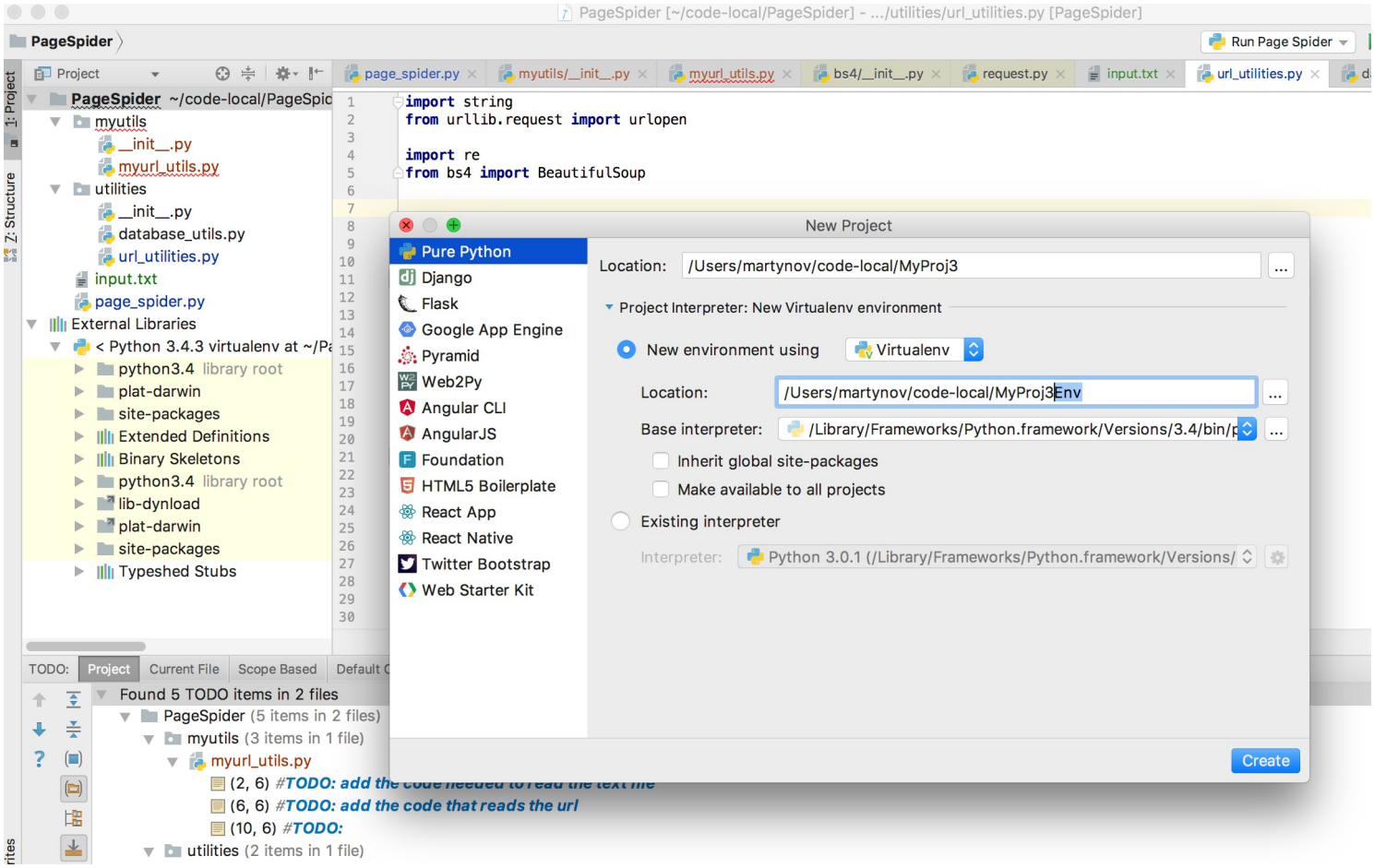
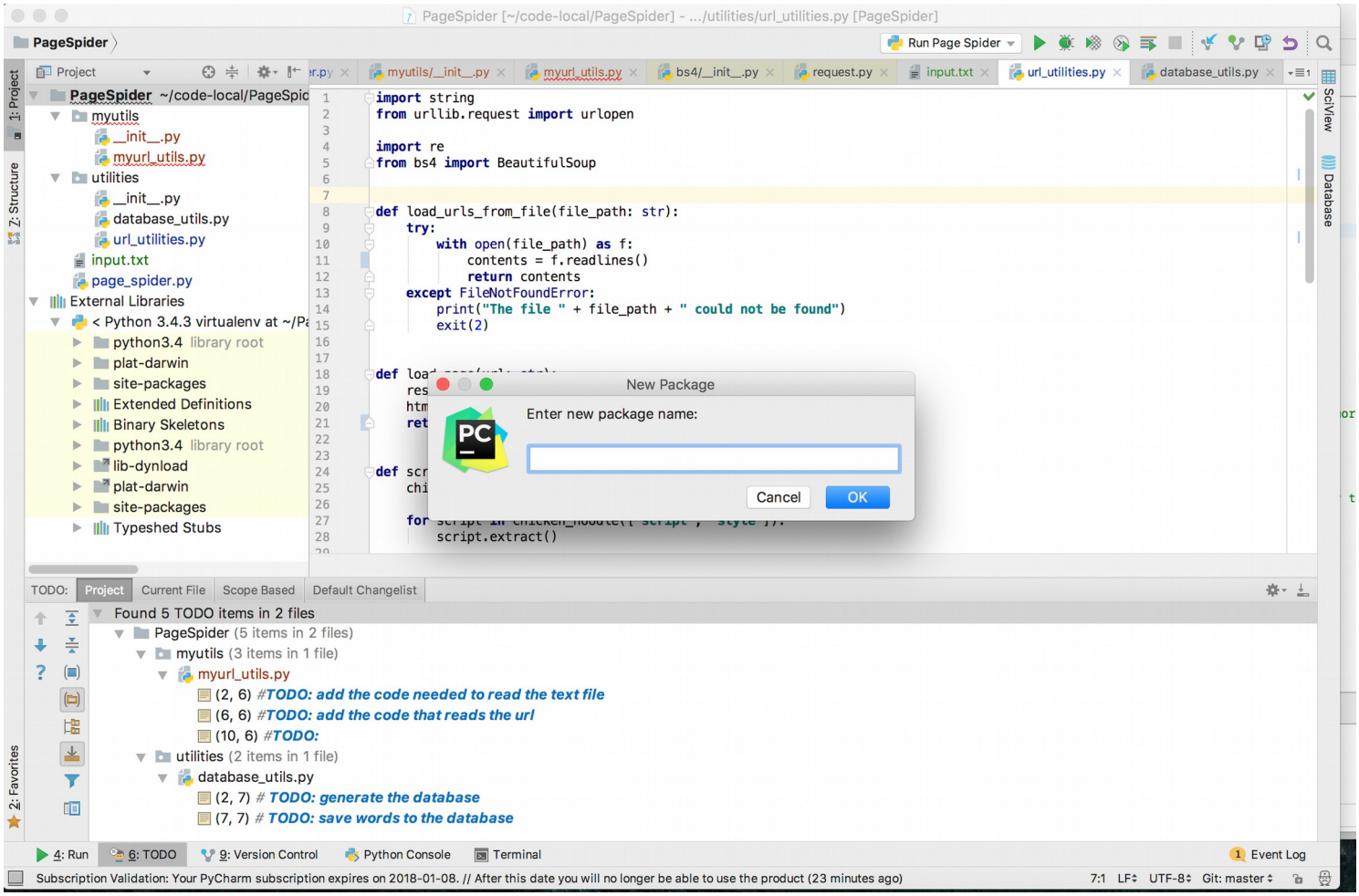
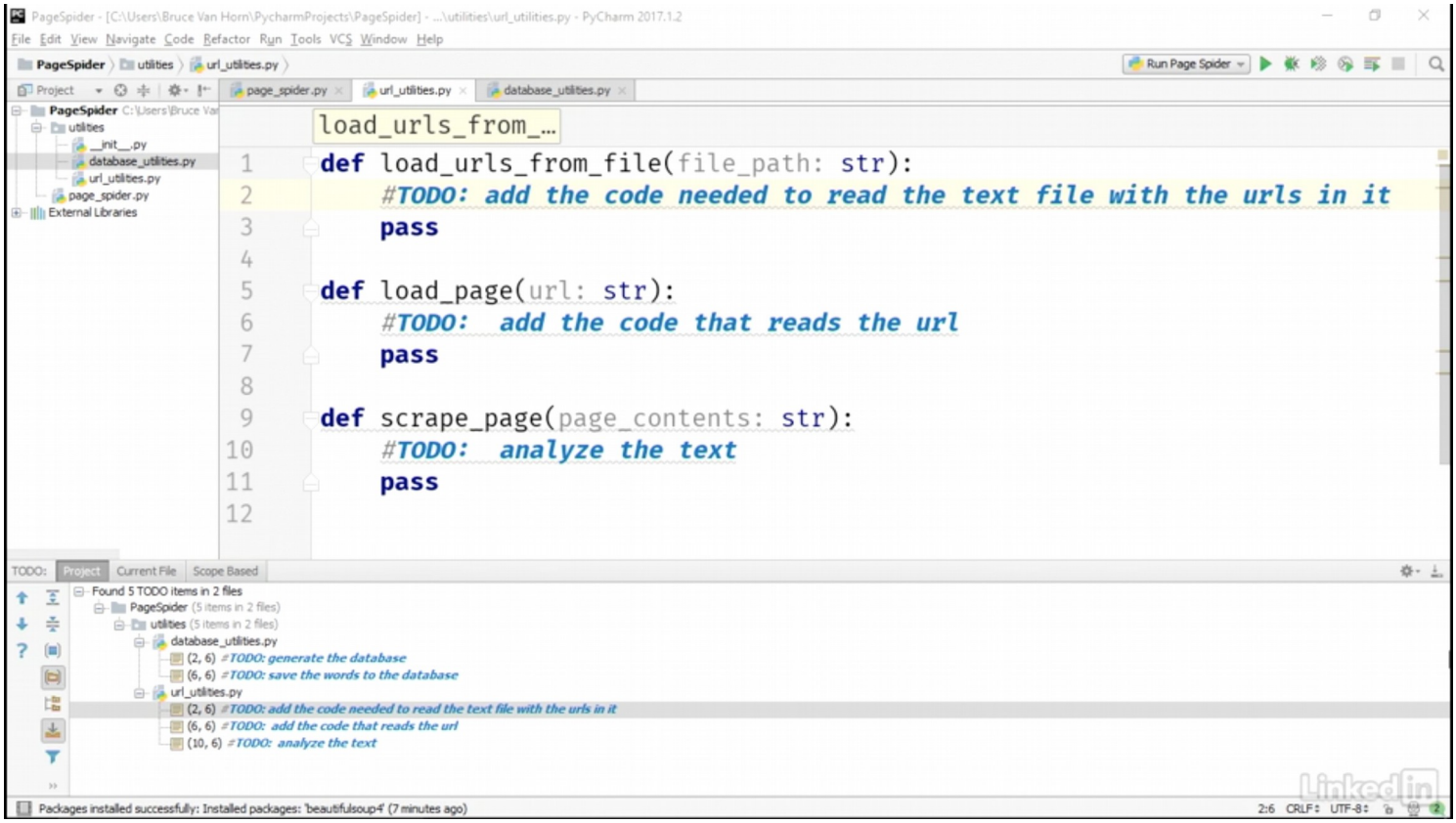
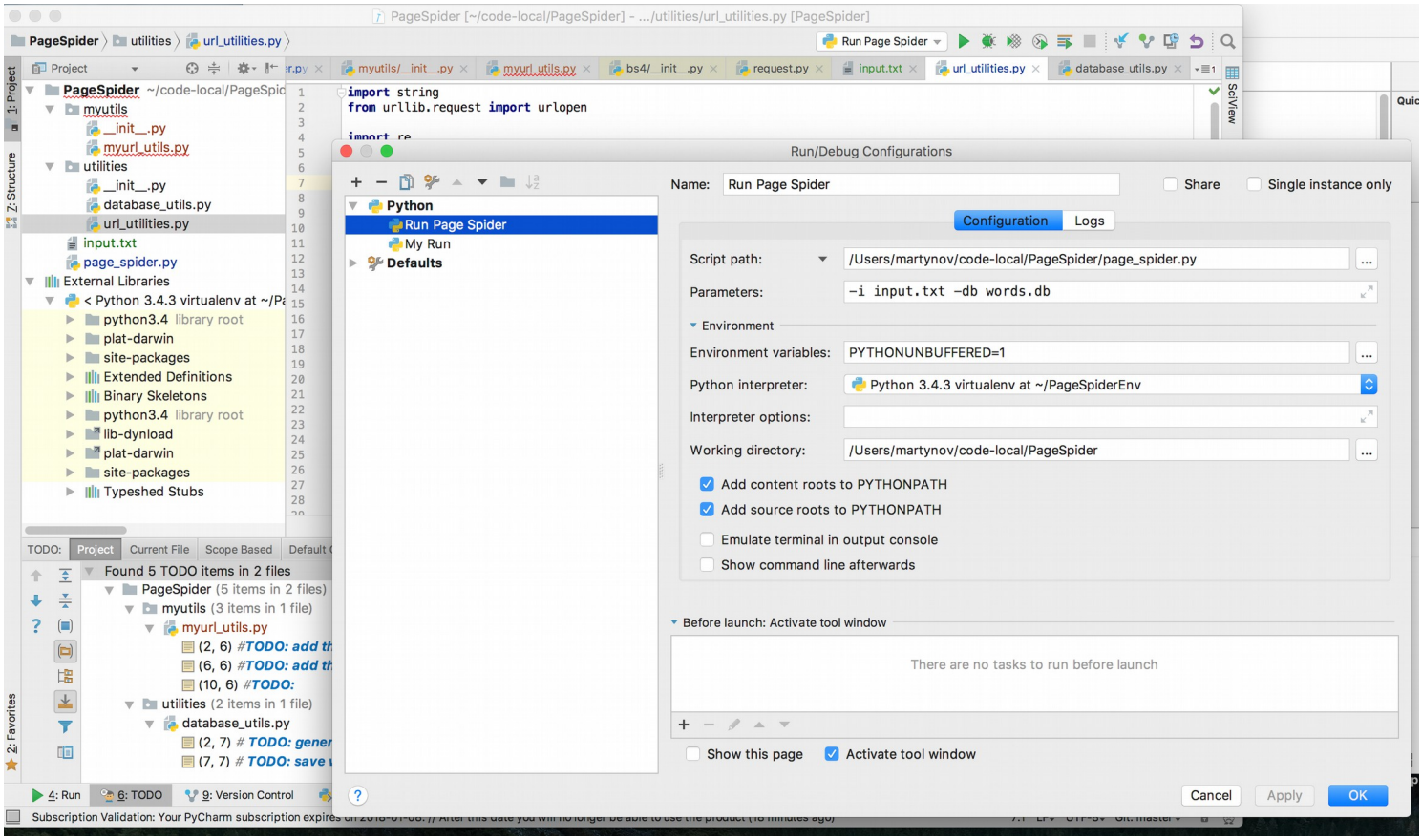
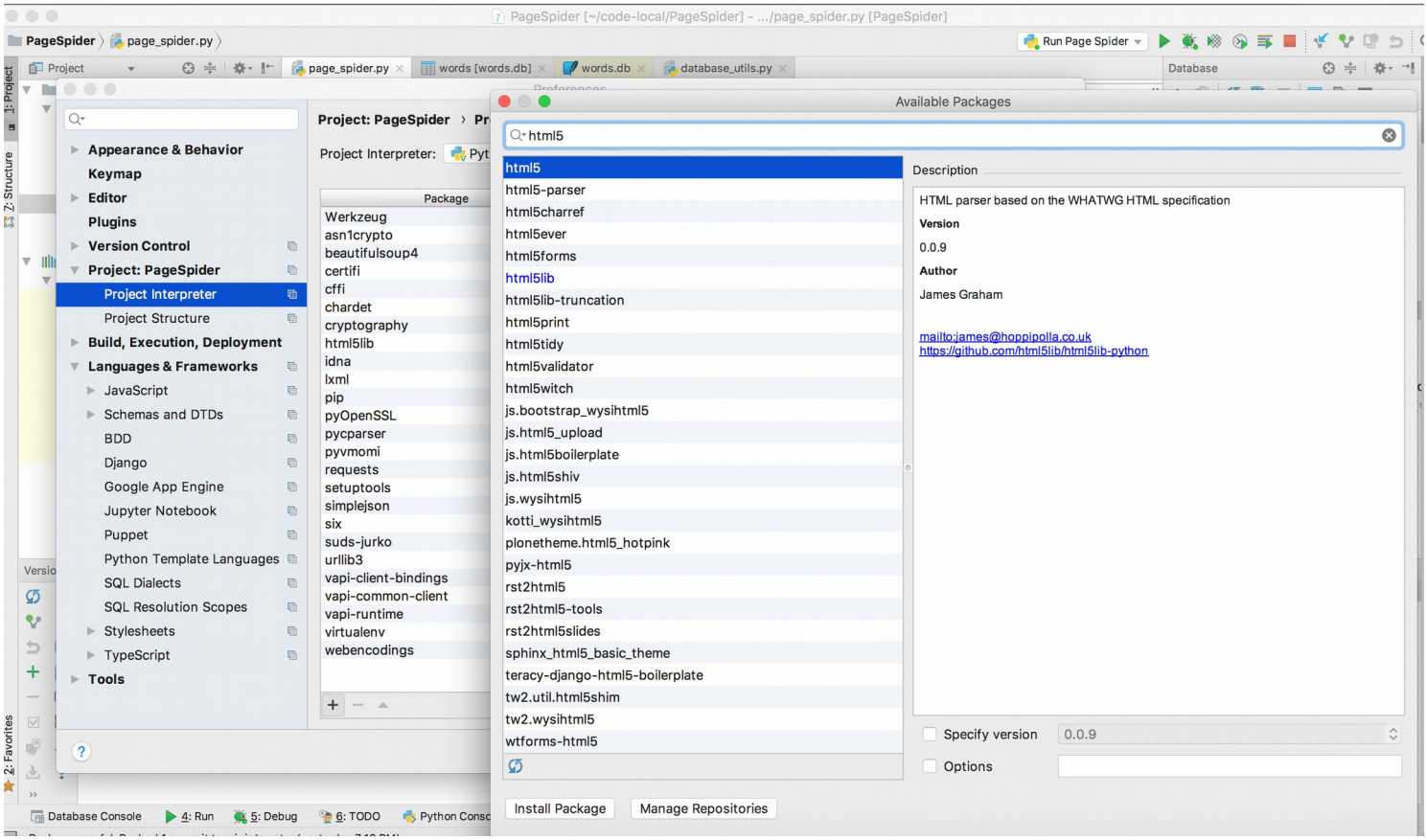
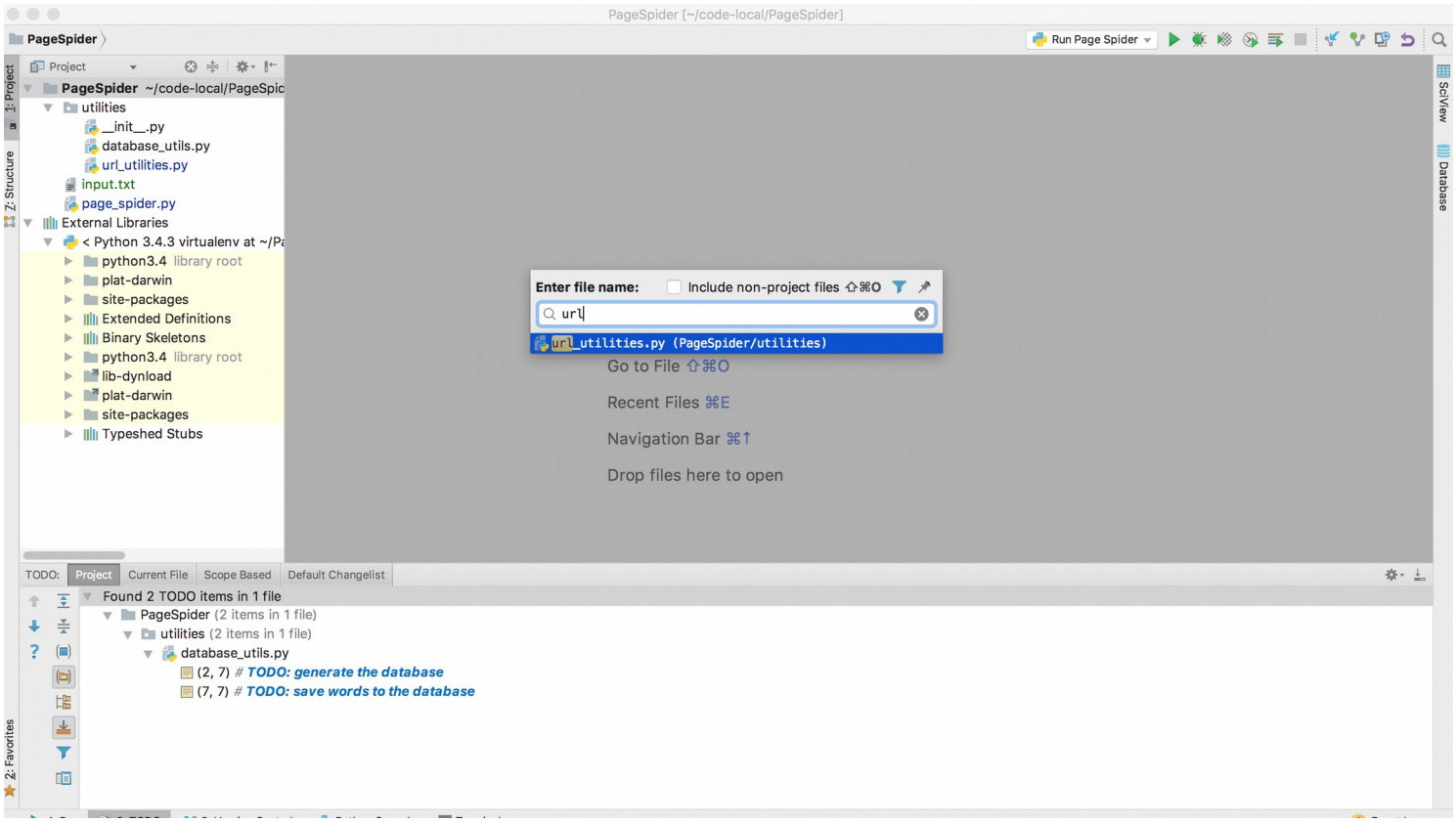

