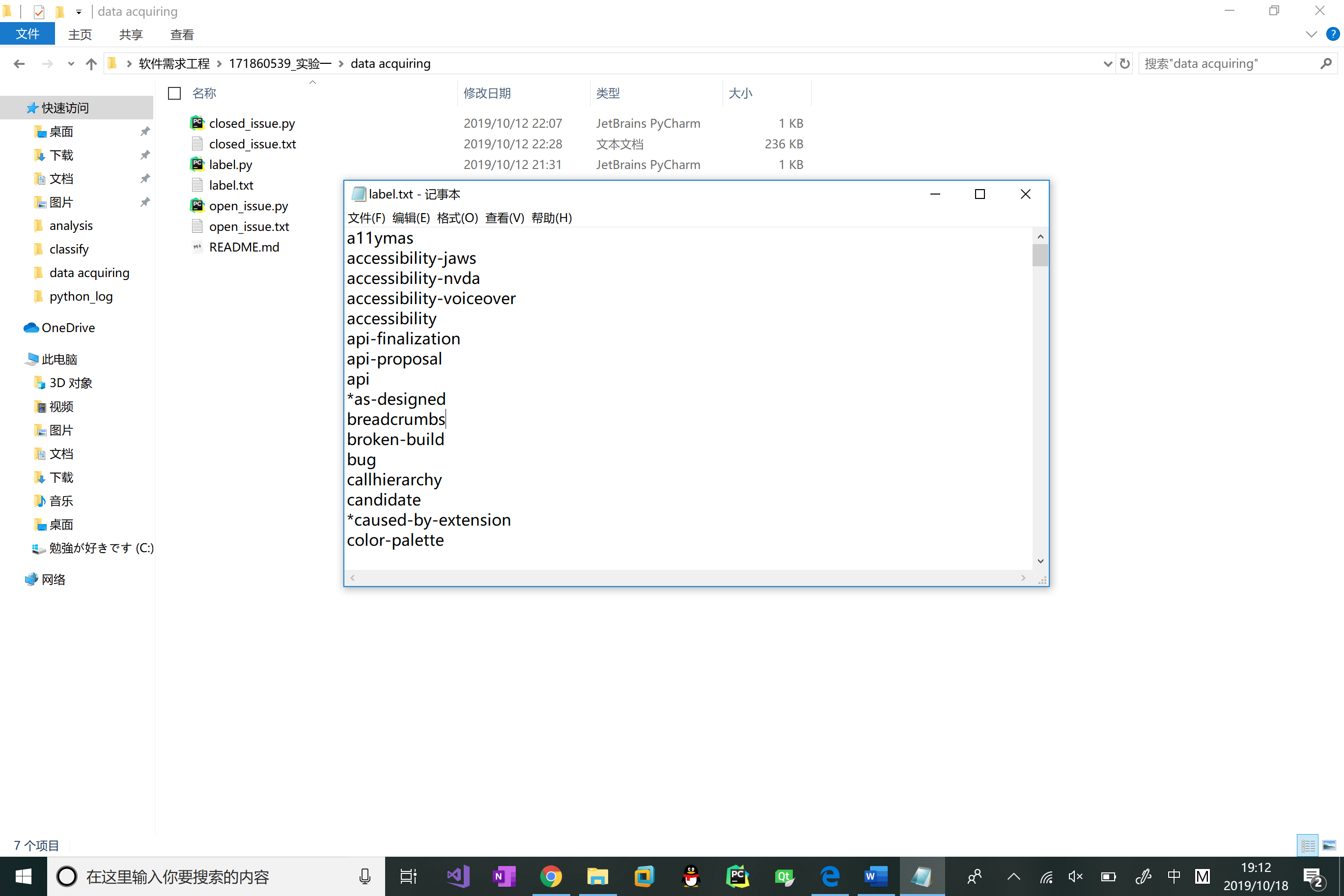


总结：

* autopep,vim,keybinding,gitlen是一些呼声比较高的工具

结果：虽然上述实验结果中的聚类内部确实能找到一些联系，但相互之间的关联仍然不够，并且对结果进行了一轮筛选的条件下效果依然并不理想。因此我们在获取了Vscode项目的issue中的labels的基础上，通过人为地对labels进行一些预处理后作为聚类分析的结果，再对issue重新进行了一次分类。步骤和结果如下：

首先是获取的labels：



如果把issues看作一个个需求，那么这些labels天然的就是issues聚类，比如accessibility，api，bug这些labels，每一个都包含了很多issues，但是像a11ymas这类词，它包含的issues可能非常少，所以它可能并不会出现在最终的需求中，但是它也出现在了labels中，所以考虑对labels进行进一步处理。

