Stack Overflow NLP Predictions

Vanessa Alvarado

TABLE OF CONTENTS

01

Background

Stack Overflow

Stack Overflow is an online community for developers to learn and share their programming knowledge.

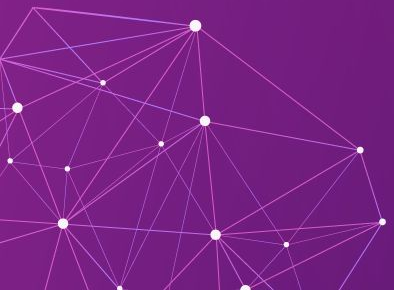
02

Problem Statement

In 2019...

2.8 Million Answers

2.6 Million **New** Questions



* Stack Overflow

Problem Statement

Questions without tags have a higher chance of not being answered.

Being able to easily suggest tags for questions will increase the likelihood of the question being answered .

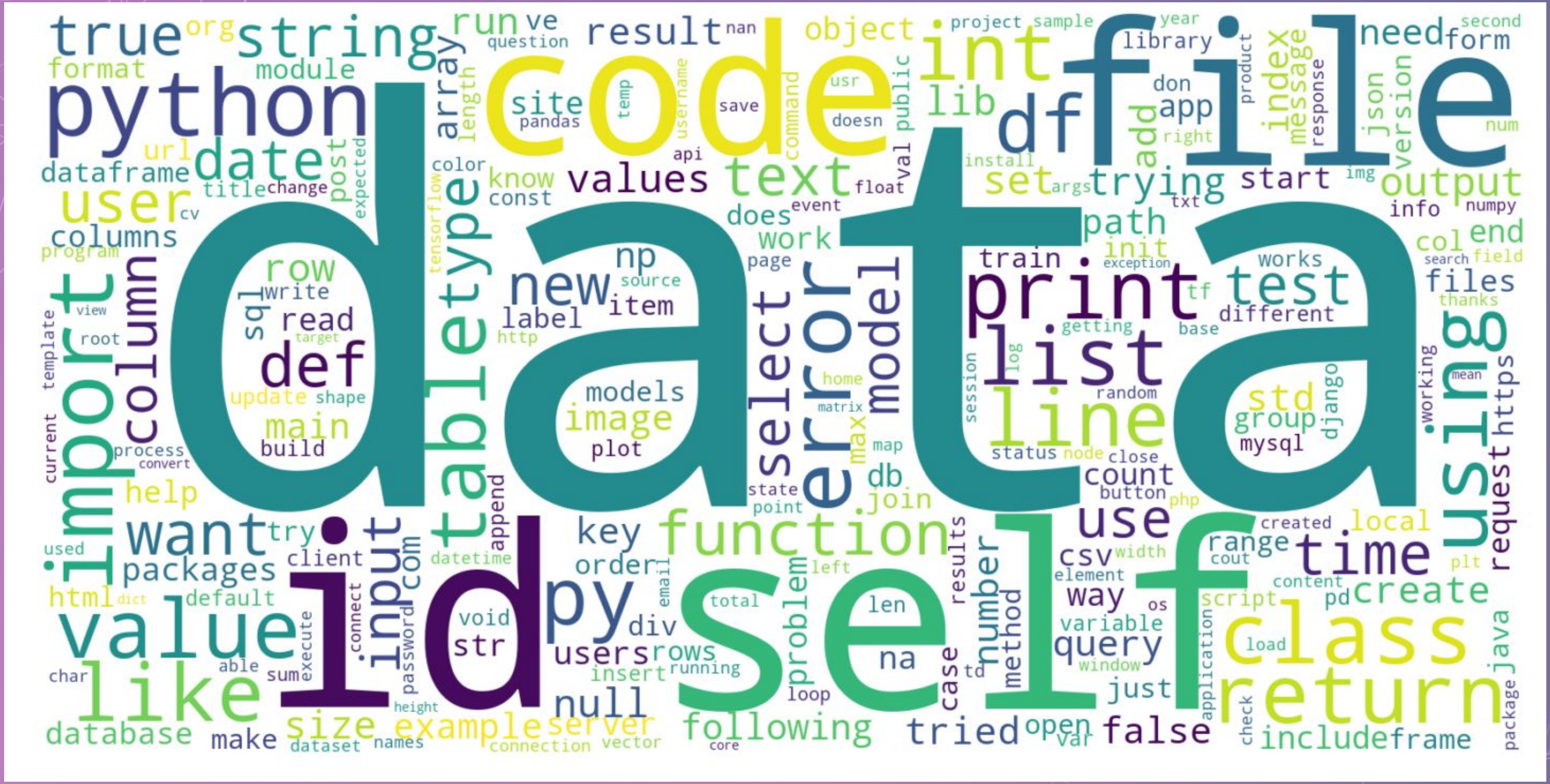
03

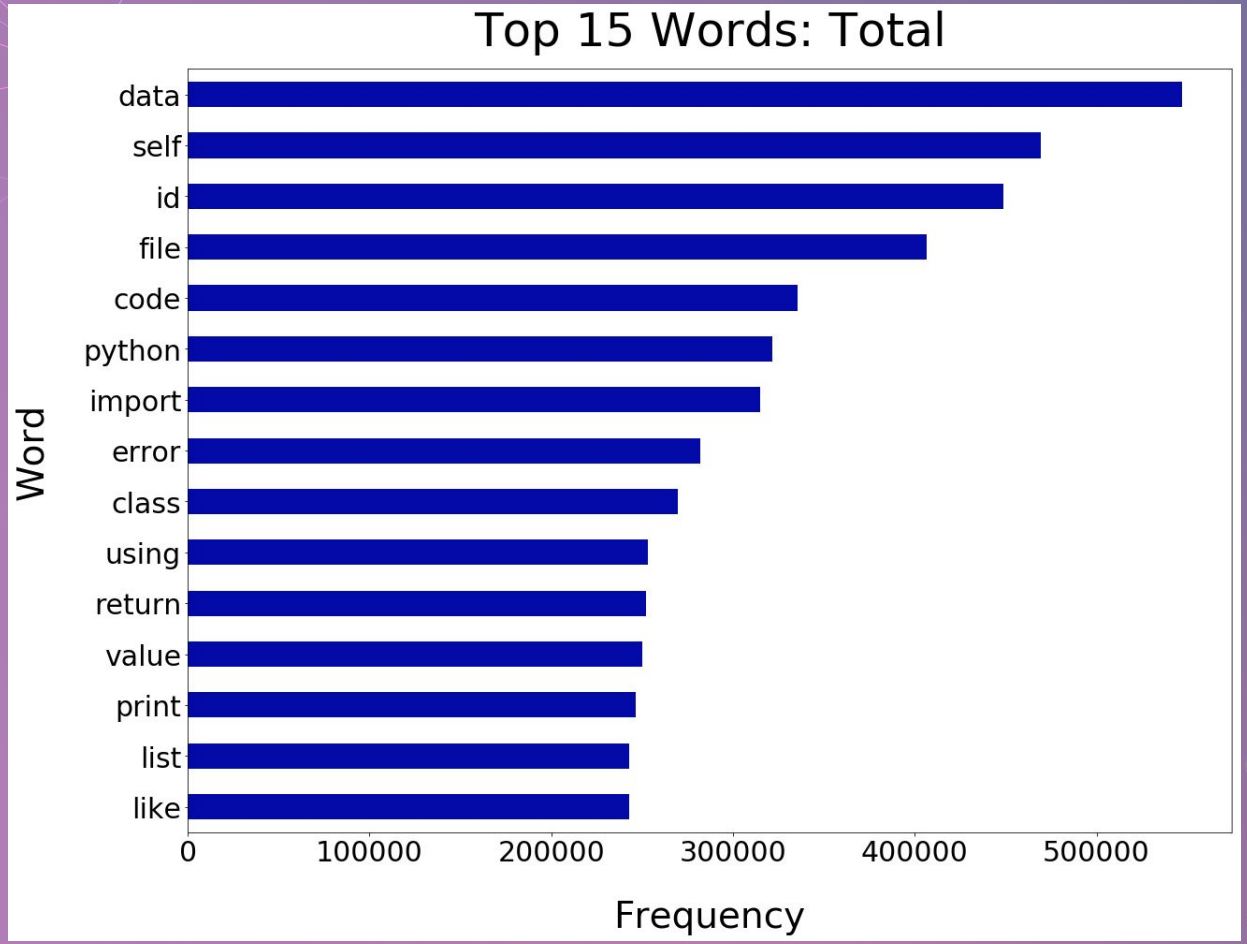
Data Overview

Queried Stackoverflow for weekly data for 2019

Target Languages:

SQL ✧ Python ✧ C++ ✧ Julia ✧ Scala ✧ R


data
self
file
code
id
python
string
true
int
df
print