PageSpider [~/code-local/PageSpider] - .../utilities/url_utilities.py [PageSpider]
PageSpider
Run Page Spider
Project
page_spider.py
myutils/__init__.py
myurl_utils.py
bs4/__init__.py
request.py
input.txt
url_utilities.py
PageSpider ~/code-local/PageSpic
myutils
__init__.py
myurl_utils.py
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/P
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
import string
from urllib.request import urlopen
import re
from bs4 import BeautifulSoup
New Project
Pure Python
Django
Flask
Google App Engine
Pyramid
Web2Py
Angular CLI
AngularJS
Foundation
HTML5 Boilerplate
React App
React Native
Twitter Bootstrap
Web Starter Kit
Location: /Users/martynov/code-local/MyProj3
Project Interpreter: New Virtualenv environment
New environment using Virtualenv
Location: /Users/martynov/code-local/MyProj3Env
Base interpreter: /Library/Frameworks/Python.framework/Versions/3.4/bin/p
Inherit global site-packages
Make available to all projects
Existing interpreter
Interpreter: Python 3.0.1 (/Library/Frameworks/Python.framework/Versions/
Create
TODO:
Project
Current File
Scope Based
Found 5 TODO items in 2 files
PageSpider (5 items in 2 files)
myutils (3 items in 1 file)
myurl_utils.py
(6, 6) #TODO: add the code that reads the url
(10, 6) #TODO:
utilities (2 items in 1 file)

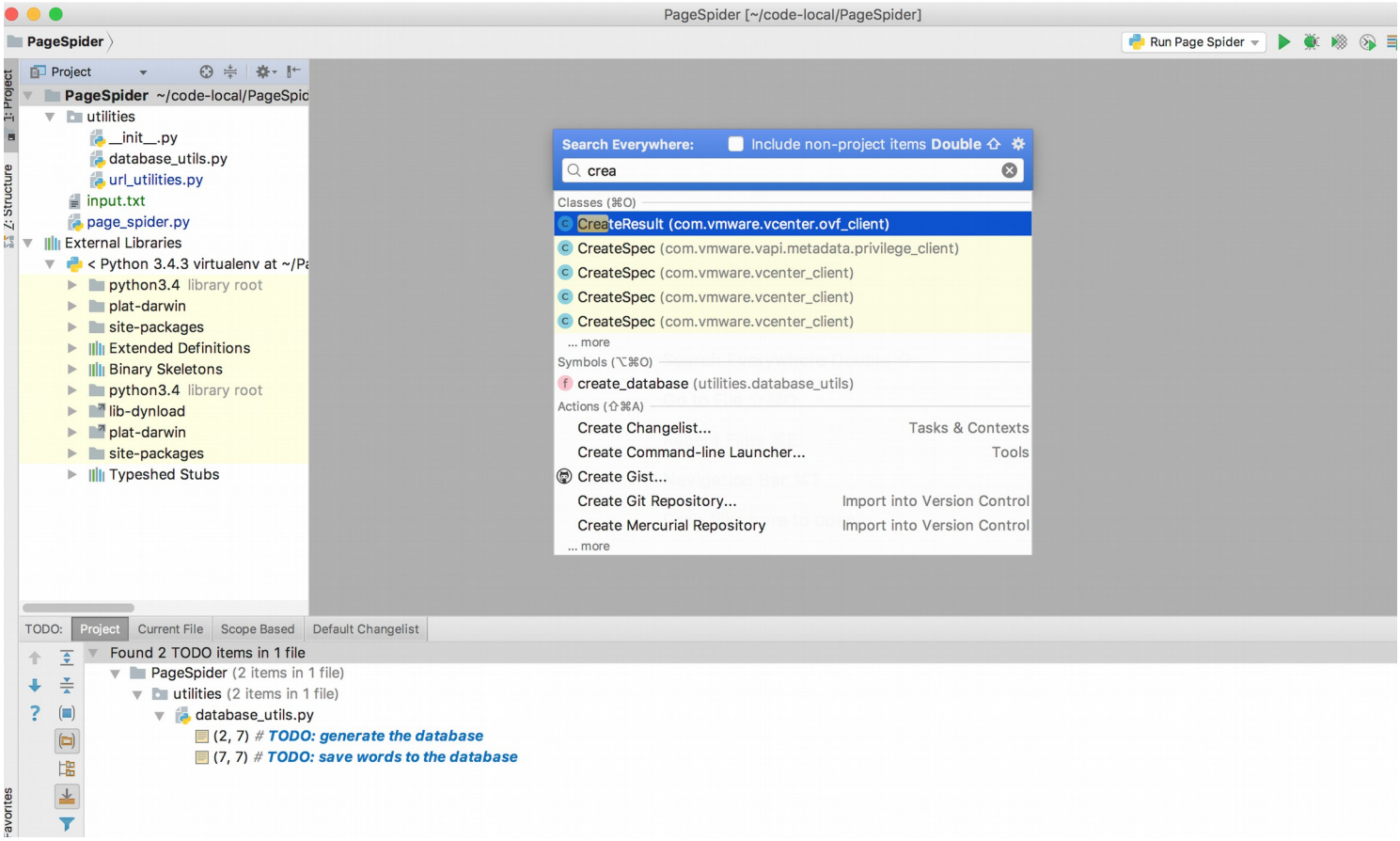

PageSpider [~/code-local/PageSpider] - .../utilities/url_utilities.py [PageSpider]
PageSpider
Run Page Spider
Project
PageSpider ~/code-local/PageSpid
myutils
__init__.py
myurl_utils.py
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
myutils/__init__.py
myurl_utils.py
bs4/__init__.py
request.py
input.txt
url_utilities.py
database_utils.py
import string
from urllib.request import urlopen
import re
from bs4 import BeautifulSoup
def load_urls_from_file(file_path: str):
try:
with open(file_path) as f:
contents = f.readlines()
return contents
except FileNotFoundError:
print("The file " + file_path + " could not be found")
exit(2)
script.extract()
New Package
Enter new package name:
Cancel
OK
TODO:
Project
Current File
Scope Based
Default Changelist
Found 5 TODO items in 2 files
PageSpider (5 items in 2 files)
myutils (3 items in 1 file)
myurl_utils.py
(2, 6) #TODO: add the code needed to read the text file
(6, 6) #TODO: add the code that reads the url
(10, 6) #TODO:
utilities (2 items in 1 file)
database_utils.py
(2, 7) # TODO: generate the database
(7, 7) # TODO: save words to the database
4: Run
6: TODO
9: Version Control
Python Console
Terminal
Event Log
Subscription Validation: Your PyCharm subscription expires on 2018-01-08. // After this date you will no longer be able to use the product (23 minutes ago)
7:1
LF
UTF-8
Git: master
1: Project
2: Favorites
7: Structure
SciView
Database