首先将类似的labels合并，比如accessibility-jaws，accessibility-nvda，accessibility-voiceover就可以合并为accessibility，同时去除连字符，将大写字母变小写字母，以便后续匹配。

然后将labels与issues配合起来，读入所有issues，若issues含有某个label，则该label含有的issue计数值加1，最后将计数值大于20的labels作为最后的general的需求（也即对issues分类的标准），称这些筛选过的labels为keyword。

这些keyword实际就是可能的需求，比如：

(1) ：代表vscode有关于辅助功能的需求，

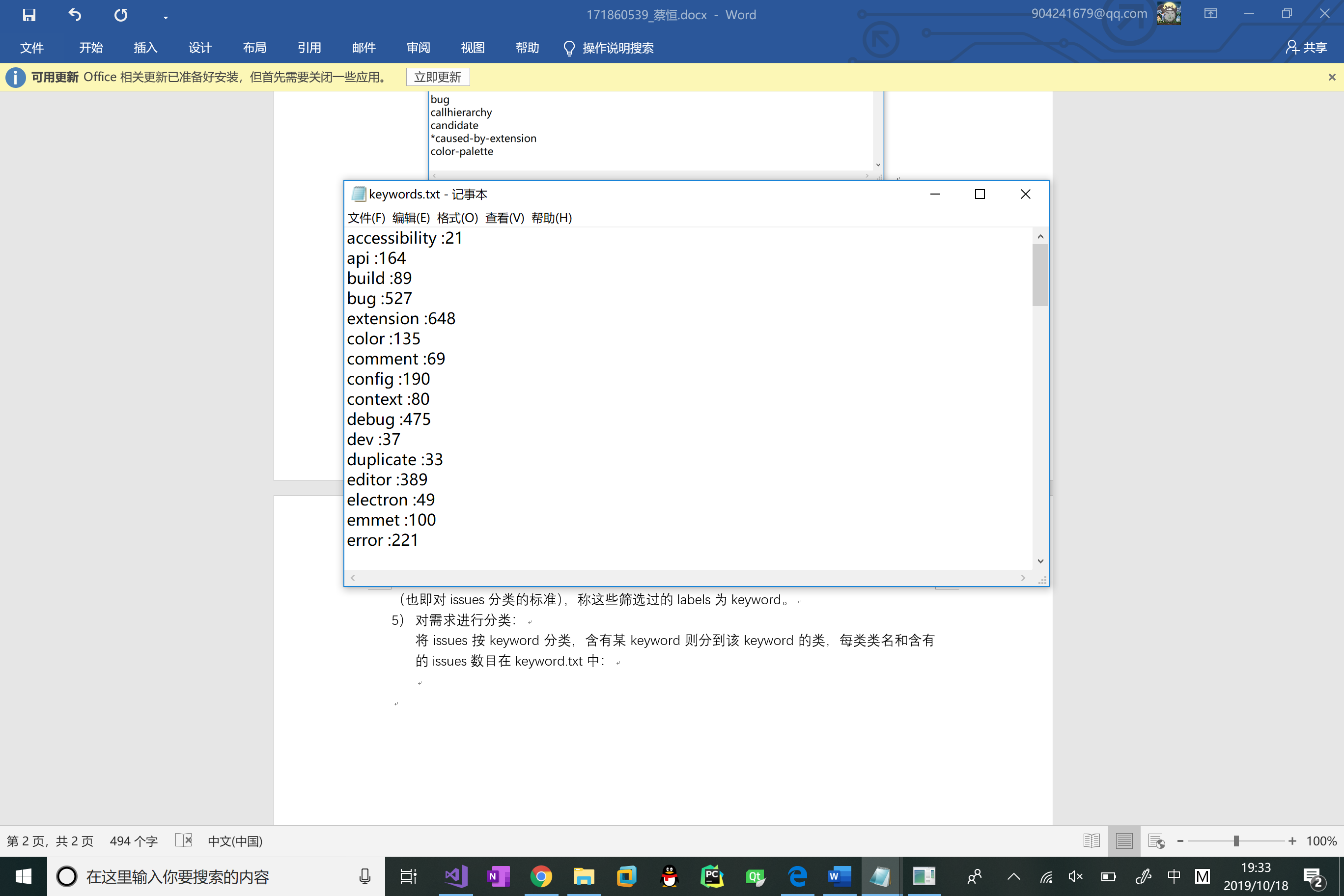
(2) ：表示vscode提供了intellisense智能提示功能，在issues中出现的次数很多，说明这个功能还不够完善，有许许多多的问题。

