

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、小组成员及分数比例：

魏宇阳 181860108---25%

罗沁禹 171860537---25%

赵宇鹏 181860146---25%

罗树汉 181860062---25%

三、小组分工：

罗沁禹：stackoverflow vscode 标签和 ide 标签问题获取，对标签词云分析，对所有问题词汇聚类分析，分析可能的需求

魏宇阳：vscode 的 open issues，closed issues，labels 数据爬取和按关键词分类，分析获取需求

赵宇鹏：对 vscode 的 PR 数据进行分析，并进行按关键词分类，分析获取需求

罗树汉：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 issue 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>

- 2) 明确信息源：

将这个项目的 open issues 和 closed issues 作为主要的数据来源。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:

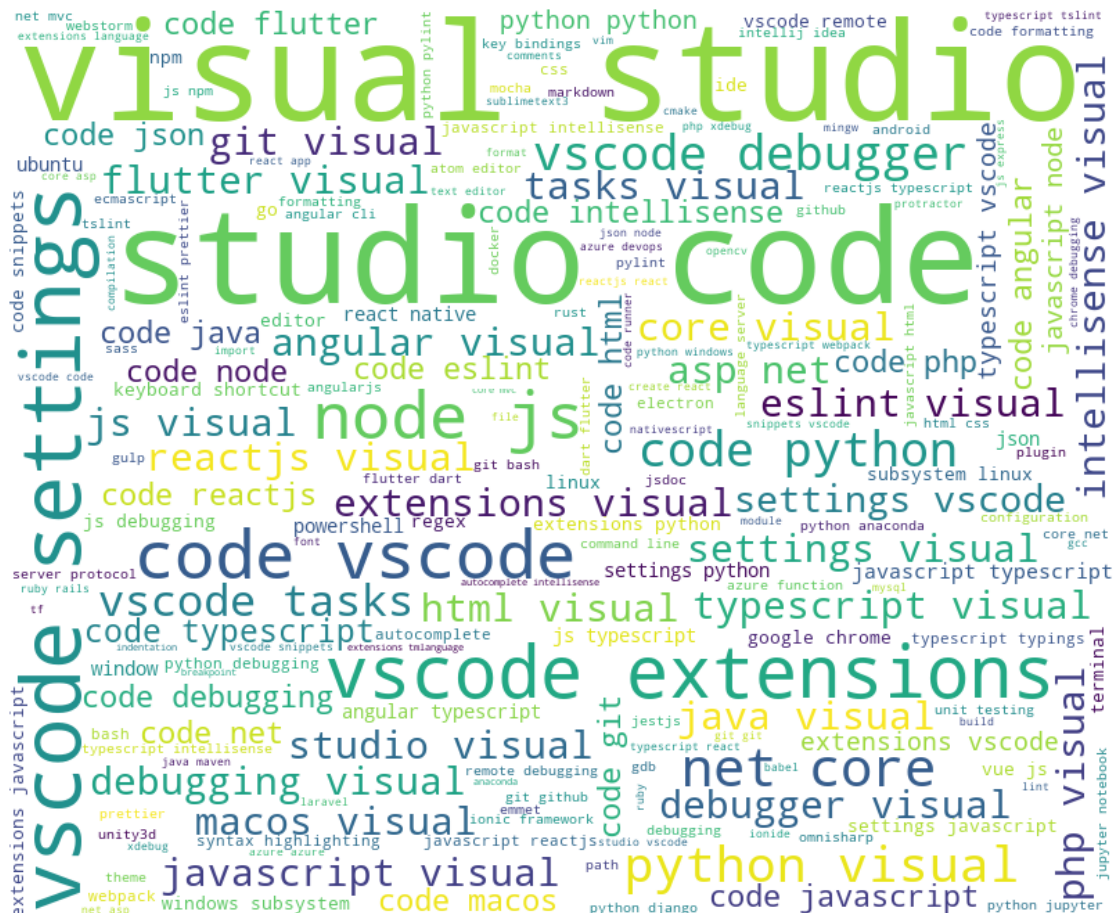
<https://github.com/microsoft/vscode/labels?page=i&sort=name-asc>

3) 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。

4) 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：



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：

```
guorutjie@ubuntu:~/Desktop/word2vec$ sudo ./word2vec -train title.txt -output vectors.bin -cbow 0 -size 200 -window 5 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1
[sudo] password for guorutjie:
Starting training using file title.txt
Vocab size: 2658
Words in train file: 171796
Alpha: 0.011897 Progress: 64.06% Words/thread/sec: 209.63k guorutjie@ubuntu:~/Desktop/word2vec$
```

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)

```
guoruijie@ubuntu:~/Desktop/word2vec$ ./distance_vectors.bin
Enter word or sentence (EXIT to break): edit
Word: edit Position in vocabulary: 401
```