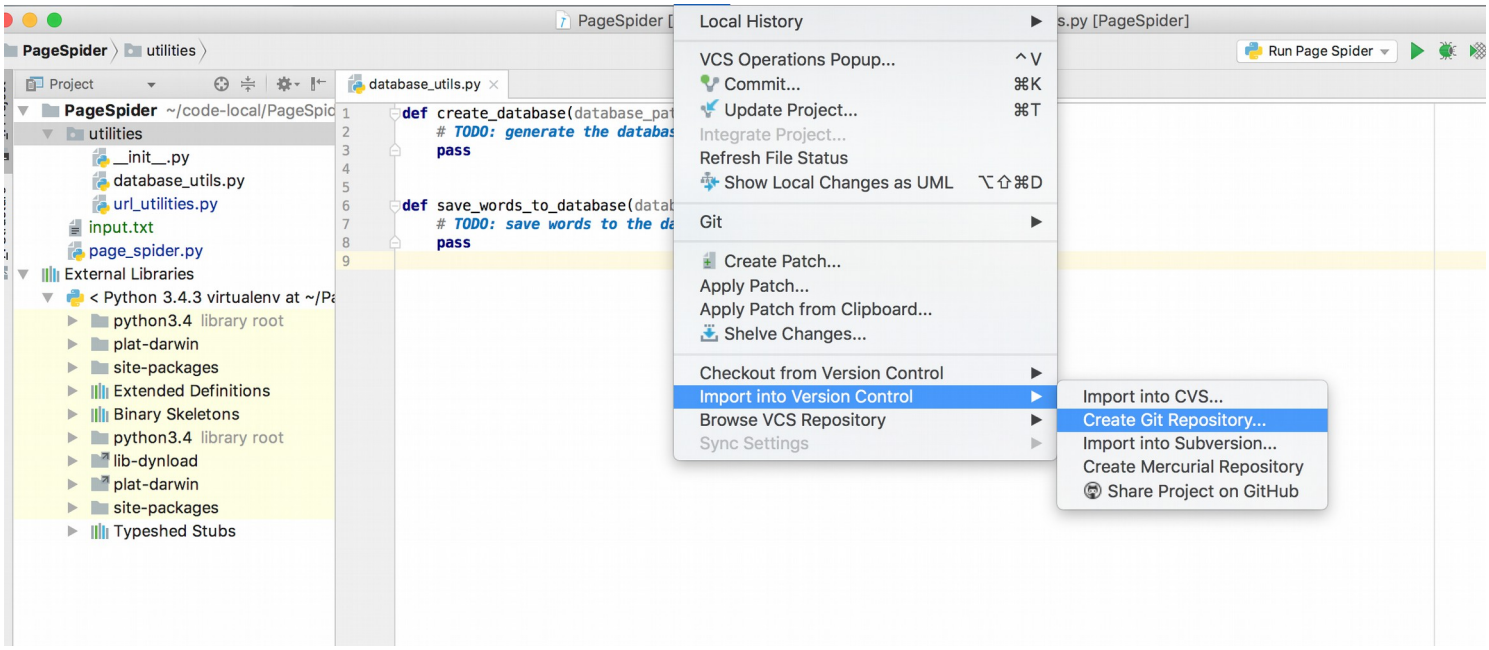

PageSpider - [C:\Users\Bruce Van Horn\PycharmProjects\PageSpider] - ...\utilities\url_utilities.py - PyCharm 2017.1.2
File Edit View Navigate Code Refactor Run Tools VCS Window Help
PageSpider utilities url_utilities.py
Run Page Spider
Project
page_spider.py
url_utilities.py
database_utilities.py
PageSpider C:\Users\Bruce Van
utilities
__init__.py
database_utilities.py
url_utilities.py
page_spider.py
External Libraries
load_urls_from_...
1 def load_urls_from_file(file_path: str):
2 #TODO: add the code needed to read the text file with the urls in it
3 pass
4
5 def load_page(url: str):
6 #TODO: add the code that reads the url
7 pass
8
9 def scrape_page(page_contents: str):
10 #TODO: analyze the text
11 pass
12
TODO:
Project
Current File
Scope Based
Found 5 TODO items in 2 files
PageSpider (5 items in 2 files)
utilities (5 items in 2 files)
database_utilities.py
(2, 6) #TODO: generate the database
(6, 6) #TODO: save the words to the database
url_utilities.py
(2, 6) #TODO: add the code needed to read the text file with the urls in it
(6, 6) #TODO: add the code that reads the url
(10, 6) #TODO: analyze the text
Packages installed successfully: Installed packages: 'beautifulsoup4' (7 minutes ago)
2:6 CRLF UTF-8

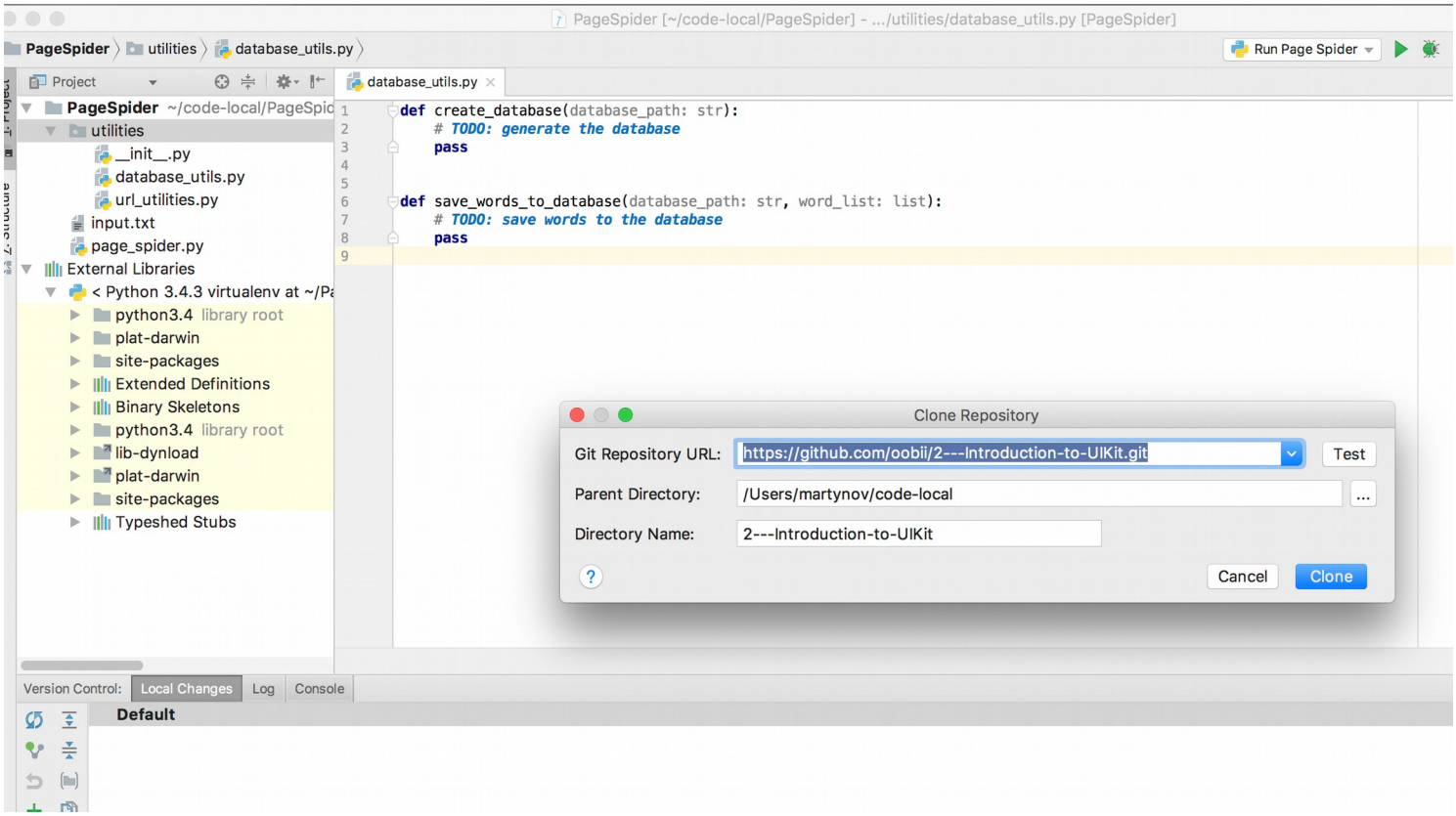

PageSpider [~/code-local/PageSpider] - .../utilities/url_utilities.py [PageSpider]
Run Page Spider
import string
from urllib.request import urlopen
Run/Debug Configurations
Name: Run Page Spider
Share
Single instance only
Python
Run Page Spider
My Run
Defaults
Configuration
Logs
Script path: /Users/martynov/code-local/PageSpider/page_spider.py
Parameters: -i input.txt -db words.db
Environment
Environment variables: PYTHONUNBUFFERED=1
Python interpreter: Python 3.4.3 virtualenv at ~/PageSpiderEnv
Interpreter options:
Working directory: /Users/martynov/code-local/PageSpider
Add content roots to PYTHONPATH
Add source roots to PYTHONPATH
Emulate terminal in output console
Show command line afterwards
Before launch: Activate tool window
There are no tasks to run before launch
Show this page
Activate tool window
Cancel
Apply
OK


PageSpider [~/code-local/PageSpider] - .../page_spider.py [PageSpider]
PageSpider
page_spider.py
Run Page Spider
Project
page_spider.py
words [words.db]
words.db
database_utils.py
Database
Preferences
Appearance & Behavior
Keymap
Editor
Plugins
Version Control
Project: PageSpider
Project Interpreter
Project Structure
Build, Execution, Deployment
Languages & Frameworks
JavaScript
Schemas and DTDs
BDD
Django
Google App Engine
Jupyter Notebook
Puppet
Python Template Languages
SQL Dialects
SQL Resolution Scopes
Stylesheets
TypeScript
Tools
Project: PageSpider
Project Interpreter:
Package
Werkzeug
asn1crypto
beautifulsoup4
certifi
cffi
chardet
cryptography
html5lib
idna
lxml
pip
pyOpenSSL
pycparser
pyvmomi
requests
setuptools
simplejson
six
suds-jurko
urllib3
vapi-client-bindings
vapi-common-client
vapi-runtime
virtualenv
webencodings
Available Packages
html5
html5
html5-parser
html5charref
html5ever
html5forms
html5lib
html5lib-truncation
html5print
html5tidy
html5validator
html5witch
js.bootstrap_wysihtml5
js.html5_upload
js.html5boilerplate
js.html5shiv
js.wysihtml5
kotti_wysihtml5
plonetheme.html5_hotpink
pyjx-html5
rst2html5
rst2html5-tools
rst2html5slides
sphinx_html5_basic_theme
teracy-django-html5-boilerplate
tw2.util.html5shim
tw2.wysihtml5
wtforms-html5
Description
HTML parser based on the WHATWG HTML specification
Version
0.0.9
Author
James Graham
mailto:james@hoppipolla.co.uk
https://github.com/html5lib/html5lib-python
Specify version
0.0.9
Options
Install Package
Manage Repositories
Database Console
4: Run
5: Debug
6: TODO
Python Console
1: Project
Z: Structure
2: Favorites


PageSpider [~/code-local/PageSpider]
PageSpider
Run Page Spider
Project
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
Enter file name:
Include non-project files ⇧⌘O
url
url_utilities.py (PageSpider/utilities)
Go to File ⇧⌘O
Recent Files ⌘E
Navigation Bar ⌘↑
Drop files here to open
TODO:
Project
Current File
Scope Based
Default Changelist
Found 2 TODO items in 1 file
PageSpider (2 items in 1 file)
utilities (2 items in 1 file)
database_utils.py
(2, 7) # TODO: generate the database
(7, 7) # TODO: save words to the database
1: Project
Z: Structure
2: Favorites
SciView
Database

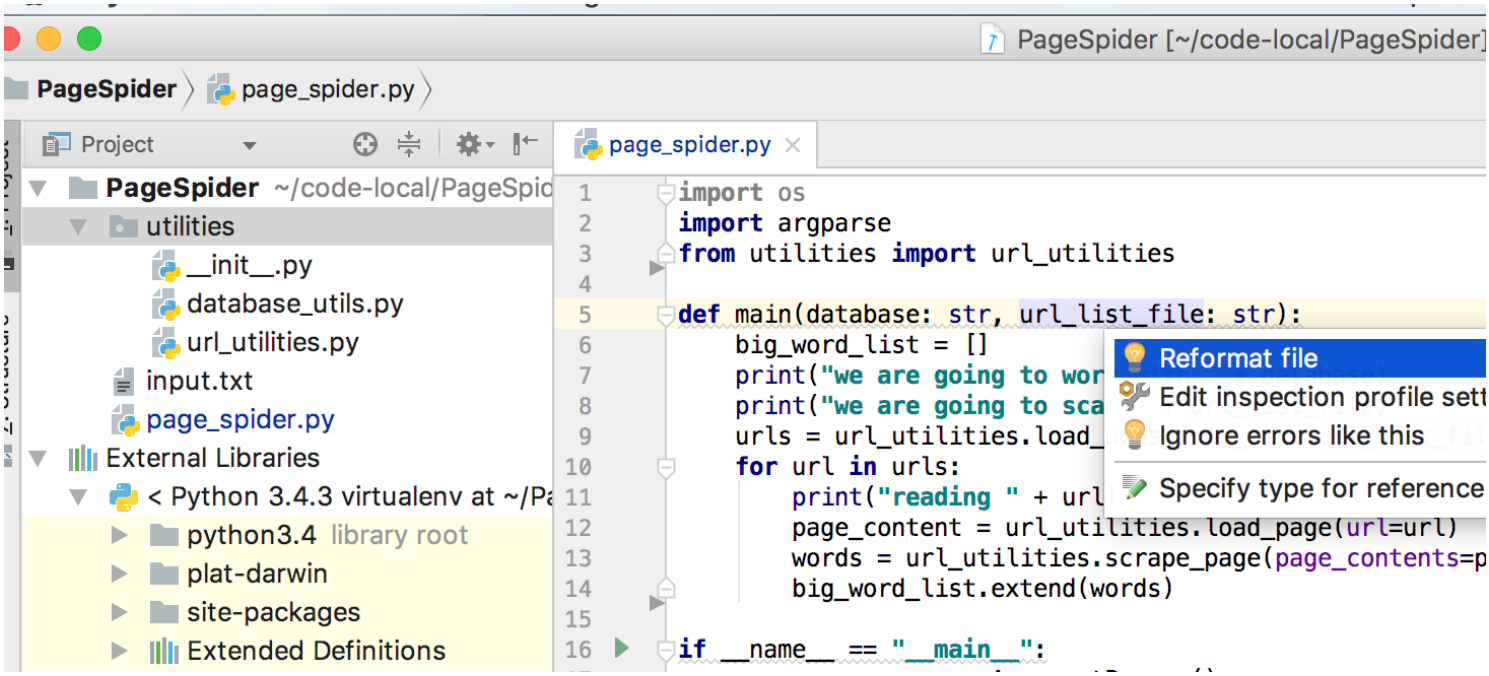

PageSpider [~/code-local/PageSpider]
PageSpider
Run Page Spider
Project
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
Search Everywhere:
Include non-project items Double
crea
Classes (⌘O)
CreateResult (com.vmware.vcenter.ovf_client)
CreateSpec (com.vmware.vapi.metadata.privilege_client)
CreateSpec (com.vmware.vcenter_client)
CreateSpec (com.vmware.vcenter_client)
CreateSpec (com.vmware.vcenter_client)
... more
Symbols (⌥⌘O)
create_database (utilities.database_utils)
Actions (⇧⌘A)
Create Changelist... Tasks & Contexts
Create Command-line Launcher... Tools
Create Gist...
Create Git Repository... Import into Version Control
Create Mercurial Repository Import into Version Control
... more
TODO:
Project
Current File
Scope Based
Default Changelist
Found 2 TODO items in 1 file
PageSpider (2 items in 1 file)
utilities (2 items in 1 file)
database_utils.py
(2, 7) # TODO: generate the database
(7, 7) # TODO: save words to the database

