Stack Overflow
Customer Behavior Analysis

Text Mining

Qianchao Liu

Outline

- Data Description
- Basic Customer Behavior Analysis
- Text Classification
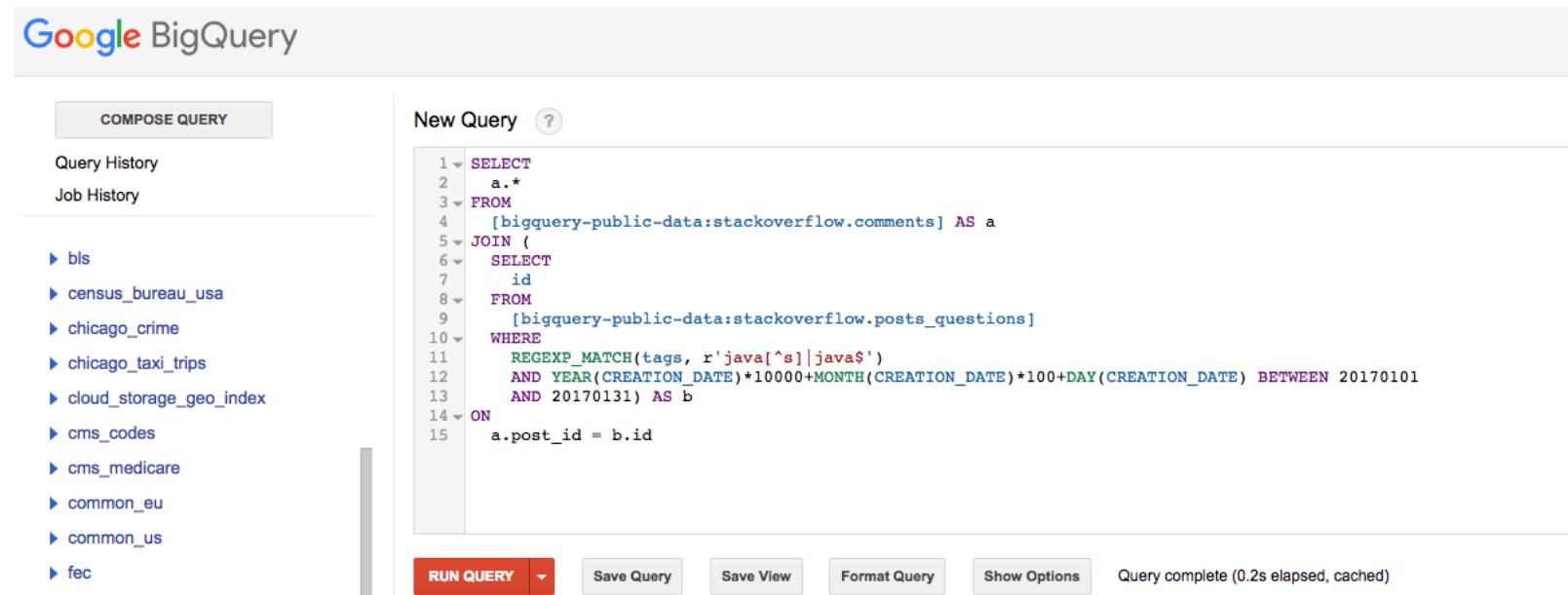- Sentiment Analysis

Part 1 Data Description

Data Description

- Questions, answers and comments from Stack Overflow
- Time period: January 2017
- Programming Languages: 7 (including Python, Ruby, Cpp ...)

Data Source:

Data Description

Part 2 Basic Customer Behavior Analysis

Basic Customer Behavior Analysis

- Question, answer and comment number over programming languages
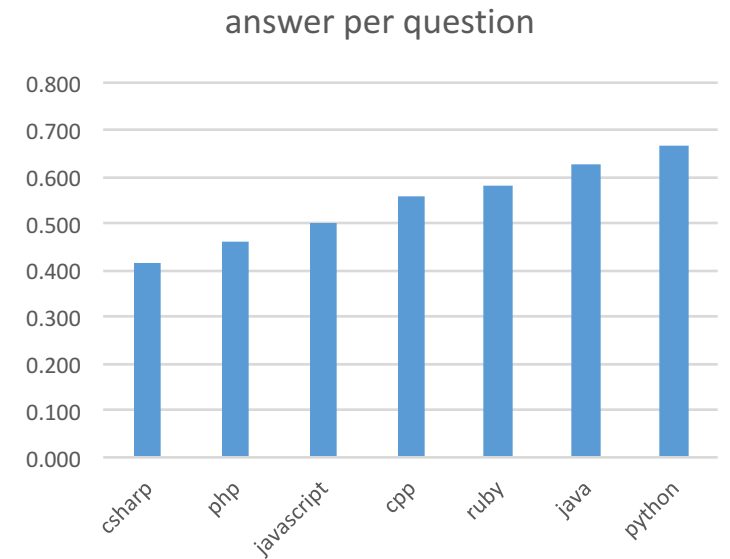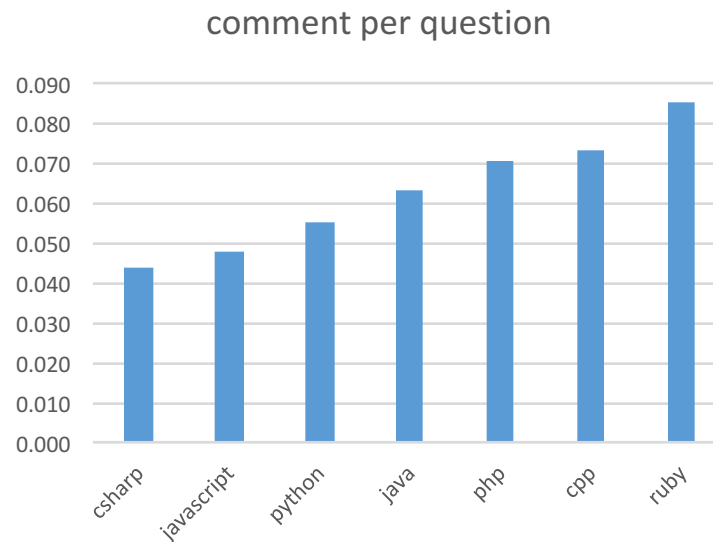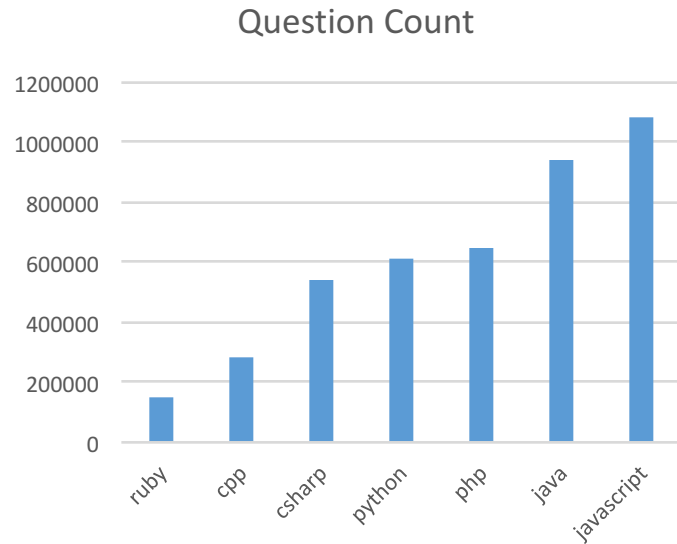
Basic Customer Behavior Analysis