error
list
line
using
table
type
import
value
like
function
return
class
text
date
user
def
py
new
want
select
model
time
use
test
output
result
values
trying
need
row
column
input
null
following
tried
false
include
frame
create
query
example
size
database
server
key
item
read
label
image
models
main
packages
dataframe
columns
array
run
object
lib
app
set
path
start
index
message
json
version
form
library
site
module
format
url
post
sql
count
csv
time
users
rows
str
na
number
method
results
case
problem
db
join
group
std
mysql
django
https
request
java
files
different
works
end
col
init
train
local
range
script
variable
just
way
open
make
help
try
client
order
div
void
html
default
plot
build
insert
len
view
root
template


Top 15 Words: Total
data
self
id
file
code
python
import
error
class
using
return
value
print
list
like
Word
0
100000
200000
300000
400000
500000
Frequency

Language Specific Top Words

SQL

1. Id
2. Table
3. Select
4. Data
5. Query

Python

1. Self
2. File
3. Python
4. Import
5. Data

Julia

1. Julia
2. Jl
3. End
4. Function
5. Array

R

1. Data
2. Na
3. Df
4. Function
5. List

C++

1. Int
2. Std
3. Include
4. Return
5. Cout

Scala

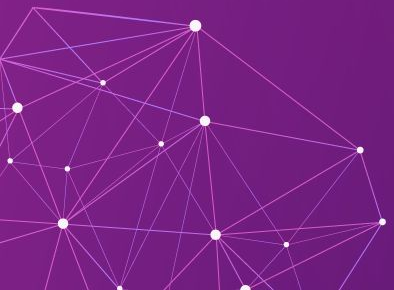
1. Scala
2. Val
3. Spark
4. Org
5. String

Top Words Used Across Languages

Code SQL, C++, Julia, R, Python

Error SQL, C++, Julia, R, Scala

Data SQL, Julia, R, Python



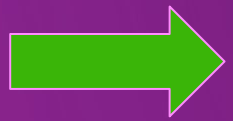
04

Modeling

Model Accuracy

Baseline: 55.23%

	Training	Testing	Diff.
Logistic + CountVect	89.96%	89.68%	0.28%
Logistic + TFIDF	91.43%	91.25%	0.18%
Decision Trees + TFIDF	82.99%	82.46%	0.53%