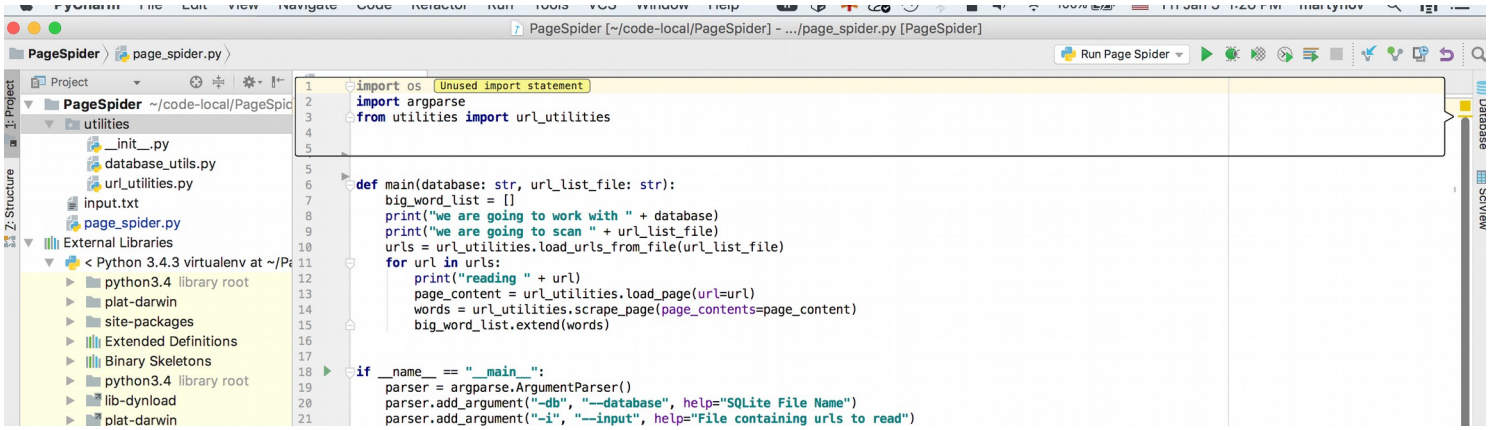

PageSpider
utilities
Project
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
database_utils.py
def create_database(database_pat
TODO: generate the databas
pass
def save_words_to_database(datab
TODO: save words to the da
pass
s.py [PageSpider]
Run Page Spider
Local History
VCS Operations Popup... ^V
Commit... ⌘K
Update Project... ⌘T
Integrate Project...
Refresh File Status
Show Local Changes as UML ⌥⇧⌘D
Git
Create Patch...
Apply Patch...
Apply Patch from Clipboard...
Shelve Changes...
Checkout from Version Control
Import into Version Control
Browse VCS Repository
Sync Settings
Import into CVS...
Create Git Repository...
Import into Subversion...
Create Mercurial Repository
Share Project on GitHub


PageSpider [~/code-local/PageSpider] - .../utilities/database_utils.py [PageSpider]
PageSpider
utilities
database_utils.py
Run Page Spider
Project
database_utils.py
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
def create_database(database_path: str):
# TODO: generate the database
pass
def save_words_to_database(database_path: str, word_list: list):
# TODO: save words to the database
pass
Clone Repository
Git Repository URL:
https://github.com/oobii/2---Introduction-to-UIKit.git
Test
Parent Directory:
/Users/martynov/code-local
Directory Name:
2---Introduction-to-UIKit
Cancel
Clone
Version Control:
Local Changes
Log
Console
Default

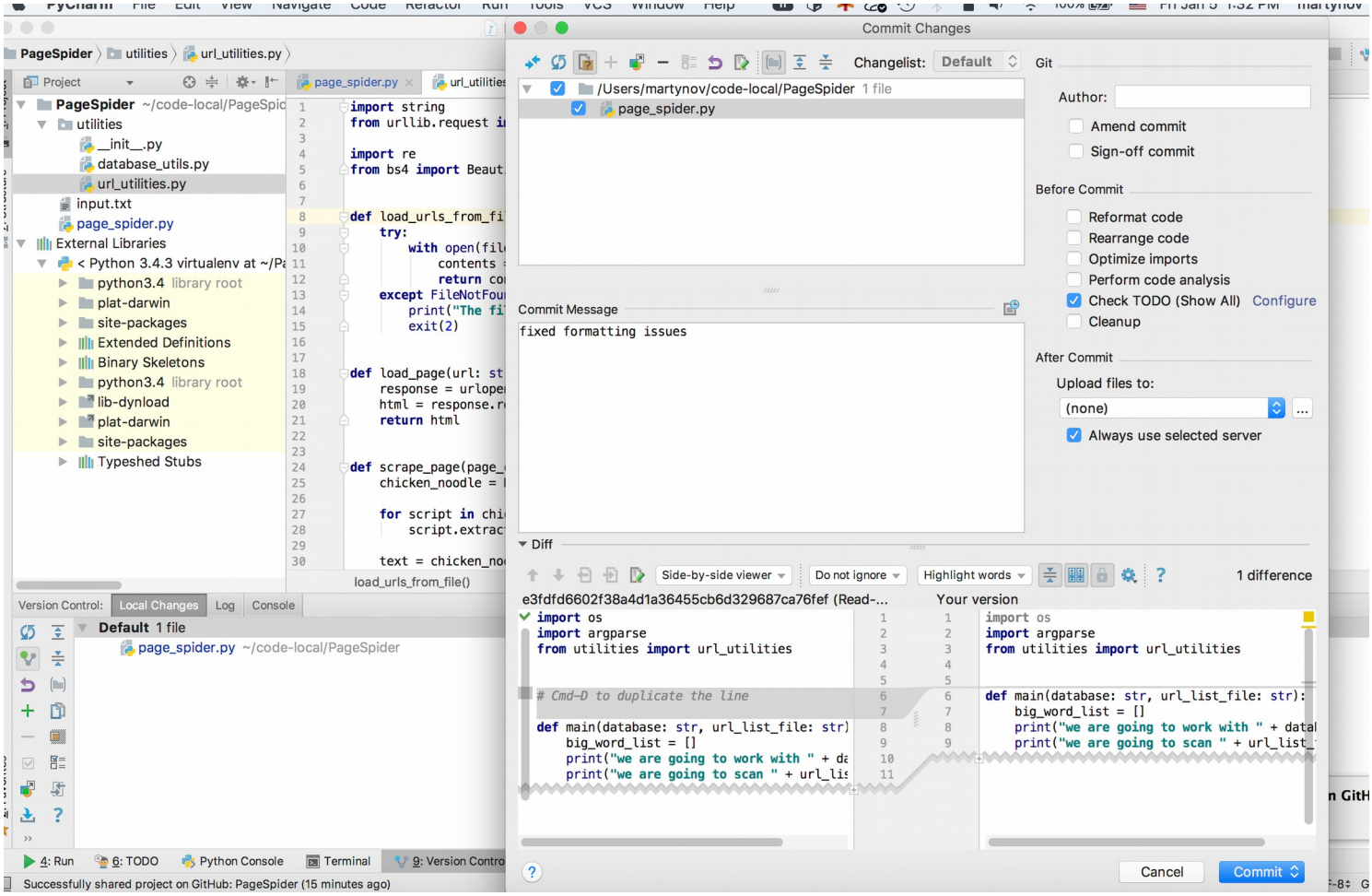
Alt-Enter on the wiggly line

Reformat Code: Alt-Cmd-L

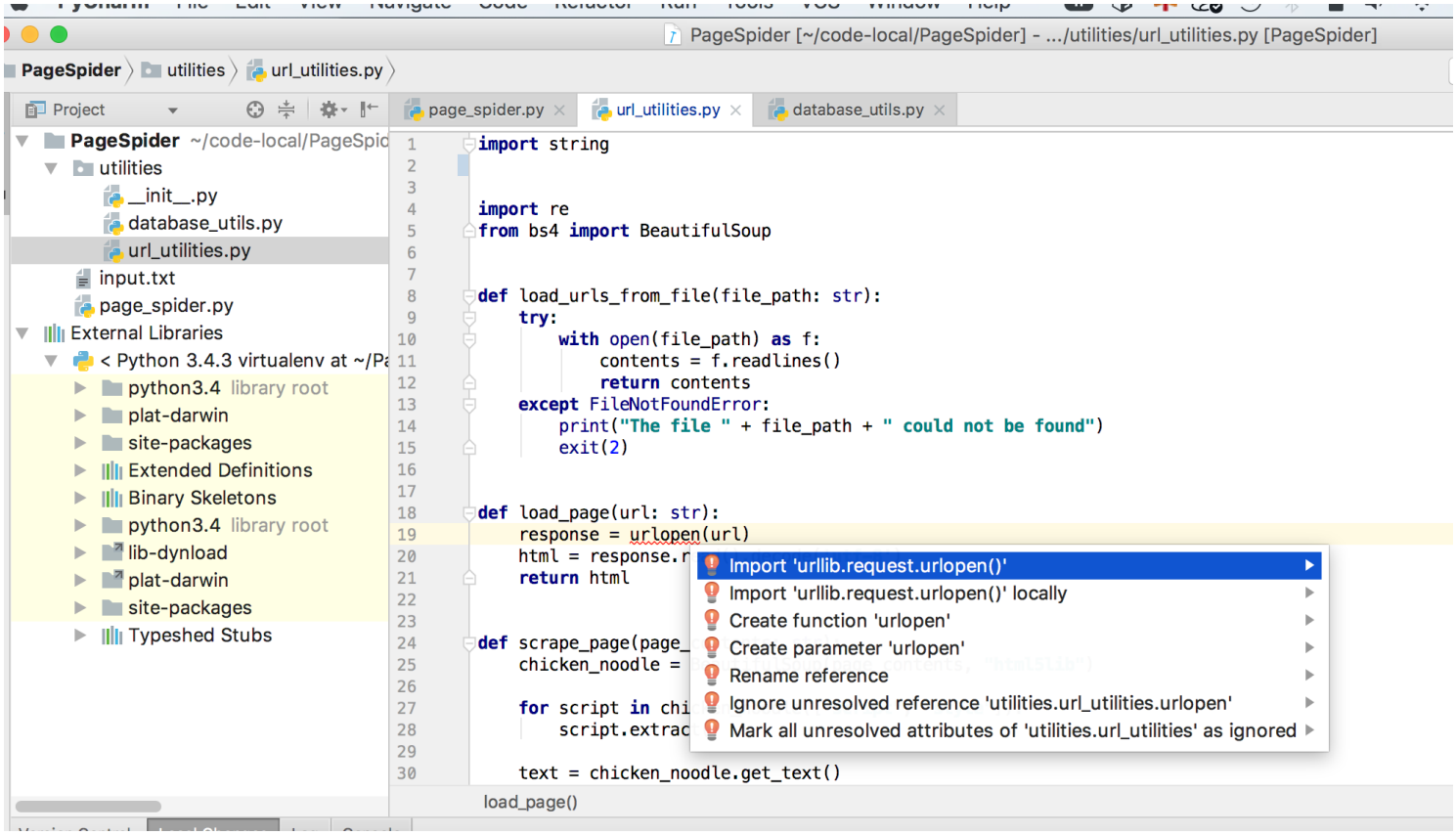


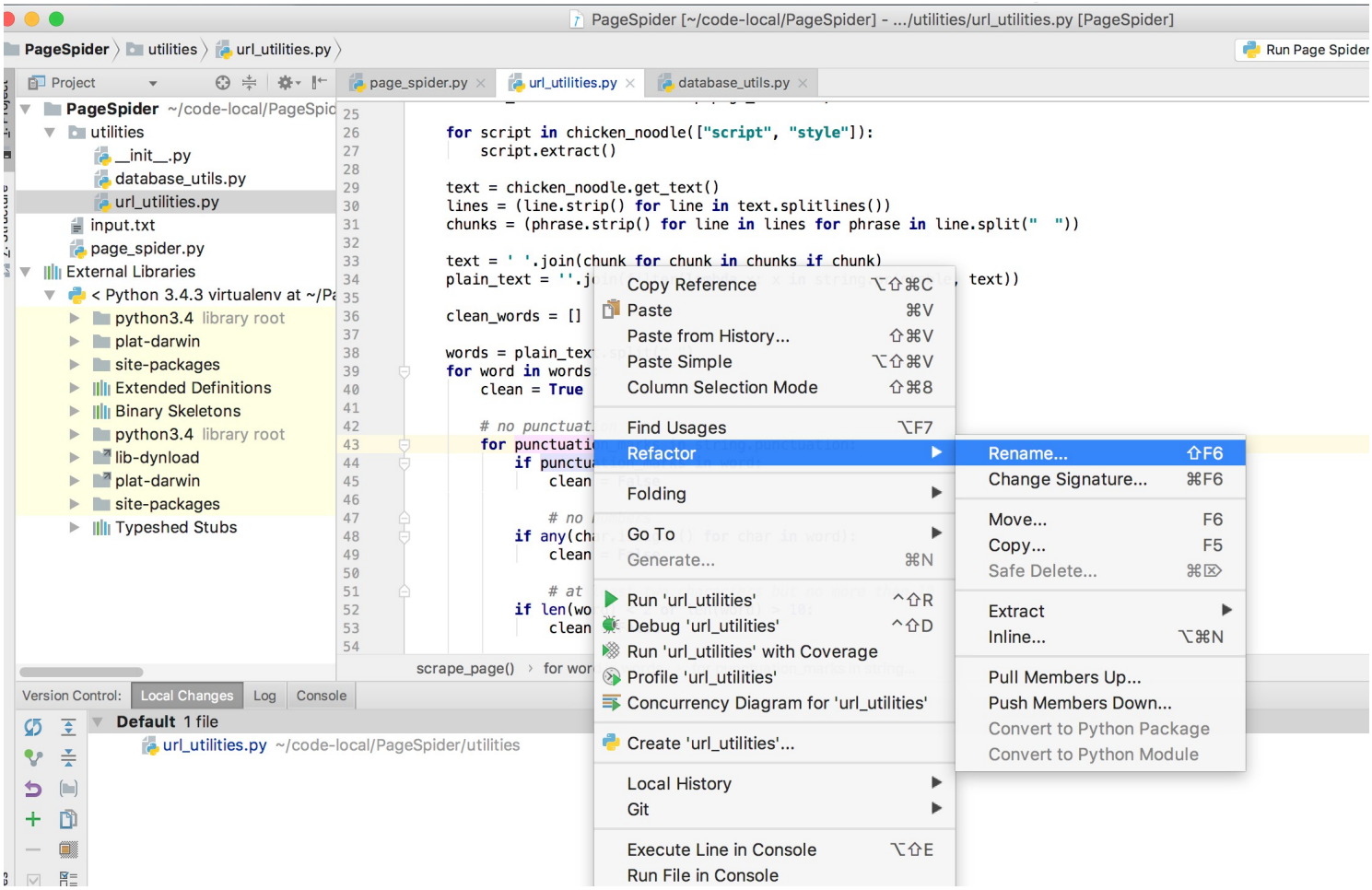

PageSpider [~/code-local/PageSpider] - .../page_spider.py [PageSpider]
PageSpider
page_spider.py
Run Page Spider
Project
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/P
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
1: Project
Z: Structure
Database
SciView
import os
Unused import statement
import argparse
from utilities import url_utilities
def main(database: str, url_list_file: str):
big_word_list = []
print("we are going to work with " + database)
print("we are going to scan " + url_list_file)
urls = url_utilities.load_urls_from_file(url_list_file)
for url in urls:
print("reading " + url)
page_content = url_utilities.load_page(url=url)
words = url_utilities.scrape_page(page_contents=page_content)
big_word_list.extend(words)
if __name__ == "__main__":
parser = argparse.ArgumentParser()
parser.add_argument("-db", "--database", help="SQLite File Name")
parser.add_argument("-i", "--input", help="File containing urls to read")