- Questions post time over Programming Language
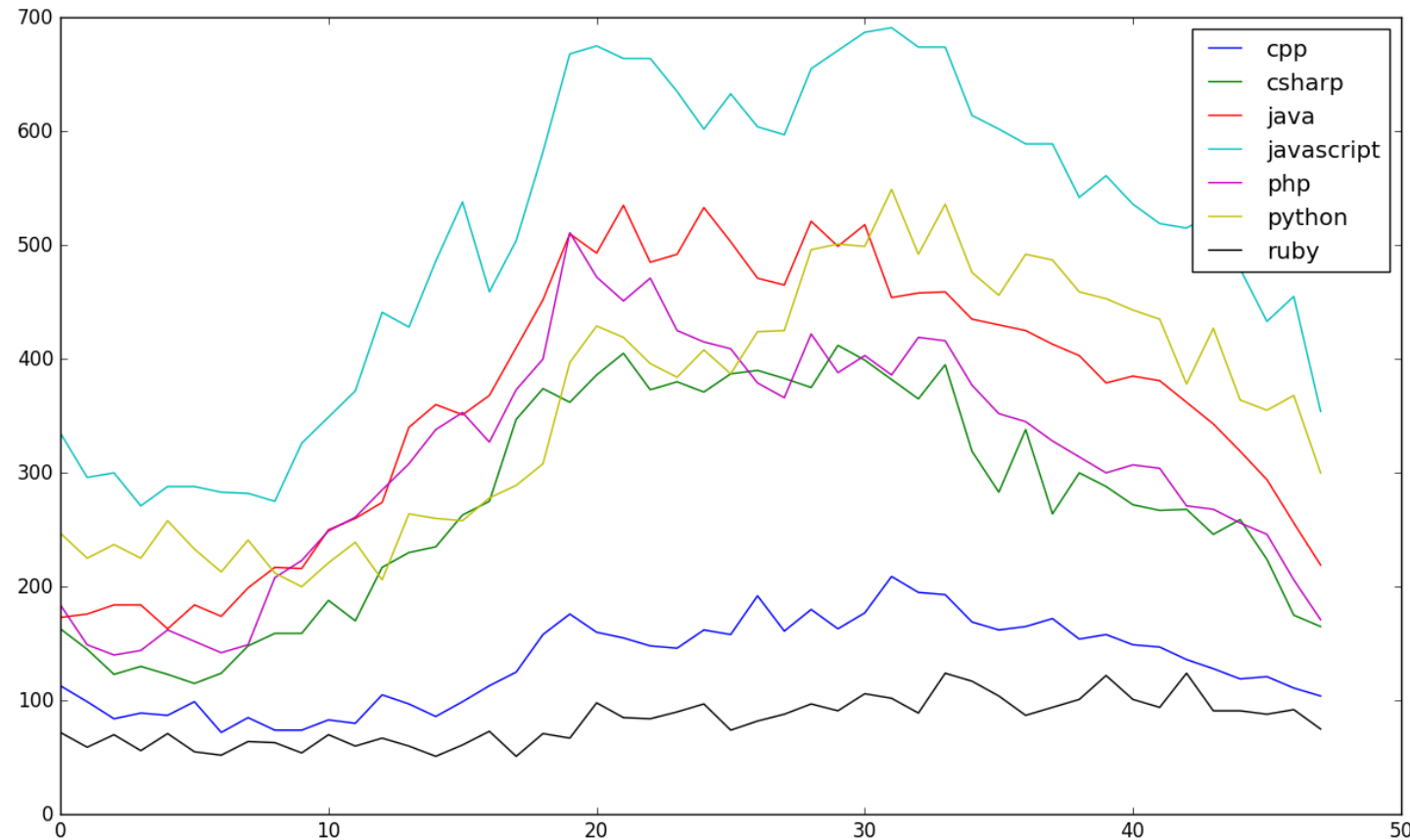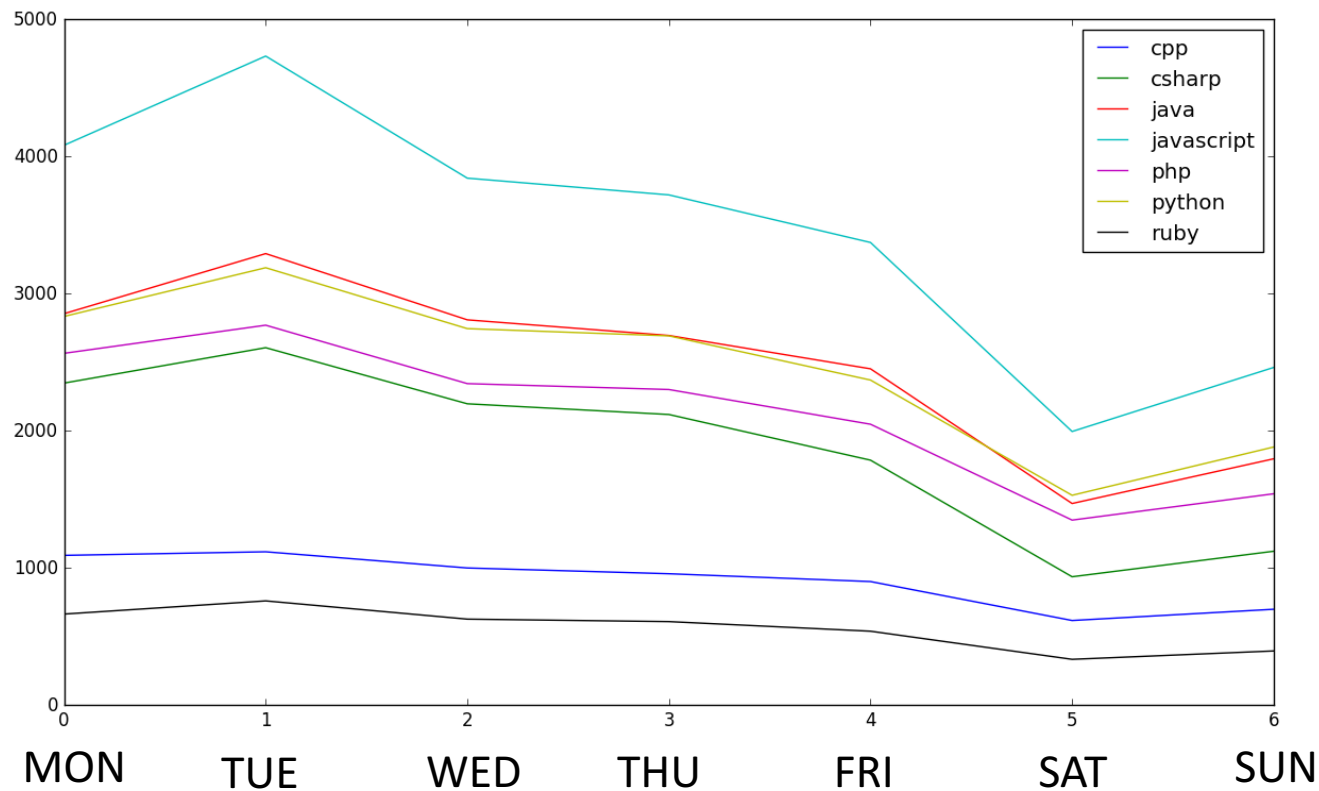

