



# SE text mining in StackExchange



Zhe

# Stackoverflow

StackExchange

sign up log in tour help stack overflow careers

search

stackoverflow

Questions Tags Users Badges Unanswered Ask Question

Stack Overflow is a community of 4.7 million programmers, just like you, helping each other. Join them, it only takes a minute: [Sign up](#)

git log format: shift commit body X columns to the right

Work on work you love. From home.

stackoverflow CAREERS

▲

0

I'm using the following custom log format to view my commits:  
Command: `git log --pretty=format:"%C(auto)%h %<(8,trunc)%aN %Cgreen%s %b"`

▼

★

3758d35 Daniel This commit does nothing  
You really should remove it before committing.

1. This is a line  
2. This is another line

a191c2b Daniel Viral helvetica lomo, typewriter fashion axe  
814a6a9 John Unam! pork belly pickled, fanny pack yr keffiyeh fap YOLO  
d5e130e Daniel Cardigan raw denim banjo  
f7187d8 Daniel 90's ramps pinterest, craft beer blue bottle

It works great except I'd like the commit body to be aligned with the commit title, and remove the last newline after the body. Is it possible to achieve using only git?

git

format

share

improve this question

edited 45 mins ago

asked 51 mins ago

Daniel Ström

57 • 9

add a comment

1 Answer

active oldest votes

asked today

viewed 6 times

active today

Upcoming Events

2015 Community Moderator Election

ends in 5 days

Blog

How To Target Job Listings Effectively

W

Art

debloo

Fidelity Investments

based in Cary, NC

20 open jobs

Principal Software Engineer -  
Hadoop, Big Data

# Tasks

---

- a. **Generate tags for posts on Stackoverflow** [1]
- b. Find most frequent topics on Stackoverflow (LDA, unsupervised) [2]
- c. Retrieve existing discussions on Stackoverflow given context [3]

[1] Stanley, C., & Byrne, M. D. (2013). Predicting tags for stackoverflow posts. *In Proceedings of ICCM* (Vol. 2013).

[2] Wang, S., Lo, D., & Jiang, L. (2013, March). An empirical study on developer interactions in StackOverflow. *In Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 1019-1024). ACM.

[3] Ponzanelli, L., Bavota, G., Di Penta, M., Oliveto, R., & Lanza, M. (2014, May). Mining stackoverflow to turn the IDE into a self-confident programming prompter. *In Proceedings of the 11th Working Conference on Mining Software Repositories* (pp. 102-111). ACM.

# Generate tags for posts on Stackoverflow

---

- a. Clayton and Byrne, 2013 [1]. ACT-R inspired Bayesian probabilistic model. 65% accuracy.
- b. Kuo, 2011 [2]. Co-occurrence model. 47% classification accuracy.
- c. Fu & Pirolli, 2007 [3]. SNIF-ACT. No result on Stackoverflow.

[1] Stanley, C., & Byrne, M. D. (2013). Predicting tags for stackoverflow posts. *In Proceedings of ICCM* (Vol. 2013).

[2] Kuo, D. (2011). On word prediction methods. *Technical report*, EECS Department, University of California, Berkeley.

[3] Fu, W. T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction*, 22(4), 355-412.