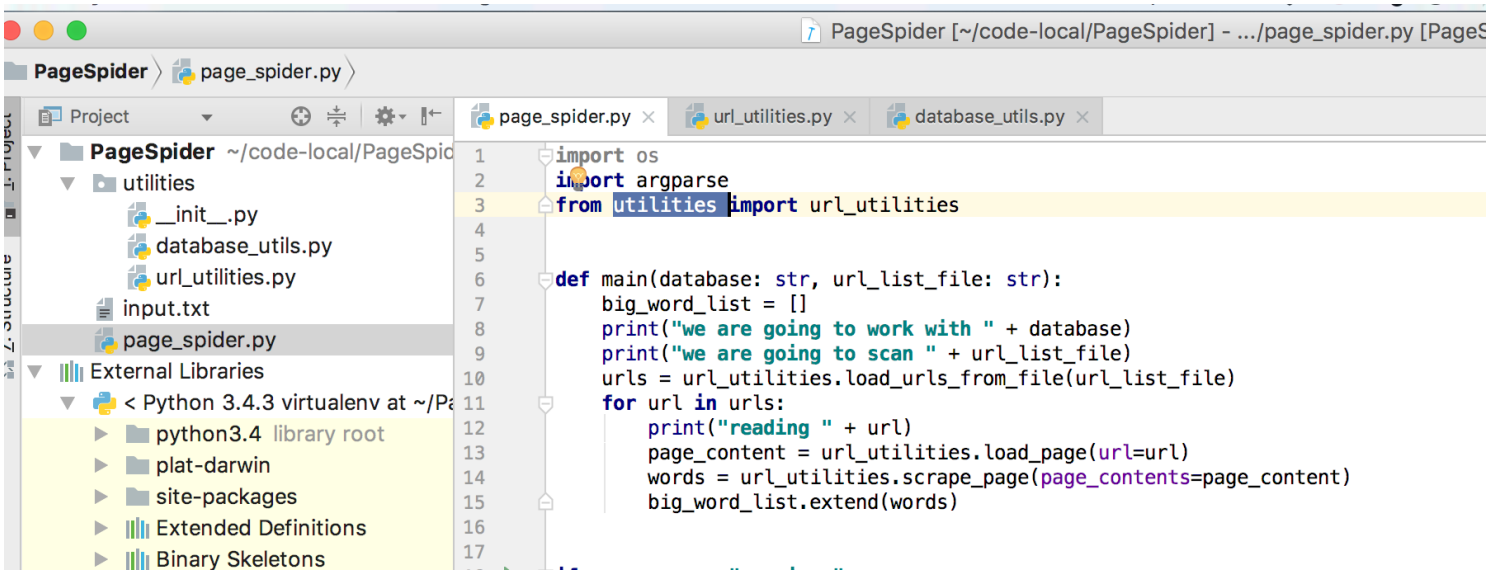
Select “Default” to Commit all at once



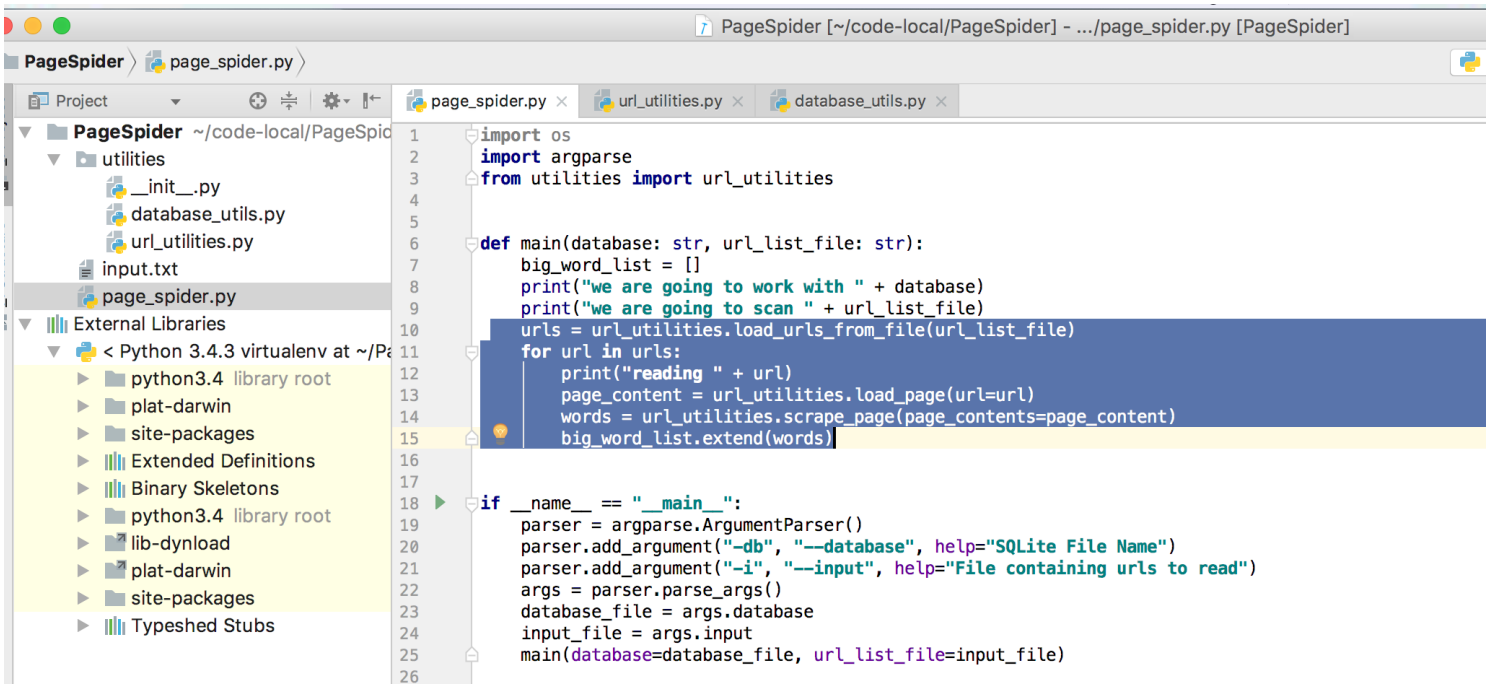
Alt-Enter on squiggly line to get suggestion how to fix the error, it will add import line at the top of the file



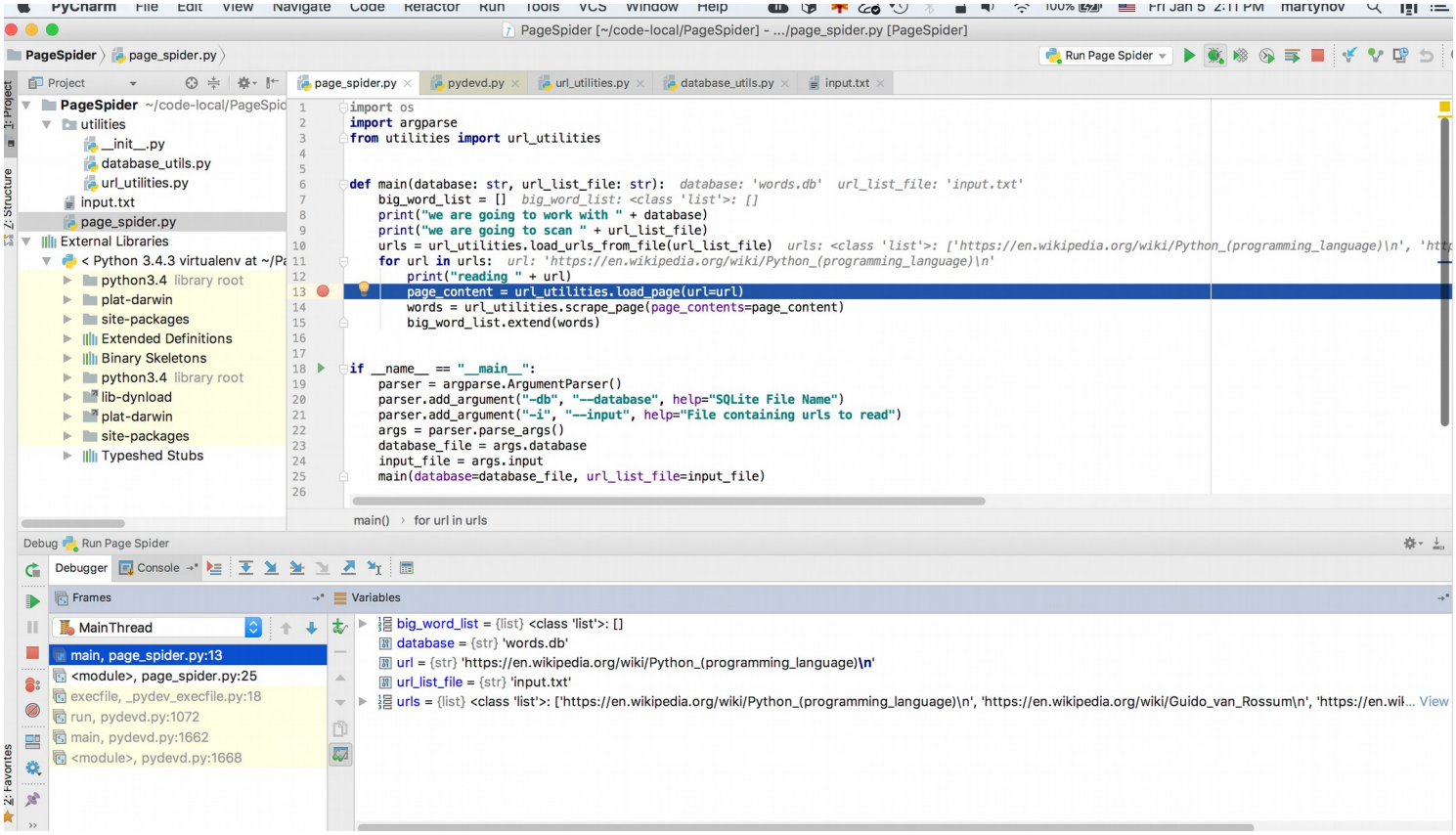

PageSpider [~/code-local/PageSpider] - .../utilities/url_utilities.py [PageSpider]
PageSpider
utilities
url_utilities.py
Run Page Spider
Project
page_spider.py
url_utilities.py
database_utils.py
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/P
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
for script in chicken_noodle(["script", "style"]):
script.extract()
text = chicken_noodle.get_text()
lines = (line.strip() for line in text.splitlines())
chunks = (phrase.strip() for line in lines for phrase in line.split(" "))
text = ' '.join(chunk for chunk in chunks if chunk)
clean_words = []
for word in words
clean = True
# no punctuat
for punctuati
if punctu
clean
no
if any(cha
clean
at
if len(wo
clean
Copy Reference ⌥⇧⌘C
Paste ⌘V
Paste from History... ⇧⌘V
Paste Simple ⌥⇧⌘V
Column Selection Mode ⇧⌘8
Find Usages ⌥F7
Refactor
Folding
Go To
Generate... ⌘N
Run 'url_utilities' ^⇧R
Debug 'url_utilities' ^⇧D
Run 'url_utilities' with Coverage
Profile 'url_utilities'
Concurrency Diagram for 'url_utilities'
Create 'url_utilities'...
Local History
Git
Execute Line in Console ⌥⇧E
Run File in Console
Rename... ⇧F6
Change Signature... ⌘F6
Move... F6
Copy... F5
Safe Delete... ⌘⌦
Extract
Inline... ⌥⌘N
Pull Members Up...
Push Members Down...
Convert to Python Package
Convert to Python Module
scrape_page()
Version Control:
Local Changes
Log
Console
Default 1 file
url_utilities.py ~/code-local/PageSpider/utilities


PageSpider [~/code-local/PageSpider] - .../page_spider.py [PageS
PageSpider
page_spider.py
Project
PageSpider ~/code-local/PageSpid
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
page_spider.py
url_utilities.py
database_utils.py
import os
import argparse
from utilities import url_utilities
def main(database: str, url_list_file: str):
big_word_list = []
print("we are going to work with " + database)
print("we are going to scan " + url_list_file)
urls = url_utilities.load_urls_from_file(url_list_file)
for url in urls:
print("reading " + url)
page_content = url_utilities.load_page(url=url)
words = url_utilities.scrape_page(page_contents=page_content)
big_word_list.extend(words)

Example how to work with Python packages
See “from … import ...”
And the selected code