➢ C# may mainly be used during working time

➢ Python is more likely be used during the latter half of a day
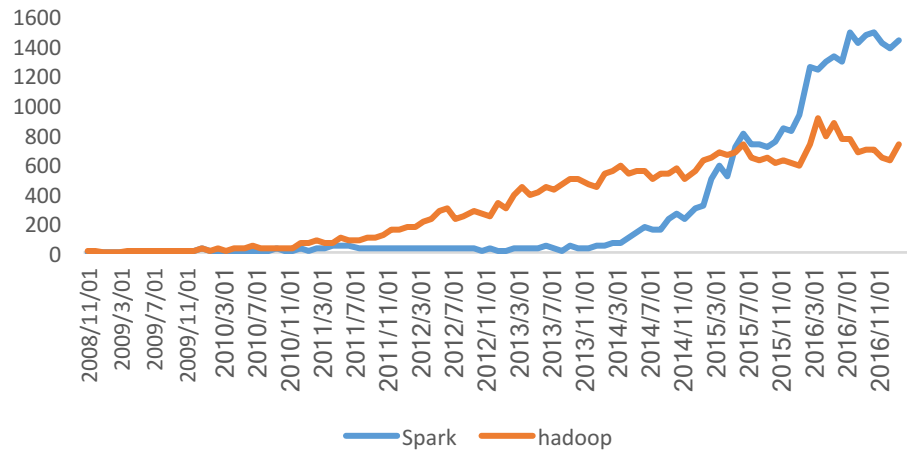
Basic Customer Behavior Analysis

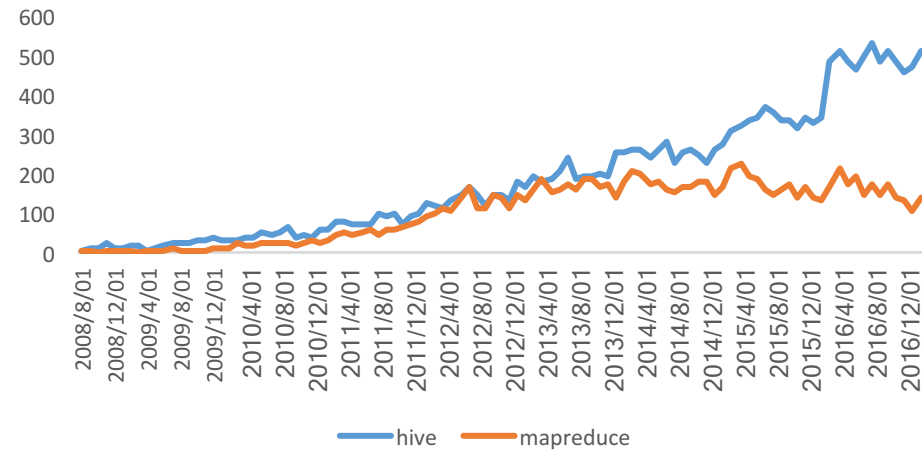- Questions post time over Programming Language