Word	Cosine distance
put	0.971528
keybinding	0.971106
define	0.961359
ignore	0.958893
link	0.953084
filter	0.948403
popup	0.945826
remove	0.943844
check	0.942889
provide	0.941597
extension's	0.941484
specify	0.941098
modify	0.940303
know	0.938038
order	0.934269
predefined	0.934071
whenever	0.933377
undo	0.931678
listen	0.929978
vscode?	0.929165
should	0.929103
view	0.929094
need	0.928339
navigate	0.928305
figure	0.927690
wrapping	0.927446
background	0.926800
access	0.926455
command?	0.926391
sync	0.925791
manually	0.925369
where	0.925318
share	0.923145
auto-indent	0.923064
config	0.921190
rid	0.920822
MySQL	0.920206
tree	0.919284
press	0.919063
used	0.918423

Word	Cosine distance
put	0.977223
link	0.972441
ignore	0.972297
semicolon	0.961285
filter	0.959366
compare	0.956584
config	0.948382
access	0.947231
where	0.947196
jsdoc	0.946532
reopen	0.945149
provide	0.943782
predefined	0.943717
trigger	0.941801

Word
repeat
Explorer
support
status
key
Access
typed
attribute
activity
formatter
api
coloring
There
Find
deployment
word
branch
What's
location
snippet
Customize
Multiple
Do
This
URI
multi
highlight
Disabling
weight
longer
update
tmLanguage
"
conflict
TextMate
highlighting
underline
directory,
Move
collapse

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

2. 嵌入式开发(embed)

Word	frameworks
debugging?	do?
related	from?
code's	twice
copying	constructor
babel-node	identifier
	configuring

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)

3. 网络编程 (web,net,.net,html)

```
reak): web
: 330
Word
-----
Delve
SFTP
os
angular
Nativescript
installation
solution
angular2
IIS
Ubuntu?
crashes
break
Laravel
Unity
x
Firebase
gdb
unverified
Rust
android
Pylint
writing
dependency
adds
dot
exception
Webpack
Problem
NPM
maven
attached
insiders
Current
deleted
actions
built
Docker
drive
create-react-app
jsconfig.json
```

```
configuration
user-defined
undo
servers
time
mono
Config
particular
folders
insertion
actually
NOT
these
first
expand
out
returned
mark
back
jsdoc
block
WebStorm?
long
let
tkinter
```

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

4. Debug

总结: (右图)

- > 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)

5. 插件(extension) (聚类分析结果不够理想)

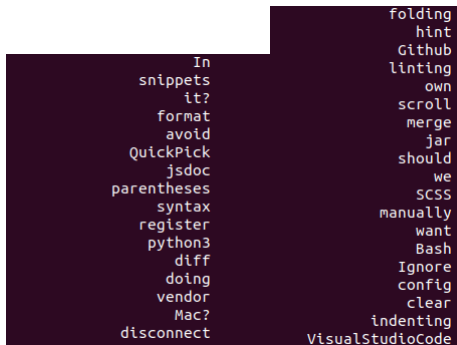
6. 版本控制(version,git,github)

```
auto-format
commit
ASP.Net
hint
Icon
```

```
break): debug
ary: 32
Word
-----
setup
run
c++
code,
compile
mac?
AngularJS
Can't
execute
application
task
clean
nodejs
build
.Net
develop
interpreter
6
electron
connect
django
install
design
existing
ubuntu?
basic
subfolder
java
Trying
attach
c#
.net
react
reset
launch
node
alignment
flutter
extension.
clear
```

> 自动格式化

7. 智能(smart,automatic)

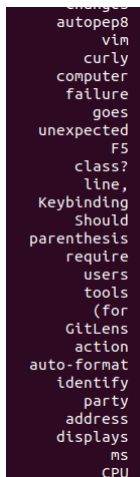


总结:

> 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

8. 必要的工具



总结:

> autopep,vim,keybinding,gitleen 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。