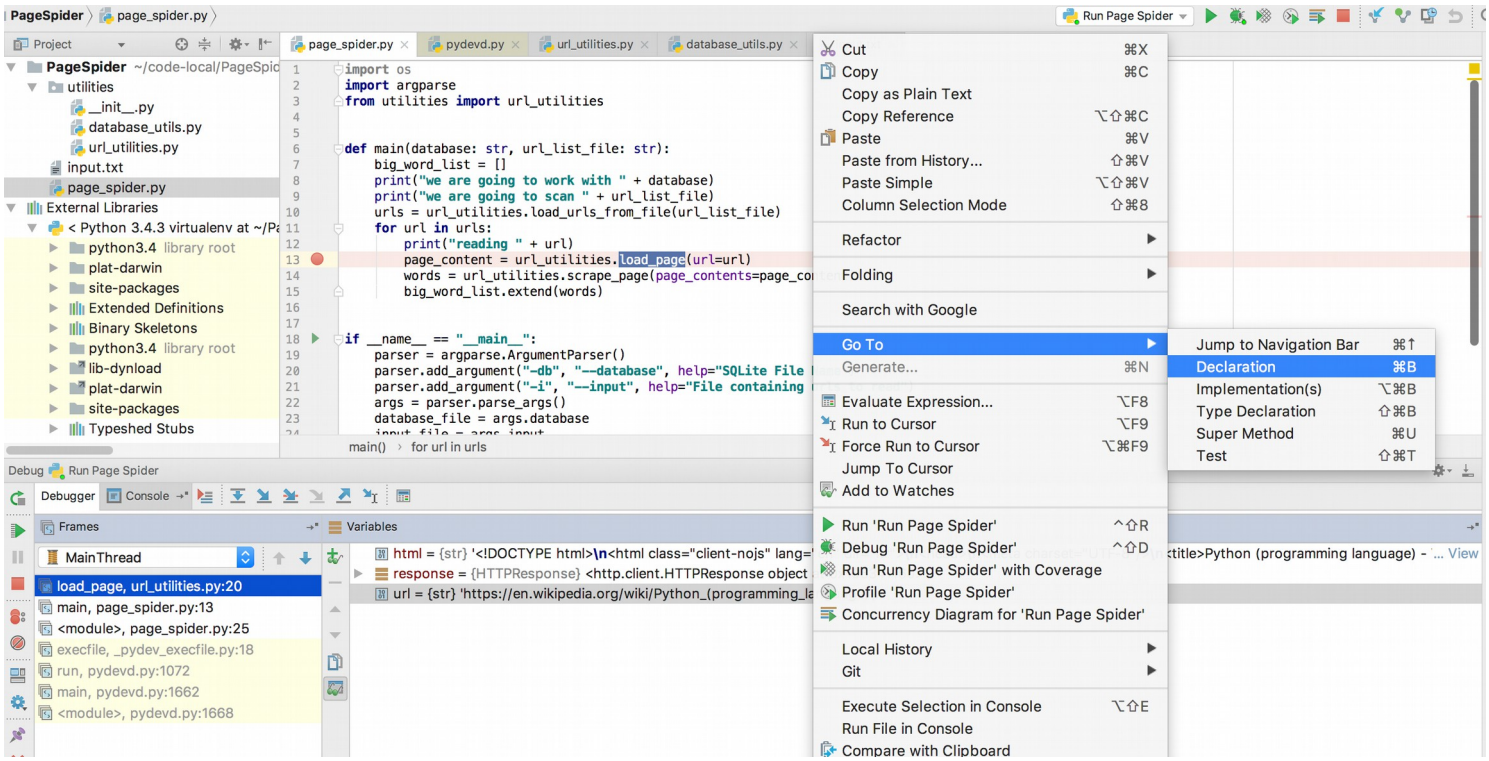


Debugging: click the bug at the top

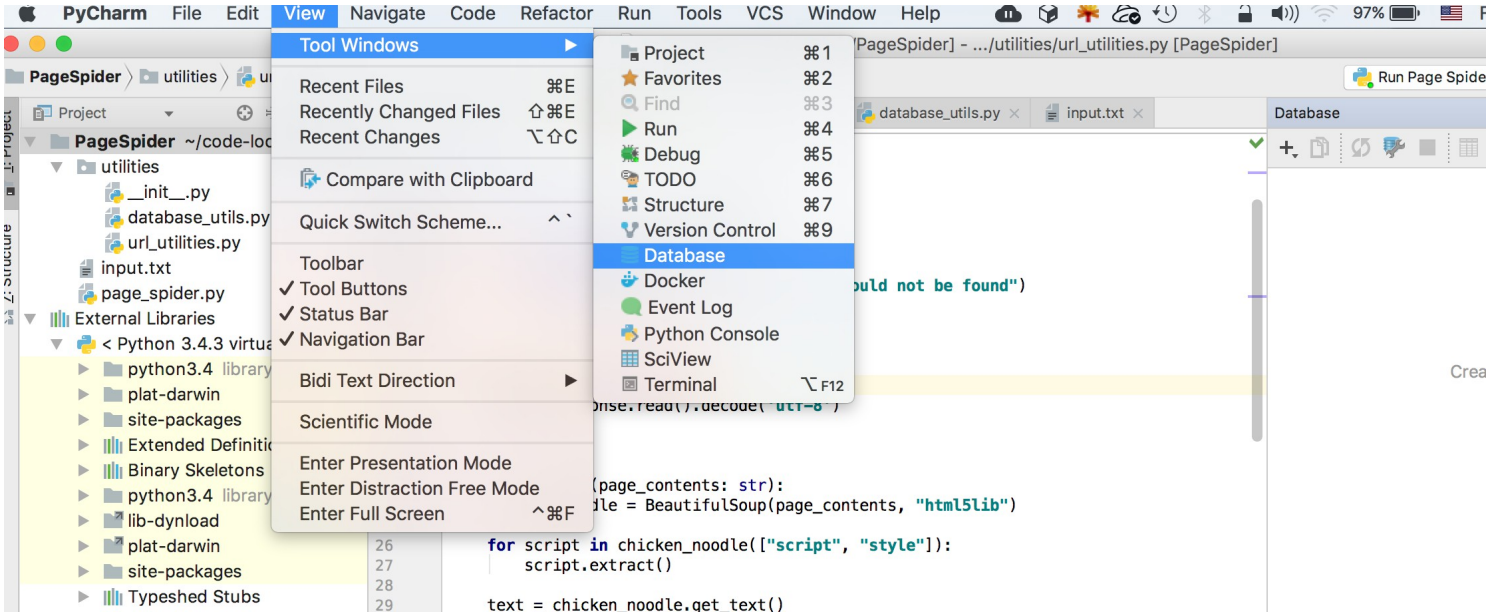


GoTo function declaration: Cmd-B

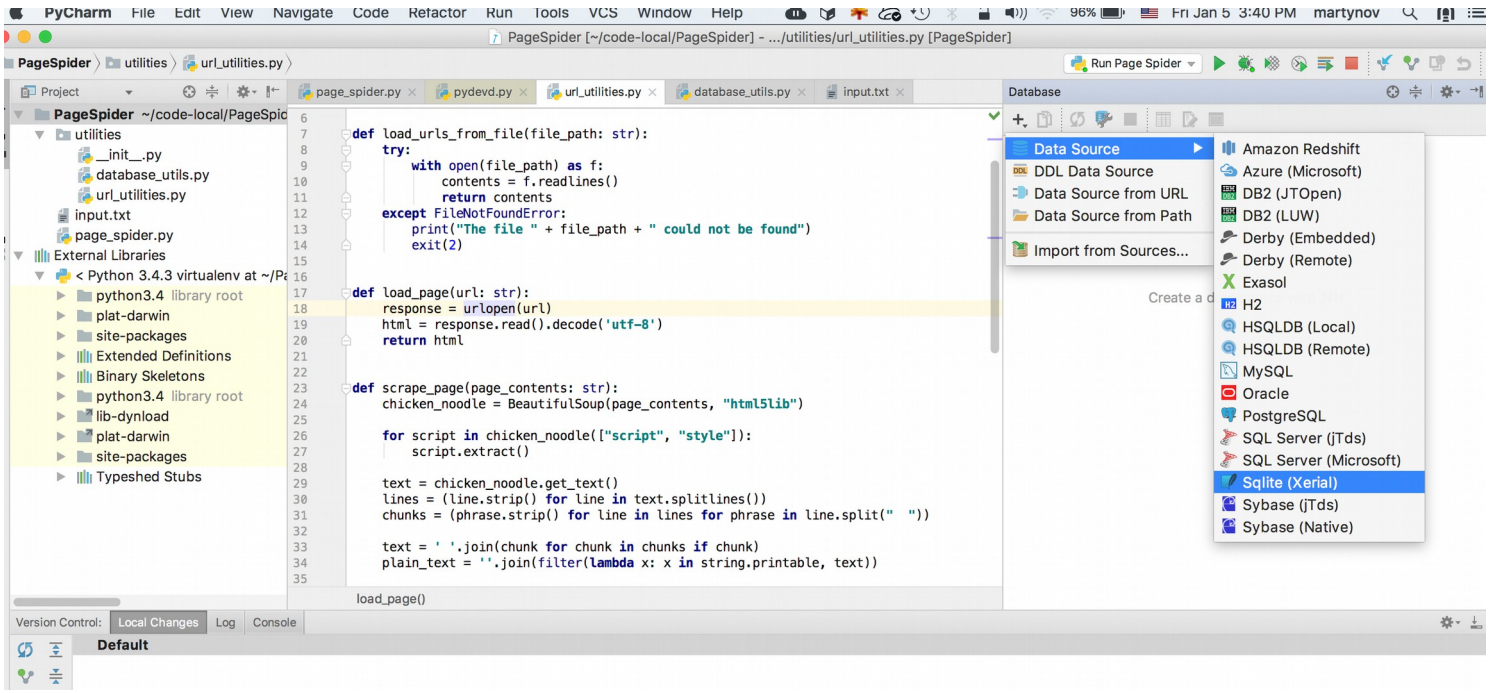
Cmd-[Back
Cmd-] Forward

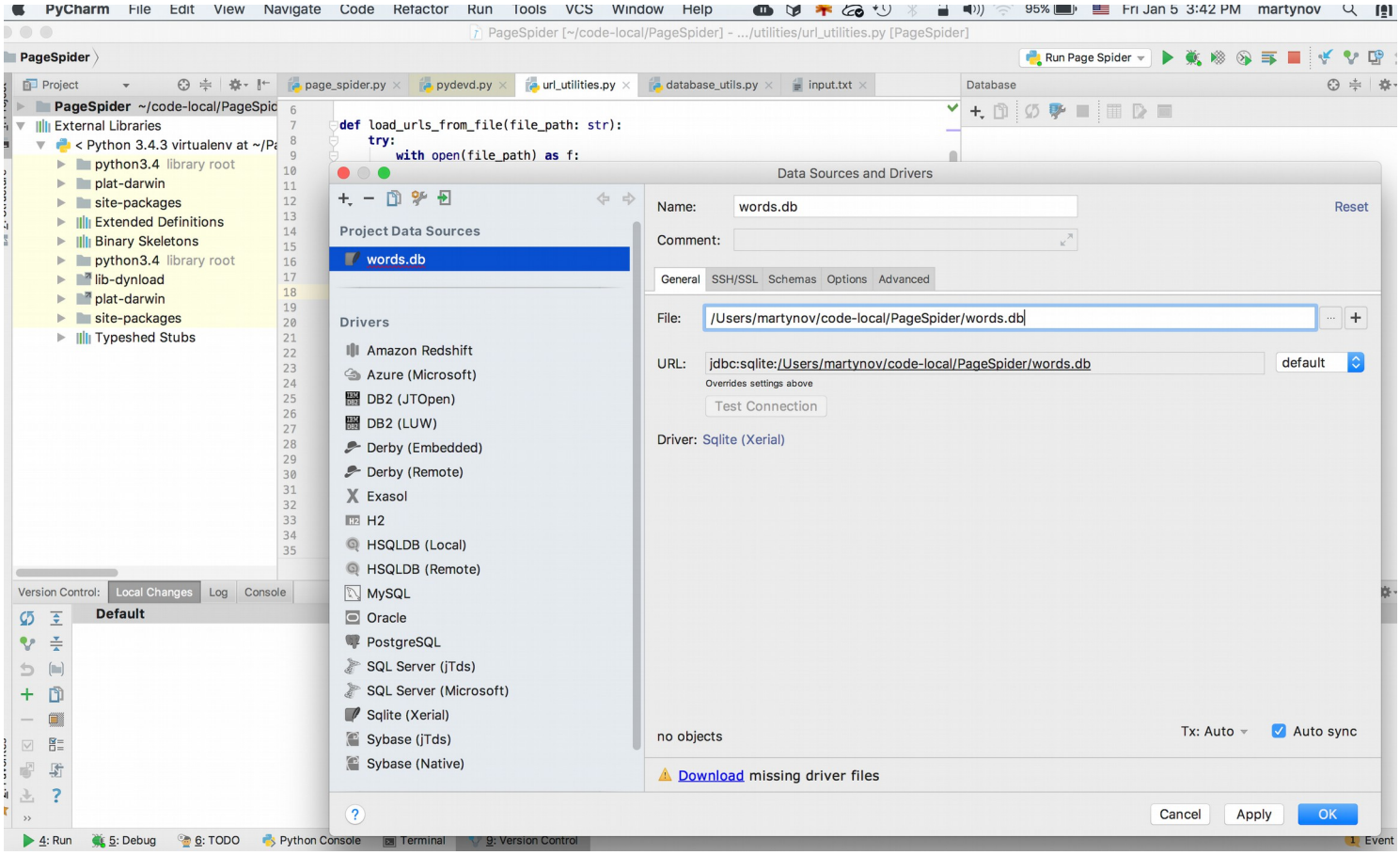


Working with Databases

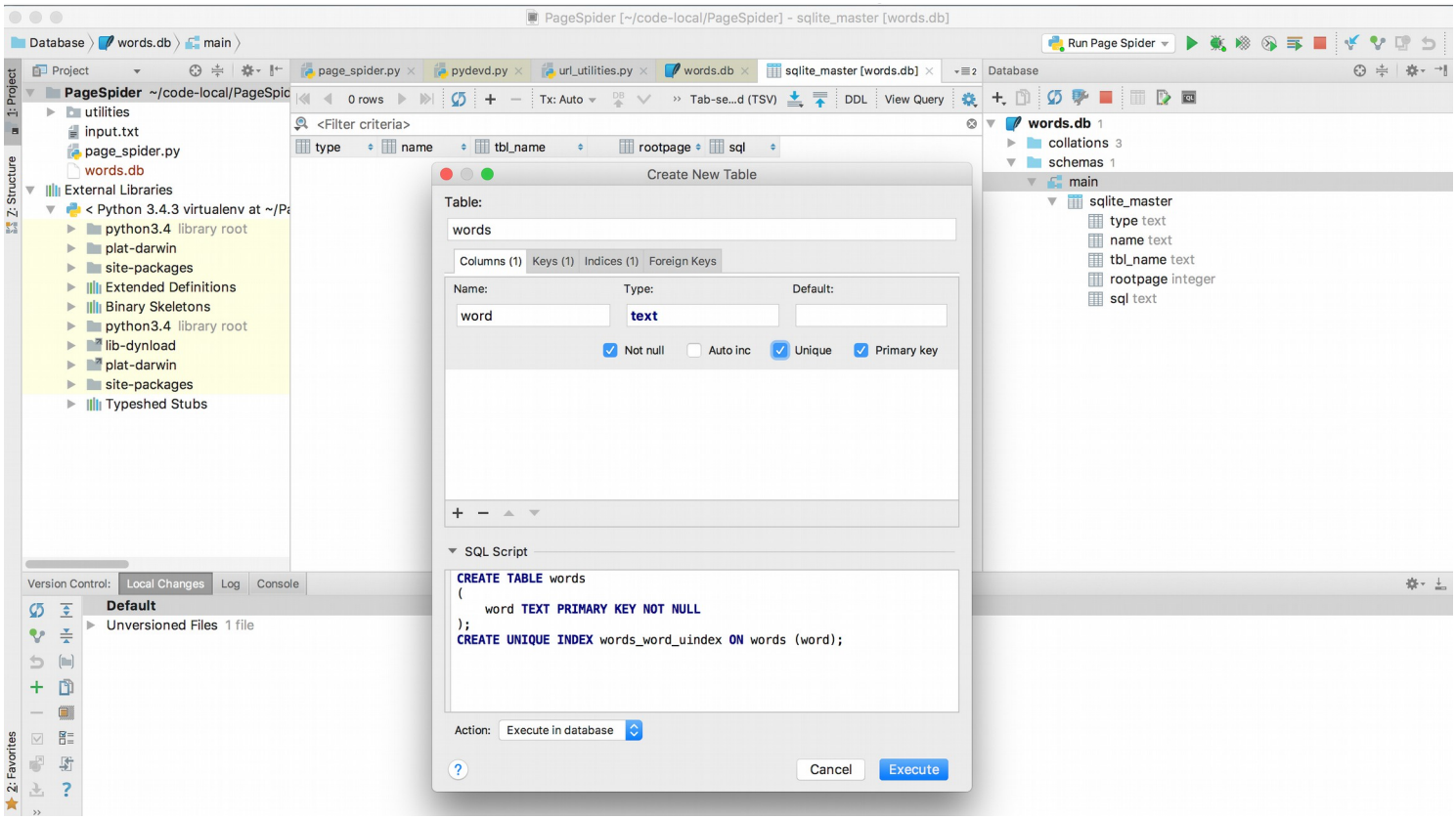


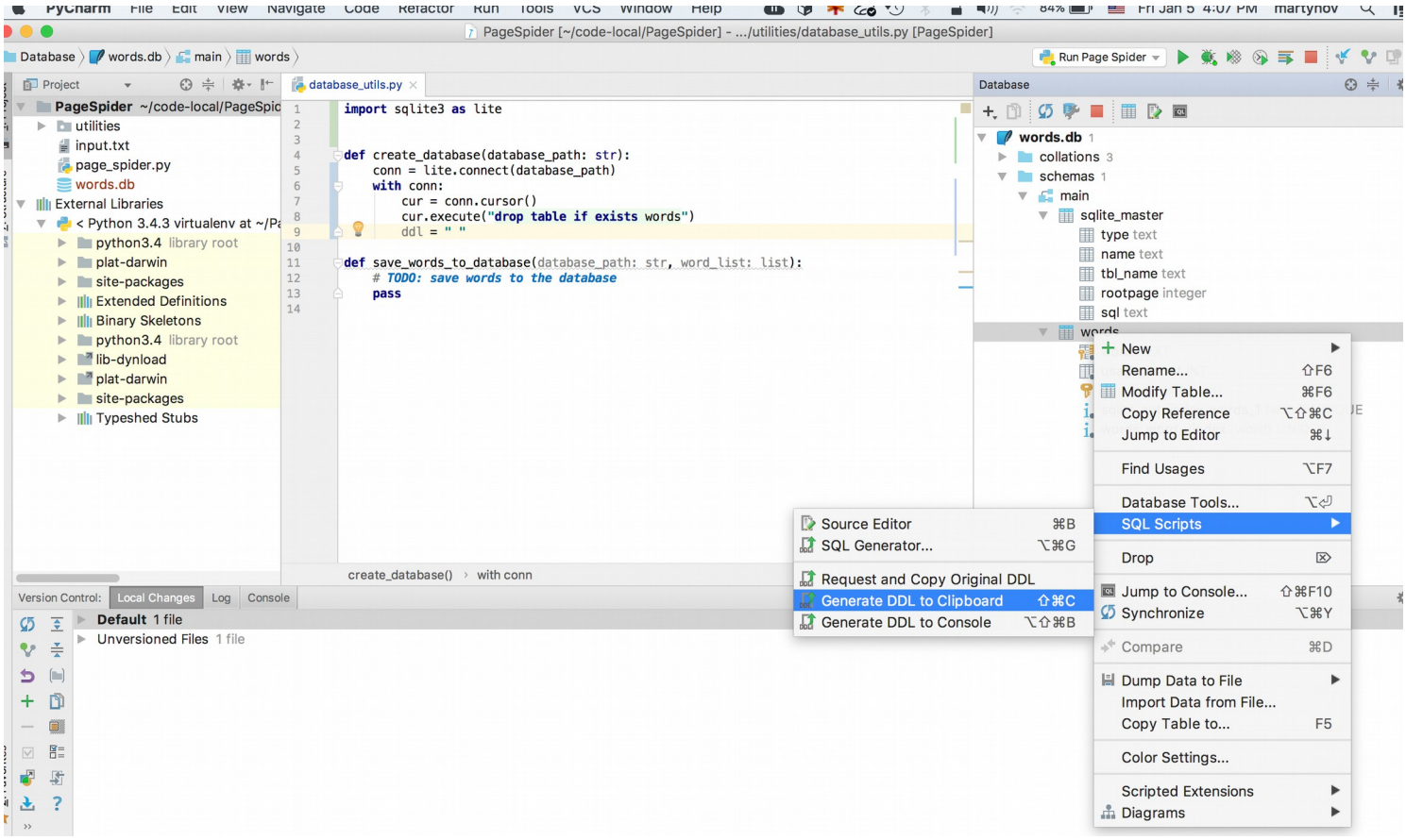
Select PageSpider, CopyPath and use it for file path
Press Download to download driver



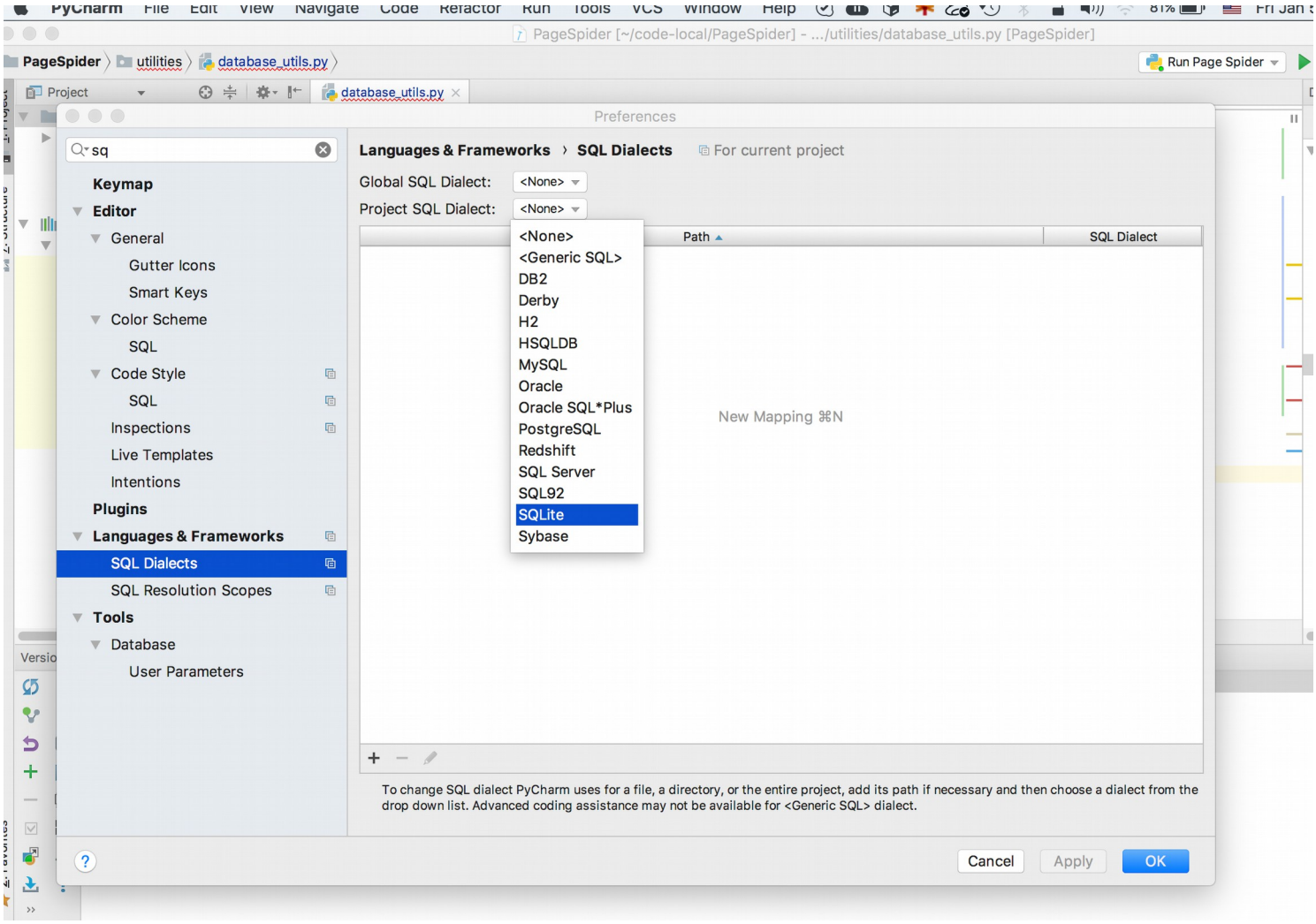

PyCharm File Edit View Navigate Code Refactor Run Tools VCS Window Help
PageSpider [~/code-local/PageSpider] - .../utilities/url_utilities.py [PageSpider]
PageSpider
Run Page Spider
Project
page_spider.py
pydevd.py
url_utilities.py
database_utils.py
input.txt
Database
PageSpider ~/code-local/PageSpic
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
def load_urls_from_file(file_path: str):
try:
with open(file_path) as f:
Data Sources and Drivers
Project Data Sources
words.db
Drivers
Amazon Redshift
Azure (Microsoft)
DB2 (JTOpen)
DB2 (LUW)
Derby (Embedded)
Derby (Remote)
Exasol
H2
HSQLDB (Local)
HSQLDB (Remote)
MySQL
Oracle
PostgreSQL
SQL Server (jTds)
SQL Server (Microsoft)
Sqlite (Xerial)
Sybase (jTds)
Sybase (Native)
Name: words.db
Reset
Comment:
General SSH/SSL Schemas Options Advanced
File: /Users/martynov/code-local/PageSpider/words.db
URL: jdbc:sqlite:/Users/martynov/code-local/PageSpider/words.db
default
Overrides settings above
Test Connection
Driver: Sqlite (Xerial)
no objects
Tx: Auto
Auto sync