> Tuesday
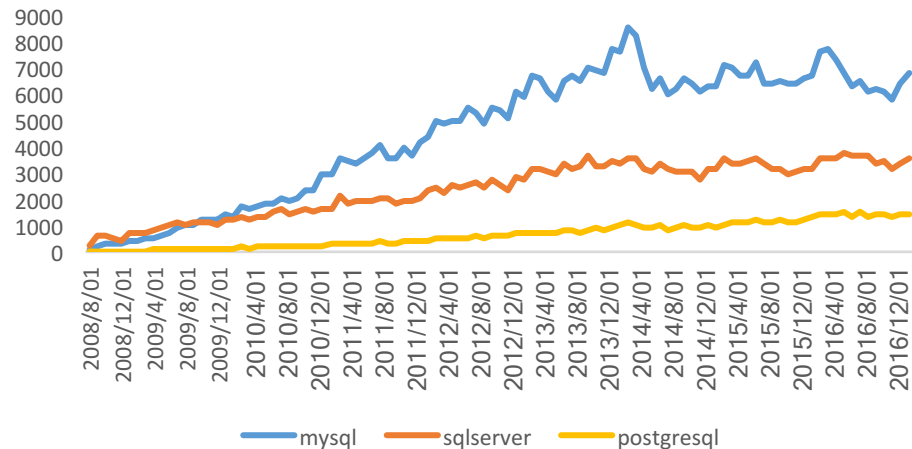> Python

Basic Customer Behavior Analysis

Apache Tools
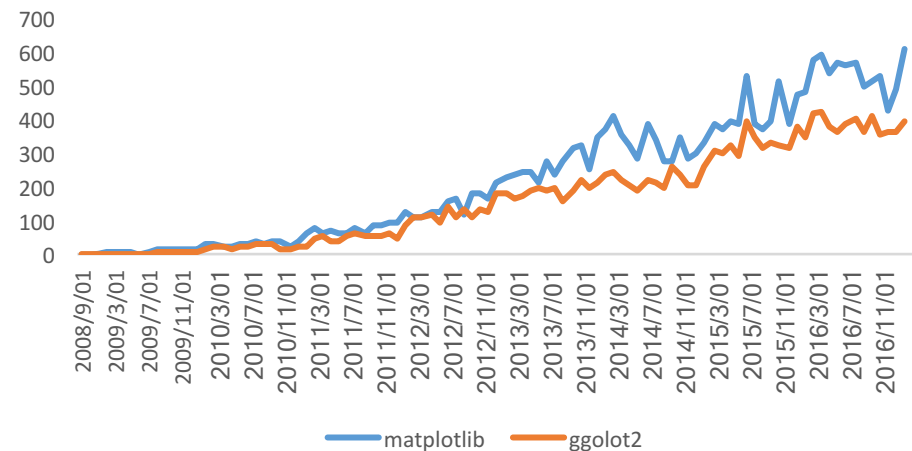
Processing data tools
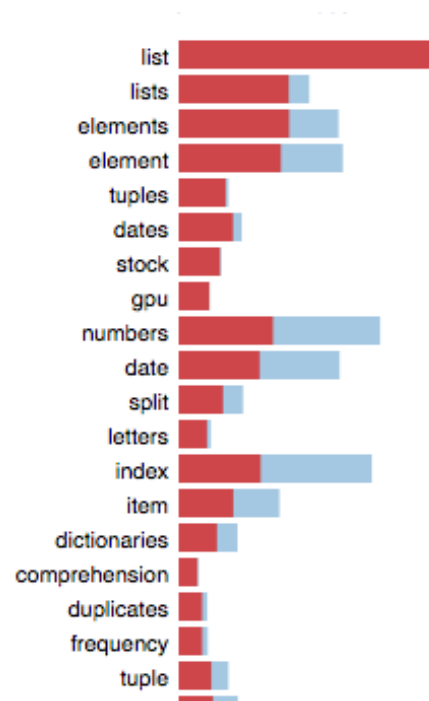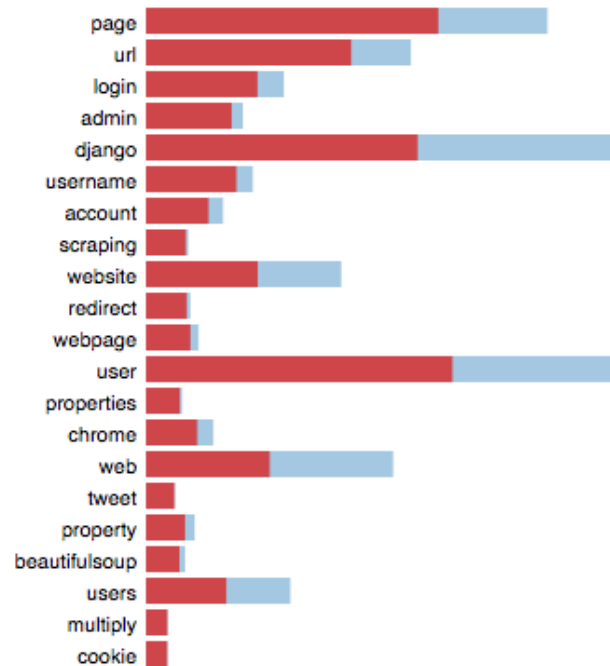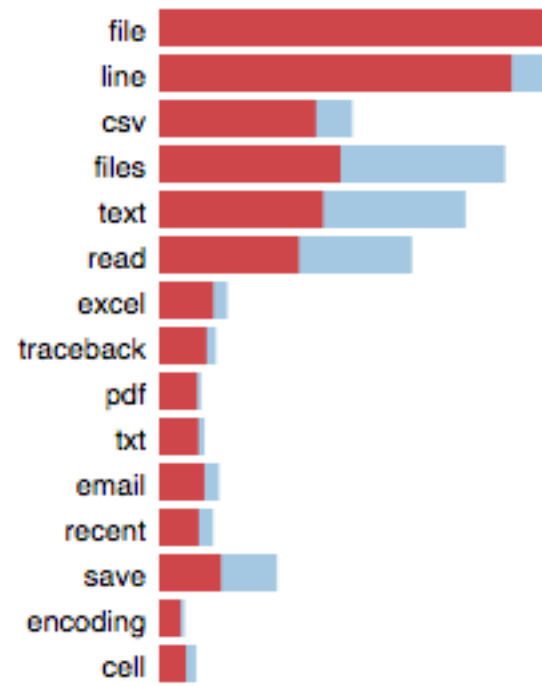
Database