(3) ：关于bug的issues有这么多，说明vscode对于debug有需求，同时，vscode IDE本身可能还有一些bug，有消除这些bug的需求。

但这些keyword只是非常粗的类别，具体的需求在issues里，下面将issues按这些keyword分类。

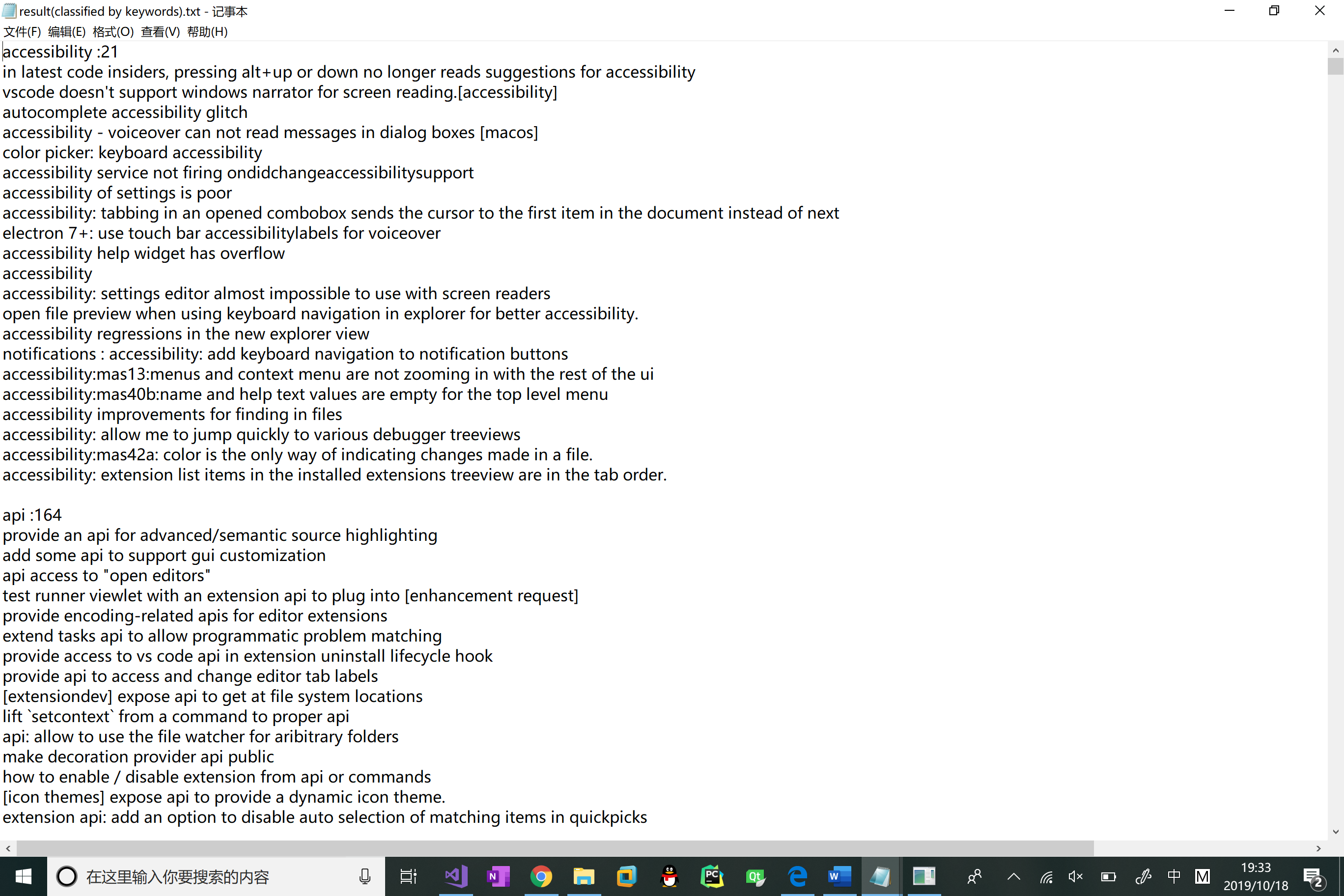
1. 对需求进行分类：

将issues按keyword分类，含有某keyword则分到该keyword的类，每类类名和含有的issues数目记录在keyword.txt中，结果概要如下：



最终将10030条issue分到了79个类别里

详细的分类结果在result.txt中：



这里能看到每个分好的类里有什么具体的需求。

此处列举部分需求的类别以及含有的部分具体需求如下：

(1) 辅助功能：

-> 语音辅助voiceover，能够播放对话框的的信息

-> 键盘导航，有丰富的命令列表，让用户可以在不使用鼠标的情况下运行vscode

-> 屏幕阅读器

(2) API:

-> 支持gui定制的api

-> 允许扩展api的本地配置更新

-> 支持调用层次结构视图的api

-> 支持用户数据同步的api

(3) 项目的构建与生成：

-> 加快生成任务的启动速度

-> 在生成的同时自动保存

-> 支持特定语言的生成任务

(4) 调试：

-> 在调试器中可以查看十进制或十六进制中的值

-> 调试控制台可以自动滚动

-> 调试控制台支持查找

-> 调试控制台支持多页选择

-> 支持节点调试器自动附加的节点选项检测

(5) 扩展：

-> 支持平台特定扩展

-> 从配置文件启用/禁用扩展名

-> 允许扩展为“快速打开”提供其他路径

(6) 颜色：

-> 支持打开颜色选择器并插入选定格式的颜色的快捷方式

-> 为文件夹添加标签/颜色

-> 在存在问题时用颜色突出显示问题指示符

除此以外还有很多，完整的需求的类别请参看keyword.txt，完整的分类后的详细需求请参看result.txt，

StackOverFlow提问需求获取：

arduino 
studio build module 
apache 
web 
type 
github 
edit6il 
v Irm 
data 
template 
auto 
python 
raspberry 
embedded 
upwuser 
android 
typescript 
firebase 
directory 
visÜa1nnstudio 
intel 
dynamic 
wqnclow 
inte 1 i J 
plugin 
installation 
[cleaerru 
virtual 
vba 
commi 
emacs 
linport E 
lisp 
environment 
for tran 
Invc 
xcodeID 
o 
spring 
deployment 
mobile 
sq 
compiler 
static 
scrip - 
webs torm 
memory 
file 
lirpux 
project 
net core 
access 
structure 
gradle 
excel 
.pa th Coo 
bash 
ou]ao 
inst 
jupyter 
keyboard 
j sapi 
Q) google react 
-esource 
connentscn 
ava 
php ser-ver- 
builder 
tool 
browser 
shortcut 
xamarln 
header character 

需求分析：

上图是我们对stackoverflow上的3000条提问所属于的标签的爬取，可以看到最醒目的是几个编程语言的名称（如java, python）、编程IDE的名称（如visual studio，eclipse，pycharm，netbeans）以及编程平台的名称（Android，Linux）。另外还有一个醒目的单词"debugging"，程序员确确实实会遇到相当多的debug困难。此外，还有关于web开发关键词web。可以猜测出，IDE要支持几种编程语言，并且要有一定的debug工具。

次醒目的是几个次常用的编程语言，如js，net，php；几个常用的代码开发框架，如maven，gradle，spring；这里可以看出IDE的基本需求肯定是支持语言。并且，要能够支持提供代码框架节省开发时间。

再细心一些，我们会发现用关于版本控制的标签，比如git,github，可能还用cloud等等。这里就可以看出IDE需要提供云存储、版本控制的功能。

再仔细看，我们可以看到比如engine,designer(有可能是QT designer等）,interface,builder。有些是开发引擎，有些是构建器（比如Qt的builder），有些是界面、接口。

这些可能体现出比较高级的需求。根据这些关键词查到对应的句子，我们可以发现：

1. 有关于engine的句子

如:

(1) ”Engine for simple client-server two-player turn-based game with random match-making and online statistics”

这是程序员有开发游戏引擎的需求，他们需要IDE提供一些现成的方便改进的引擎

(2) “WebStorm syntax highlighting for Swig template engine”

这是js模板引擎

2. 有关于designer的句子

“VB6 designer doesn't display opened modules”

这是VB6提供的设计器，问答显示他的软件出现了双击打不开模块的bug

这些能体现一些IDE比较smart的部分，他们需要提供比较丰富的组件

3. 我们还能从词云图当中看到mysql，Oracle,nosql等词汇。其中有一条十分简洁的问题是这样的：

“How to connect mysql with eclipse?”

说明IDE应该能够支持一些访问数据库的操作，提供访问数据库的接口。

此外，

“is there a valid extension for connect **mysql** via ssh on visual code studio? I can't find nothing!; I've been tryed with **Mysql** managment tool but seem there are not options for ssh …”

可以看出vscode对数据库的访问支持还不够，此需求可以作为smart需求的一部分

我们还可以发现vscode与数据库连接的操作问题比较多, 比如:

* i'm not able to connect new version of vscode and new version of mysql
* Suggestions for NodeJS+MySQL
* Connect to a MySQL datatbase using ssl certificates in using C# Dotnet Core 3 vs code error when i run some simple mysql connecting code with code runner
* …

除去连接问题以外，我们还能看到一些直接的建议：

[Suggestions for NodeJS+MySQL](https://stackoverflow.com/questions/46902076/suggestions-for-nodejsmysql?r=SearchResults)

可以看出数据库对语言的支持还需要完善

4. 在前几次课的讨论访谈中，我们也讨论到IDE对于插件的支持。在此图中出现了extension这个字眼，其中的需求有：如何导入插件？如何离线下载插件？等等