Download missing driver files
Cancel
Apply
OK
Version Control:
Local Changes
Log
Console
Default
4: Run
5: Debug
6: TODO
Python Console
Terminal
9: Version Control
Event

Creating Table and columns

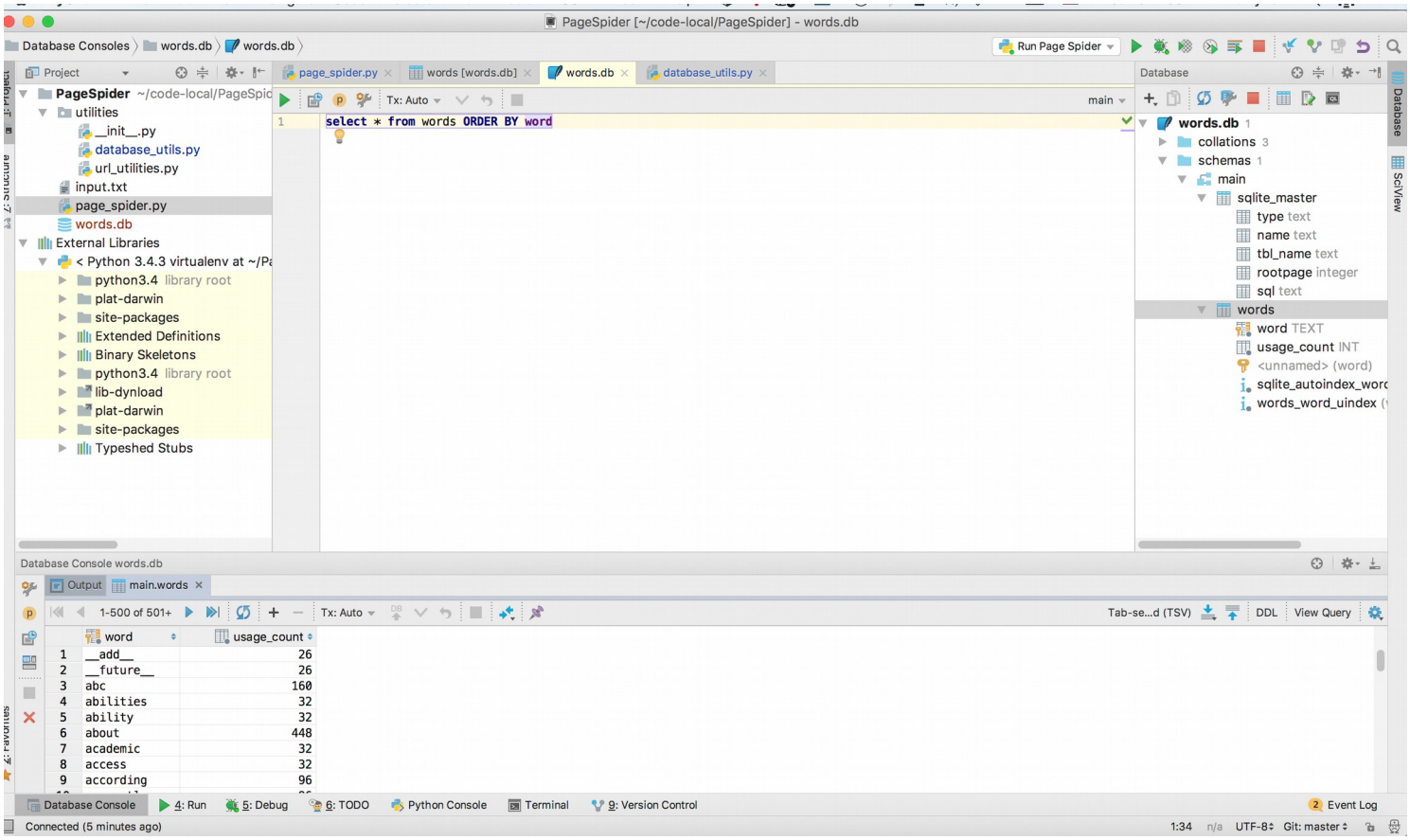



PyCharm File Edit View Navigate Code Refactor Run Tools VCS Window Help
PageSpider [~/code-local/PageSpider] - .../utilities/database_utils.py [PageSpider]
Database words.db main words
Run Page Spider
Project
PageSpider ~/code-local/PageSpic
utilities
input.txt
page_spider.py
words.db
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
database_utils.py
import sqlite3 as lite
def create_database(database_path: str):
conn = lite.connect(database_path)
with conn:
cur = conn.cursor()
cur.execute("drop table if exists words")
ddl = " "
def save_words_to_database(database_path: str, word_list: list):
# TODO: save words to the database
pass
create_database() with conn
Database
words.db 1
collations 3
schemas 1
main
sqlite_master
type text
name text
tbl_name text
rootpage integer
sql text
words
New
Rename... ⇧F6
Modify Table... ⌘F6
Copy Reference ⌥⇧⌘C
Jump to Editor ⌘↓
Find Usages ⌥F7
Database Tools... ⌥⏎
SQL Scripts
Drop ⌦
Jump to Console... ⇧⌘F10
Synchronize ⌥⌘Y
Compare ⌘D
Dump Data to File
Import Data from File...
Copy Table to... F5
Color Settings...
Scripted Extensions
Diagrams
Source Editor ⌘B
SQL Generator... ⌥⌘G
Request and Copy Original DDL
Generate DDL to Clipboard ⇧⌘C
Generate DDL to Console ⌥⇧⌘B
Version Control: Local Changes Log Console
Default 1 file
Unversioned Files 1 file