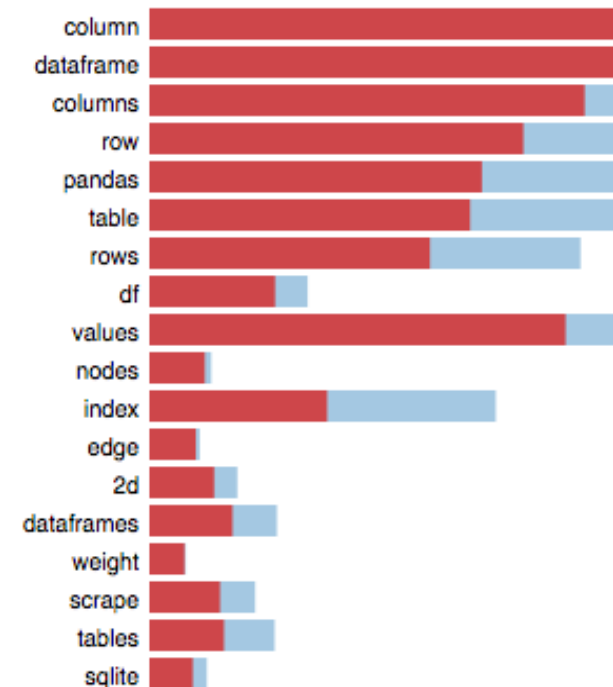Plot Tools for python and R

Topics for python questions(LDA)

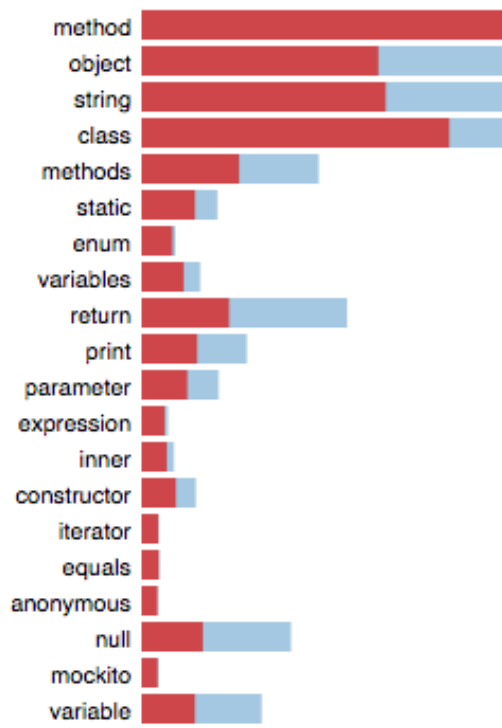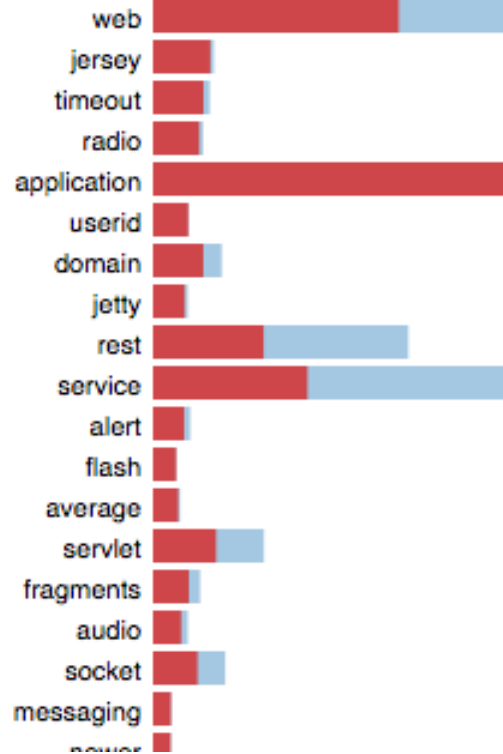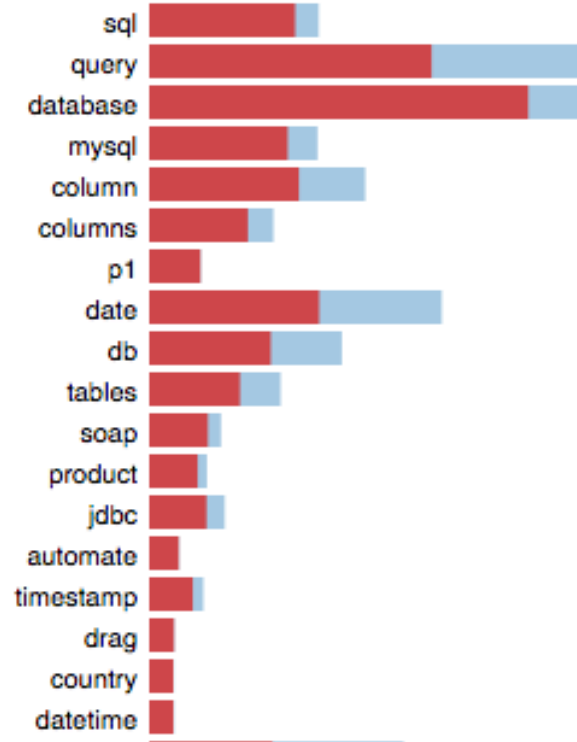Python Syntax   Web   File   Pandas