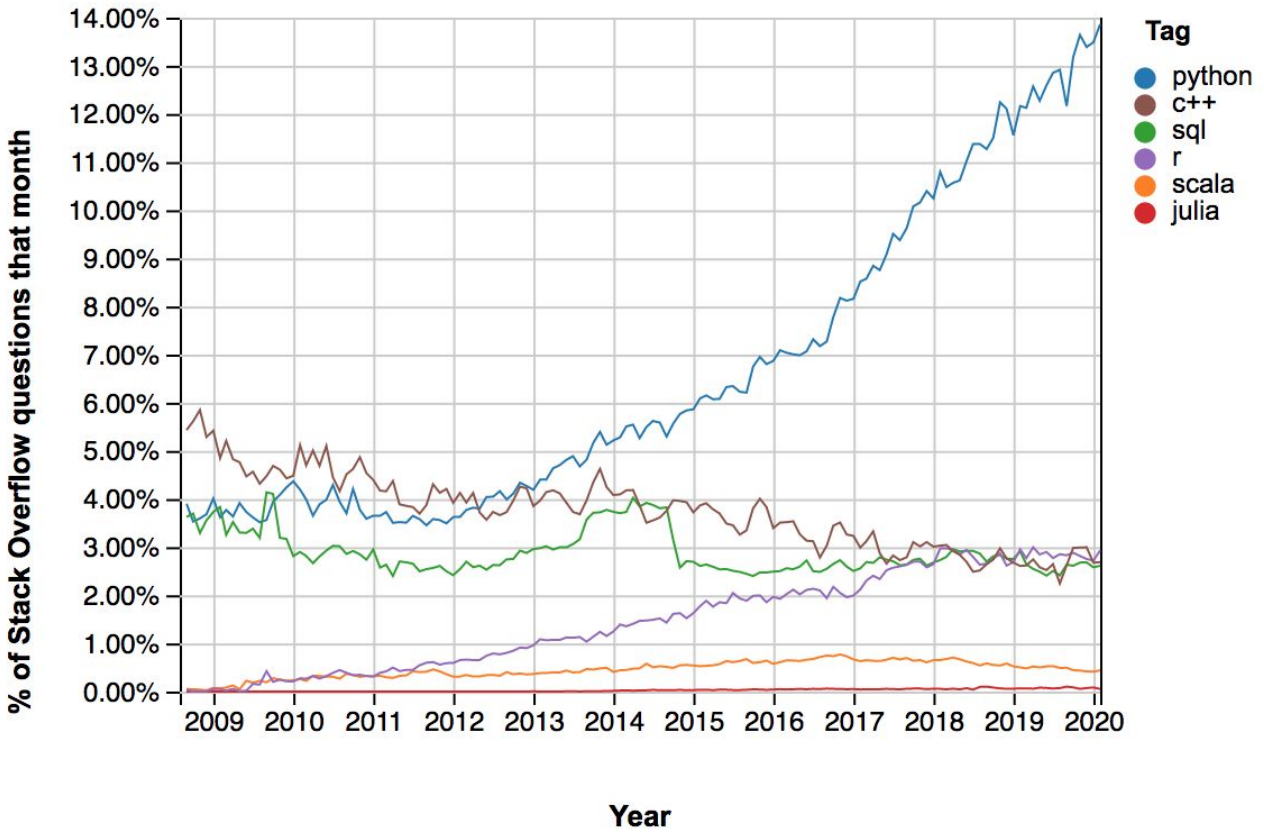
05
Conclusions

55%	Python
21%	SQL
11%	R

Question
%

10%	C++
1.9%	Scala
0.3%	Julia

Between these 6 languages, in 2019, **Python** and **SQL** accounted for more than **75%** of the tagged questions on Stack Overflow's website.


% of Stack Overflow questions that month
14.00%
13.00%
12.00%
11.00%
10.00%
9.00%
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year
Tag
python
c++
sql
r
scala
julia

06

Next Steps

Future Work

Data

Add more years to data used for modeling

More coding languages

Modeling

Hyperparameters

Random Forests

Word2Vect

Global Vectors (GloVe)

Resources

Cloud Computing to help fit and run more computationally heavy models.


“The world is one big data problem.”
– Andrew McAfee

THANKS

Does anyone have any questions?

RESOURCES

VECTORS

- Technology background with gradient colors
- Blue 5g concept background
- Abstract landing pages with technology devices

PHOTOS

- Woman using smartphone with hologram
- Happy businesswoman looking at camera with holding pencil and diary
- Portrait of smiling man holding digital tablet looking at camera
- Smiling bearded man holding disposable coffee cup while opening door
- Portrait of pretty woman holding laptop looking at camera
- Motherboard with optical fiber