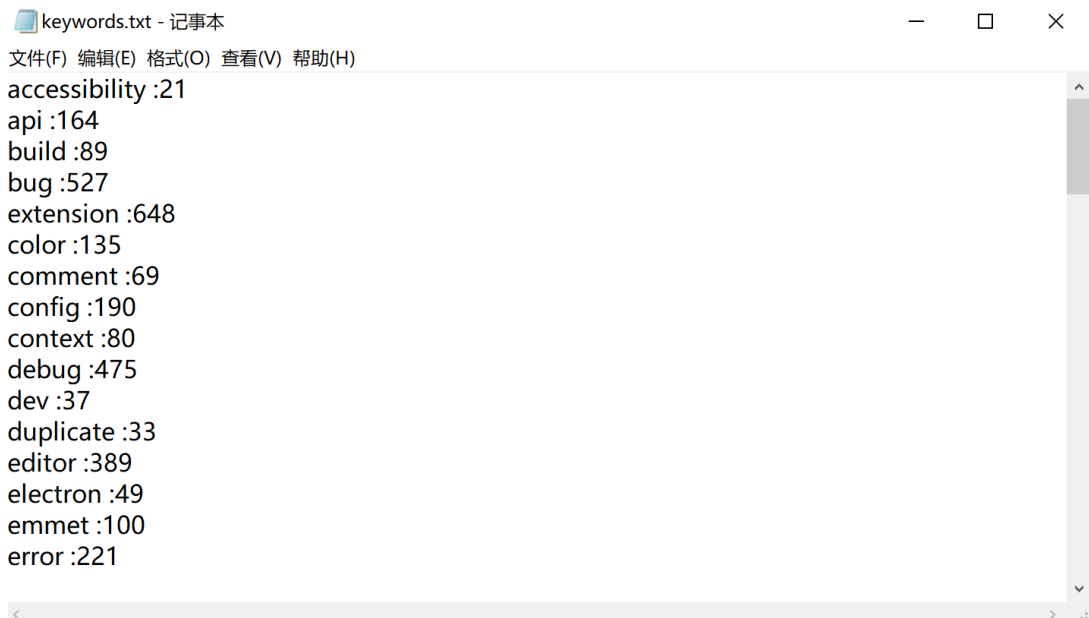
这些 keyword 实际就是可能的需求, 比如:

- (1) **accessibility :21** : 代表 vscode 有关于辅助功能的需求,
- (2) **intellisense :129** : 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) **bug :527** : 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。

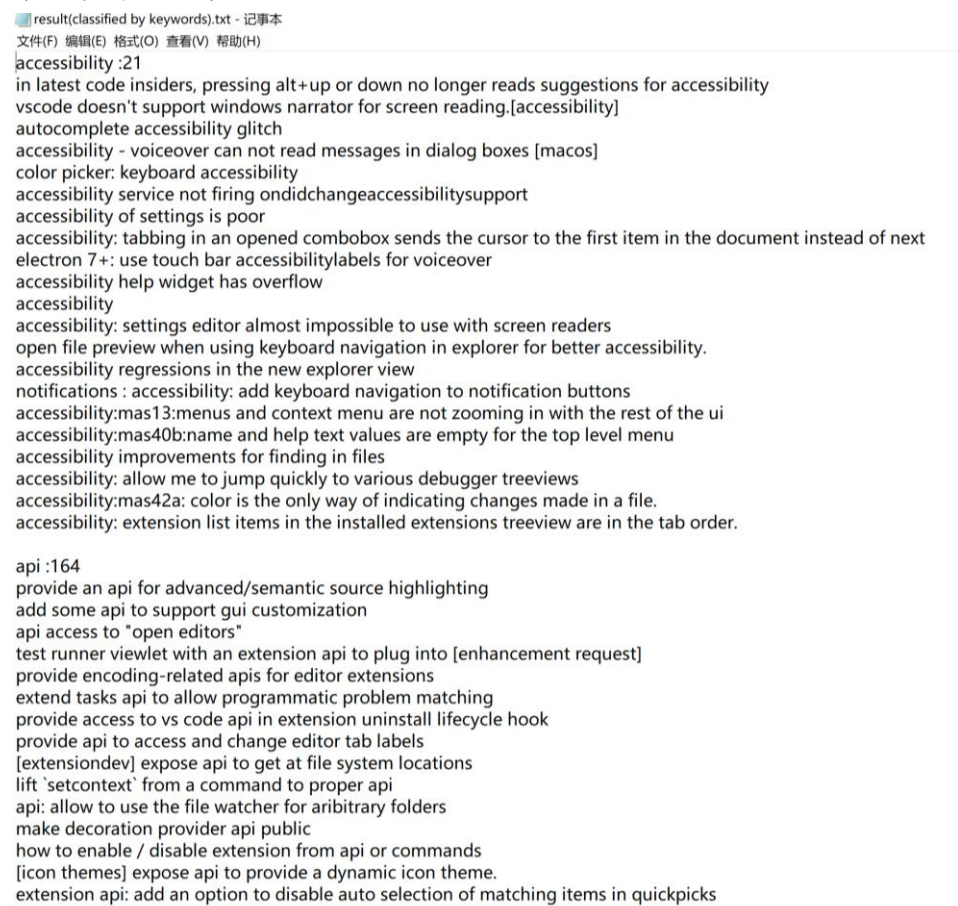
5) 对需求进行分类:

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,

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 tried 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](#)

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