Topics for Java questions(LDA)

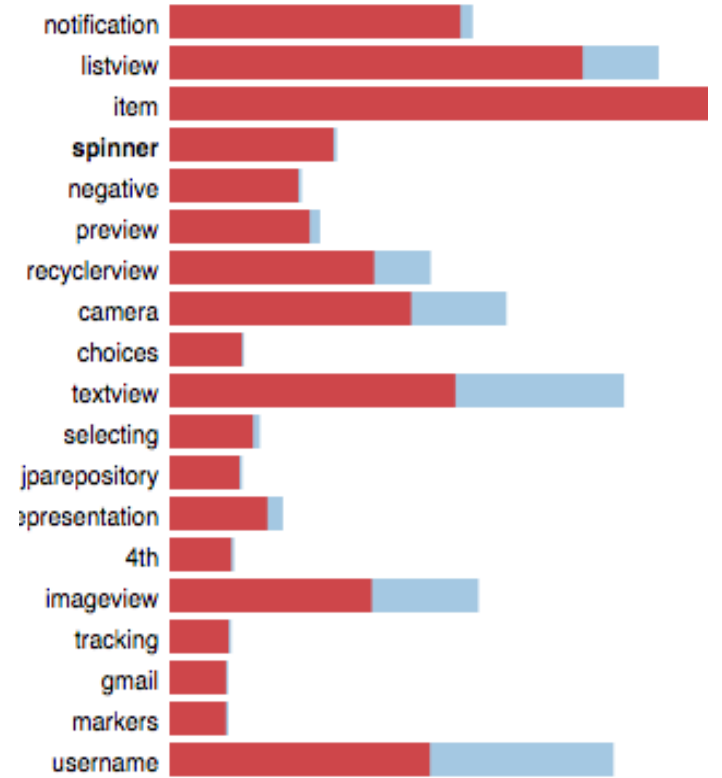
Java Syntax   Web   Database   Android

Part 3 Text Classification for Better Tags

Classification

- Only 20% of problems' description contain the programming language name.

Tags should be added by hand

Classification

- Using TF-IDF Vector + Classification Model

Error Rate

Model	Error Rate
XGBoost	0.286886
Logistic Regession	0.301942
Ramdom Forest	0.379721
Adaboost	0.456956
MultinomialNB	0.534883
Decision Tree	0.606107