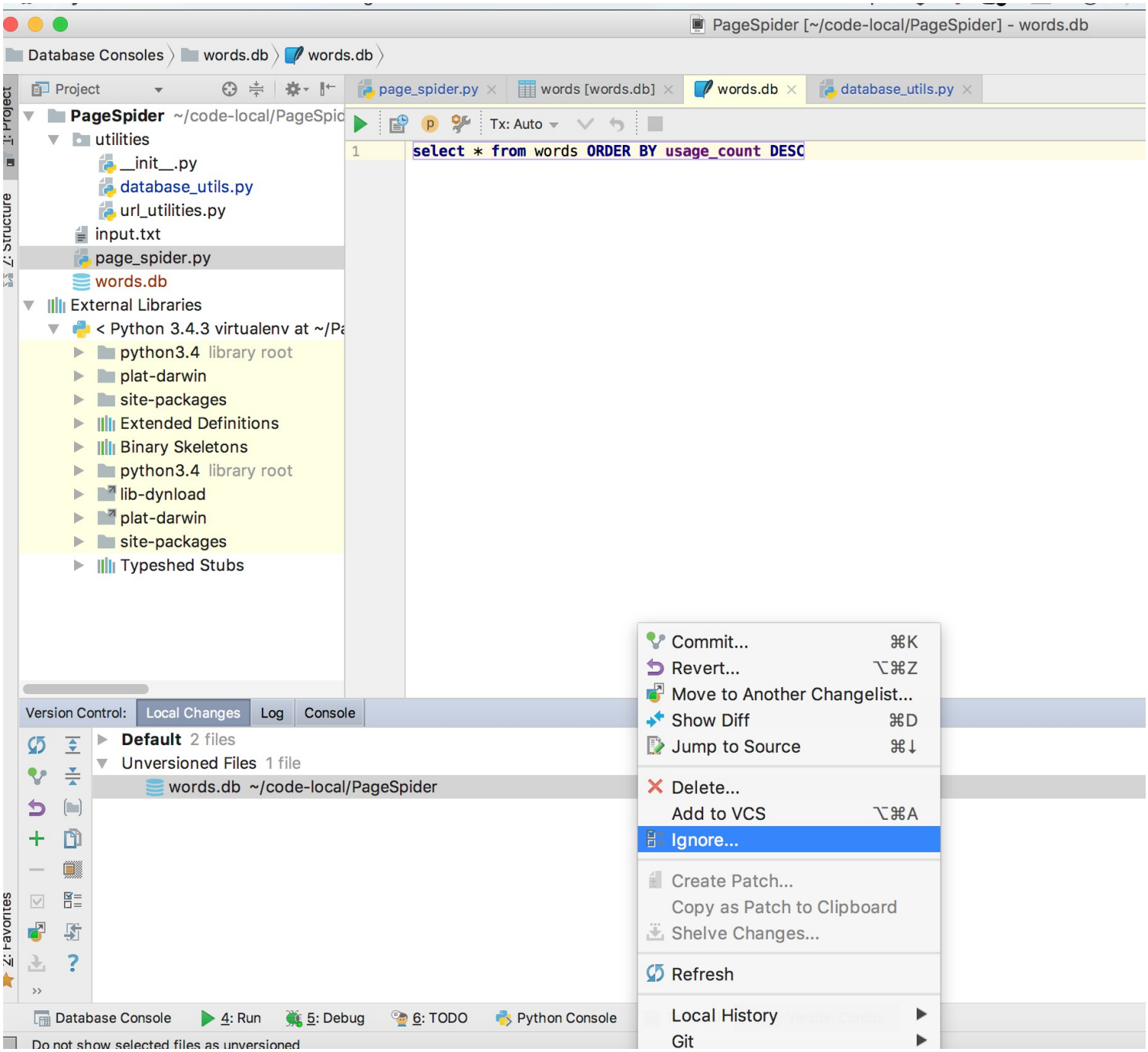
Setting Project wide SQL dialect

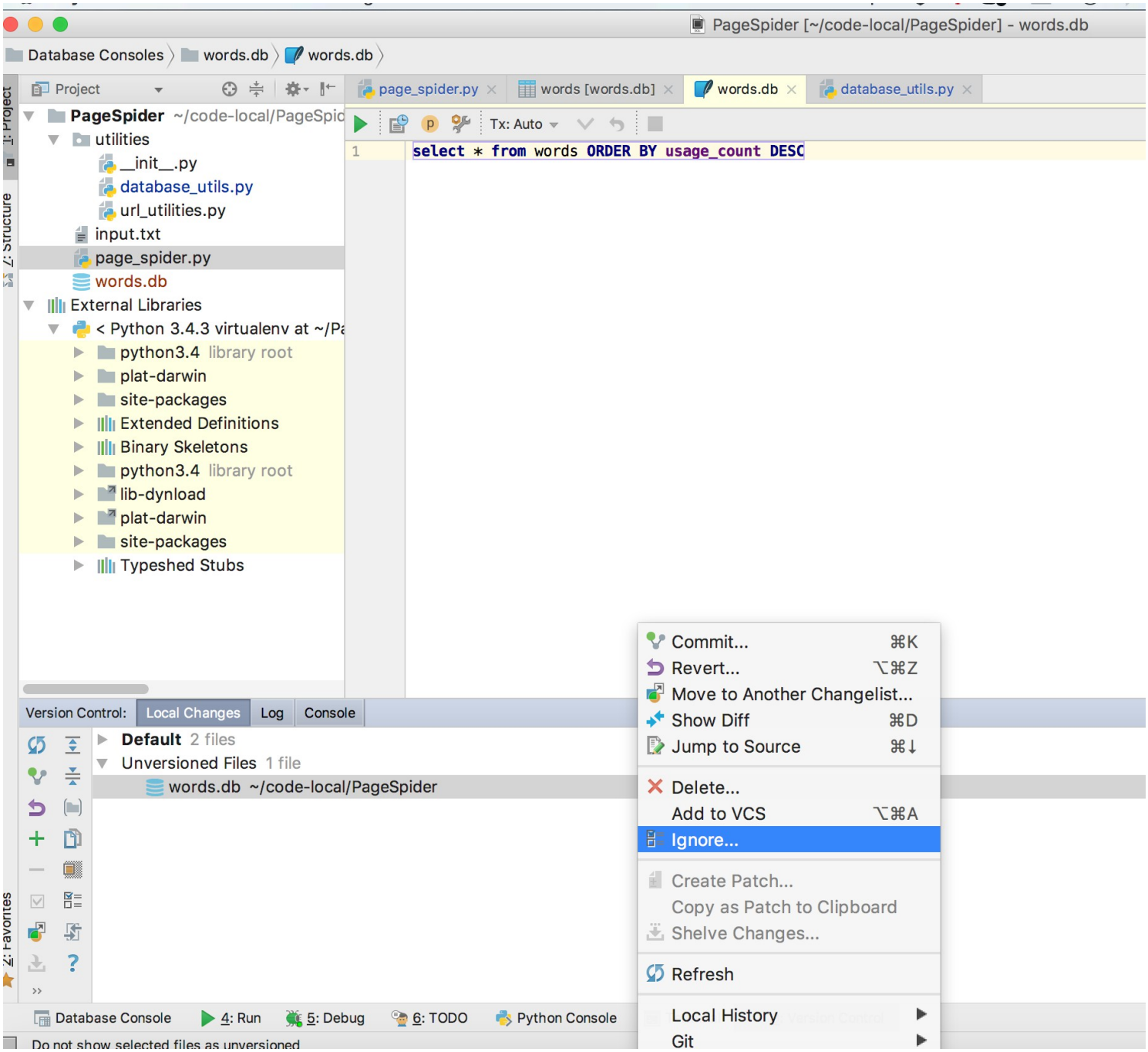


Open SQL console to run query



Ignore file in the Git version control

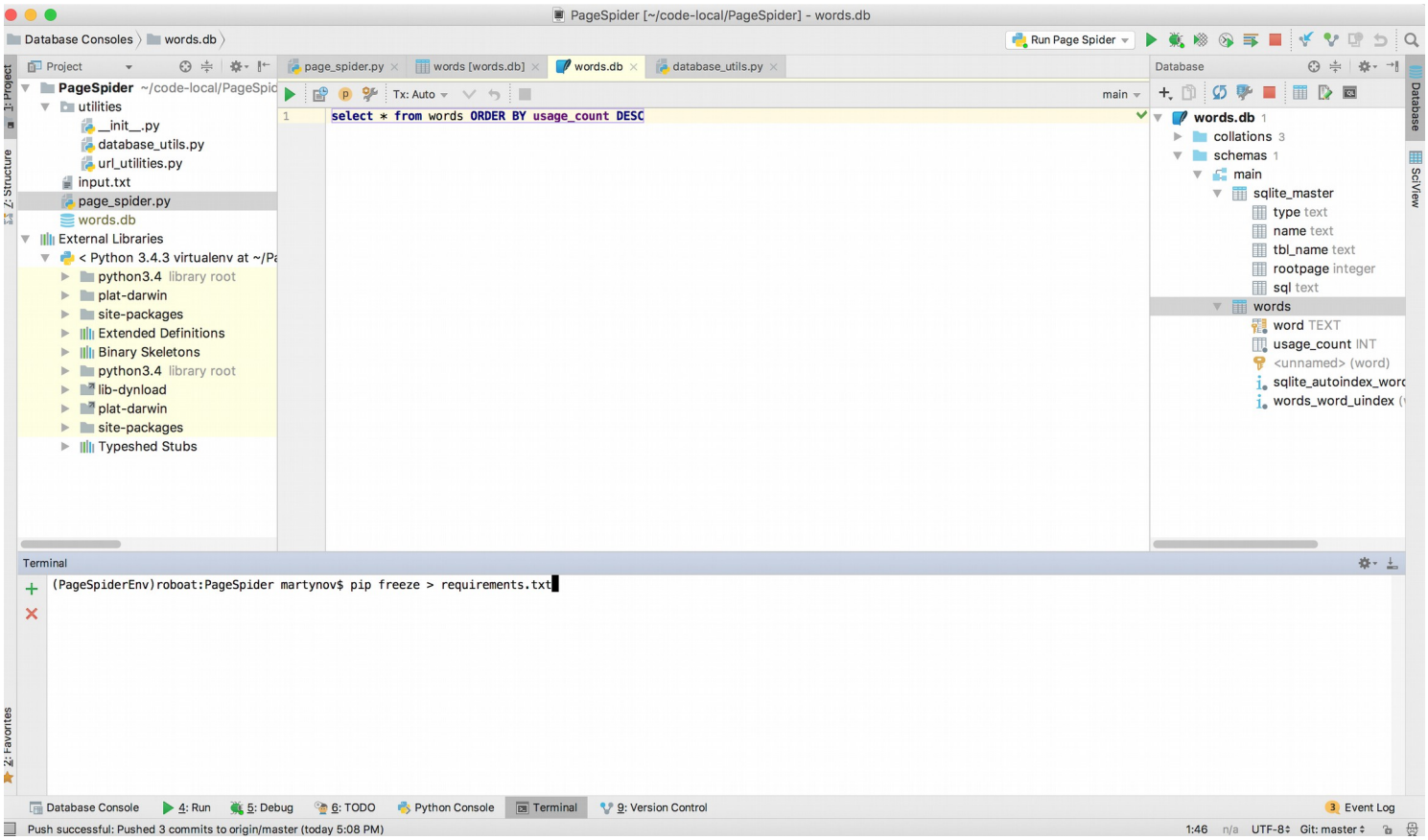


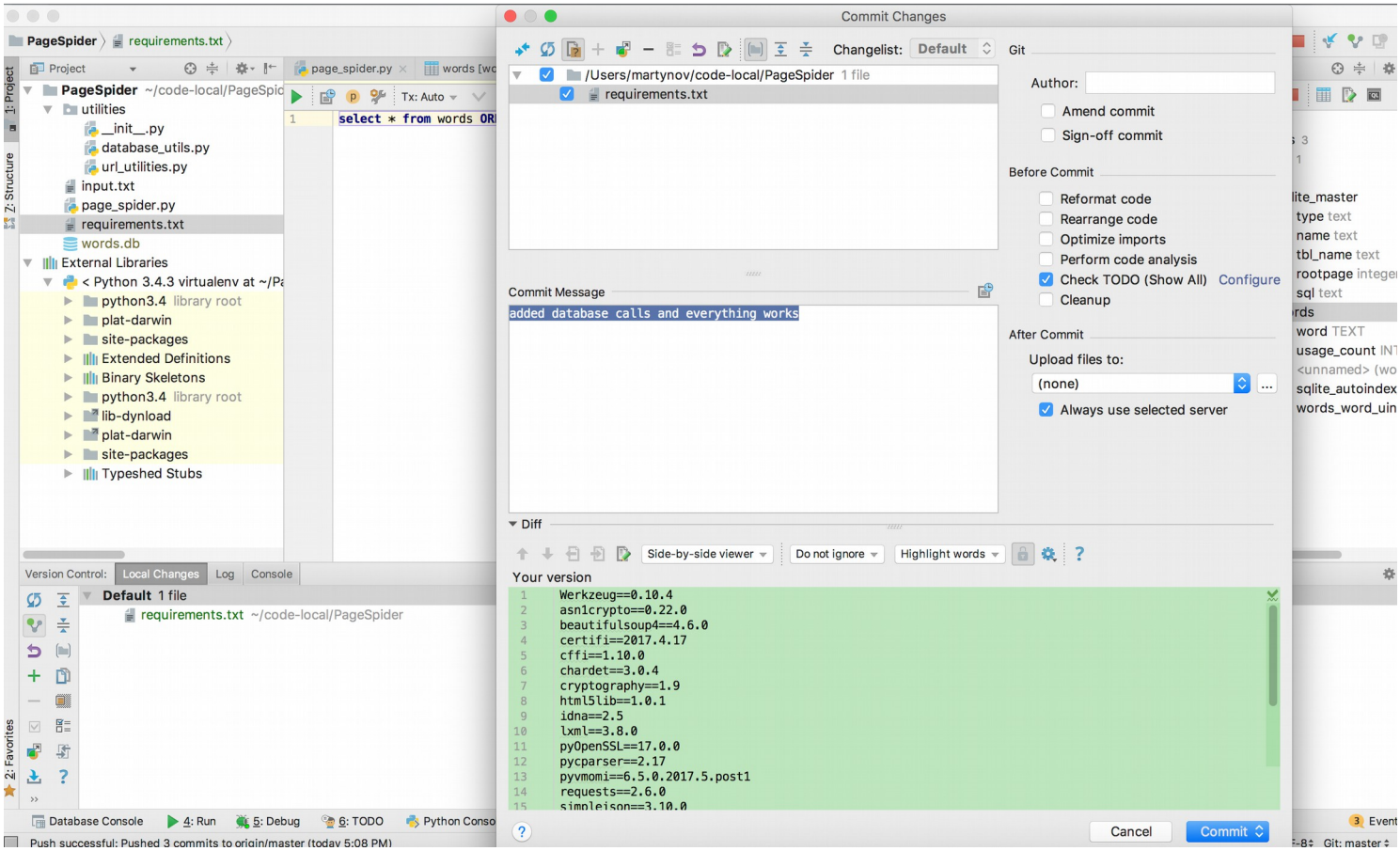

PageSpider [~/code-local/PageSpider] - words.db
Database Consoles
words.db
words.db
Project
page_spider.py
words [words.db]
words.db
database_utils.py
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
words.db
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
Tx: Auto
1
select * from words ORDER BY usage_count DESC
Version Control:
Local Changes
Log
Console
Default 2 files
Unversioned Files 1 file
words.db ~/code-local/PageSpider
Commit... ⌘K
Revert... ⌥⌘Z
Move to Another Changelist...
Show Diff ⌘D
Jump to Source ⌘↓
Delete...
Add to VCS ⌥⌘A
Ignore...
Create Patch...
Copy as Patch to Clipboard
Shelve Changes...
Refresh
Local History
Git
Database Console
4: Run
5: Debug
6: TODO
Python Console
Do not show selected files as unversioned
1: Project
7: Structure
2: Favorites

This is done through Terminal

pip freeze > requirements.txt

Then add it to Git repo and commit




Commit Changes
PageSpider
requirements.txt
Project
page_spider.py
words [wo
PageSpider ~/code-local/PageSpic
utilities
__init__.py
database_utils.py
url_utilities.py
input.txt
page_spider.py
requirements.txt
words.db
External Libraries
< Python 3.4.3 virtualenv at ~/Pa
python3.4 library root
plat-darwin
site-packages
Extended Definitions
Binary Skeletons
python3.4 library root
lib-dynload
plat-darwin
site-packages
Typeshed Stubs
Tx: Auto
select * from words OR
Changelist: Default
/Users/martynov/code-local/PageSpider 1 file
requirements.txt
Git
Author:
Amend commit
Sign-off commit
Before Commit
Reformat code
Rearrange code
Optimize imports
Perform code analysis
Check TODO (Show All) Configure
Cleanup
Commit Message
added database calls and everything works
After Commit
Upload files to:
(none)
Always use selected server
Diff
Side-by-side viewer
Do not ignore
Highlight words
Your version
Werkzeug==0.10.4
asn1crypto==0.22.0
beautifulsoup4==4.6.0
certifi==2017.4.17
cffi==1.10.0
chardet==3.0.4
cryptography==1.9
html5lib==1.0.1
idna==2.5
lxml==3.8.0
pyOpenSSL==17.0.0
pycparser==2.17
pyvmomi==6.5.0.2017.5.post1
requests==2.6.0
Version Control:
Local Changes
Log
Console
Default 1 file
requirements.txt ~/code-local/PageSpider
Database Console
4: Run
5: Debug
6: TODO
Python Conso
Push successful: Pushed 3 commits to origin/master (today 5:08 PM)
Cancel
Commit
1: Project
7: Structure
2: Favorites