Part 4 Sentiment Analysis for Comments

Sentiment Analysis Using Word2vec and LSTM

Text Source: Comments for questions

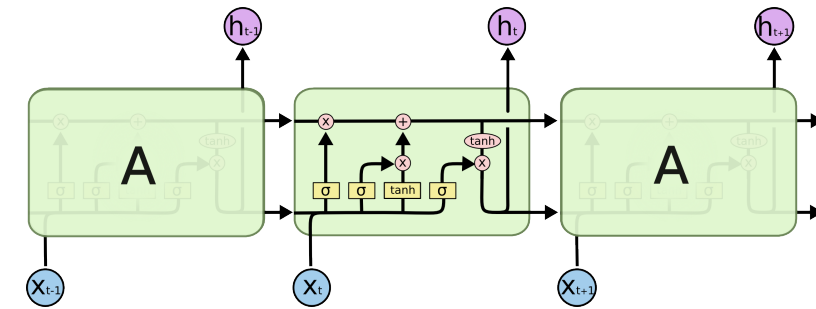# Sentiment Analysis Using Word2vec and LSTM
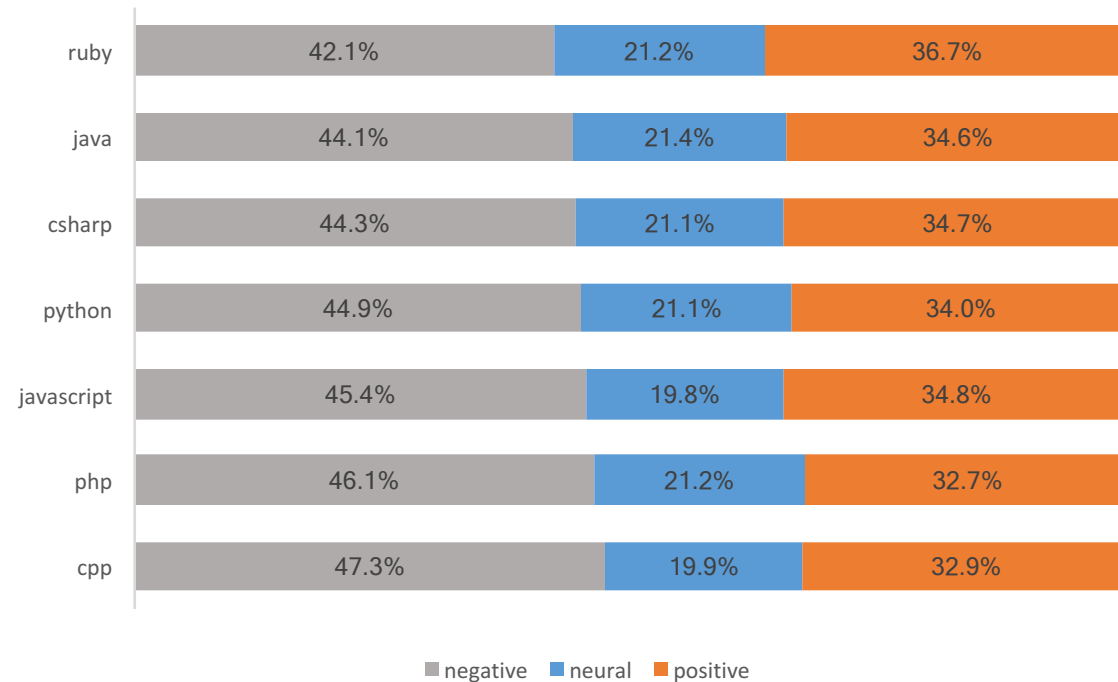
How Google sentiment works?
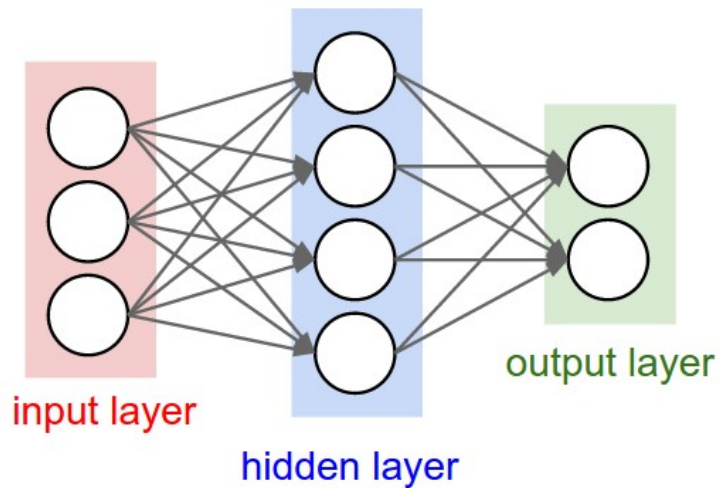
Neural Network
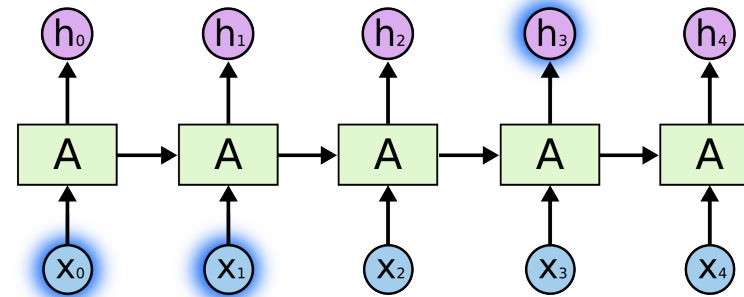
Recurrent Neural Network

Long Short Term Memory

Sentiment Analysis Using Word2vec and LSTM

Sentiment Score