# Why accuracy

## Evaluation methods of multi-class tasks [1]

Measure	Formula	Evaluation focus
Average Accuracy	$\frac{\sum_{i=1}^I \frac{tp_i + tn_i}{tp_i + fp_i + fn_i + tn_i}}{I}$	The average per-class effectiveness of a classifier
Error Rate	$\frac{\sum_{i=1}^I \frac{fp_i + fn_i}{tp_i + fp_i + fn_i + tn_i}}{I}$	The average per-class classification error
Precision <sub>μ</sub>	$\frac{\sum_{i=1}^I tp_i}{\sum_{i=1}^I (tp_i + fp_i)}$	Agreement of the data class labels with those of a classifiers if calculated from sums of per-text decisions
Recall <sub>μ</sub>	$\frac{\sum_{i=1}^I tp_i}{\sum_{i=1}^I (tp_i + fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions
Fscore <sub>μ</sub>	$\frac{(\beta^2 + 1) Precision_{\mu} Recall_{\mu}}{\beta^2 Precision_{\mu} + Recall_{\mu}}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions
Precision <sub>M</sub>	$\frac{\sum_{i=1}^I tp_i}{\sum_{i=1}^I (tp_i + fp_i)}$	An average per-class agreement of the data class labels with those of a classifiers
Recall <sub>M</sub>	$\frac{\sum_{i=1}^I tp_i}{\sum_{i=1}^I (tp_i + fn_i)}$	An average per-class effectiveness of a classifier to identify class labels
Fscore <sub>M</sub>	$\frac{(\beta^2 + 1) Precision_M Recall_M}{\beta^2 Precision_M + Recall_M}$	Relations between data's positive labels and those given by a classifier based on a per-class average

[1] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

# Evaluation

---

## a. Fscore\_mu

weighted mean of F-scores of each class

$$\text{Sum} ( F(i) * \text{Population}(i) ) / \text{Sum} ( \text{Population}(i) )$$

## b. Fscore\_M

unweighted mean of F-scores of each class

$$\text{Sum} ( F(i) ) / \text{Number of Classes}$$

# What can SMOTE do

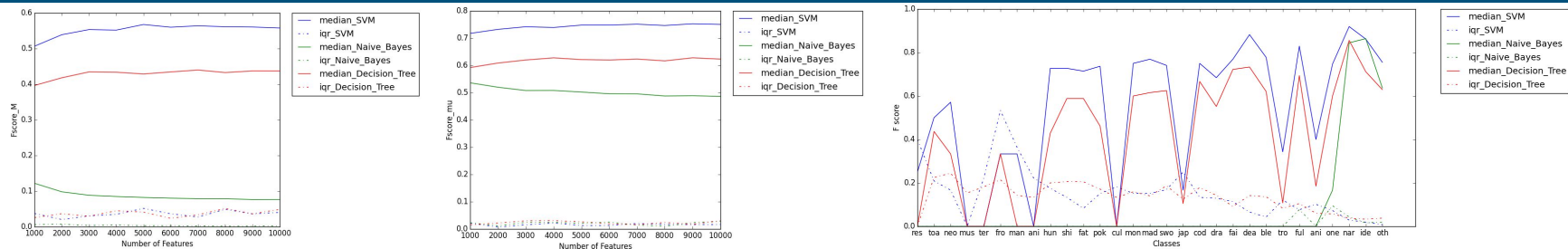
---

After SMOTE:

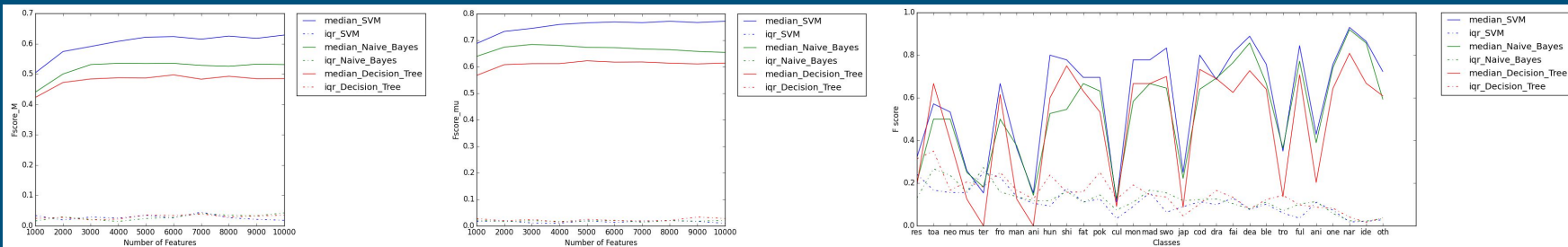
- a. Fscore\_mu goes down (not really maybe...)
- b. Fscore\_M goes up (significantly)

# SMOTE

Before:

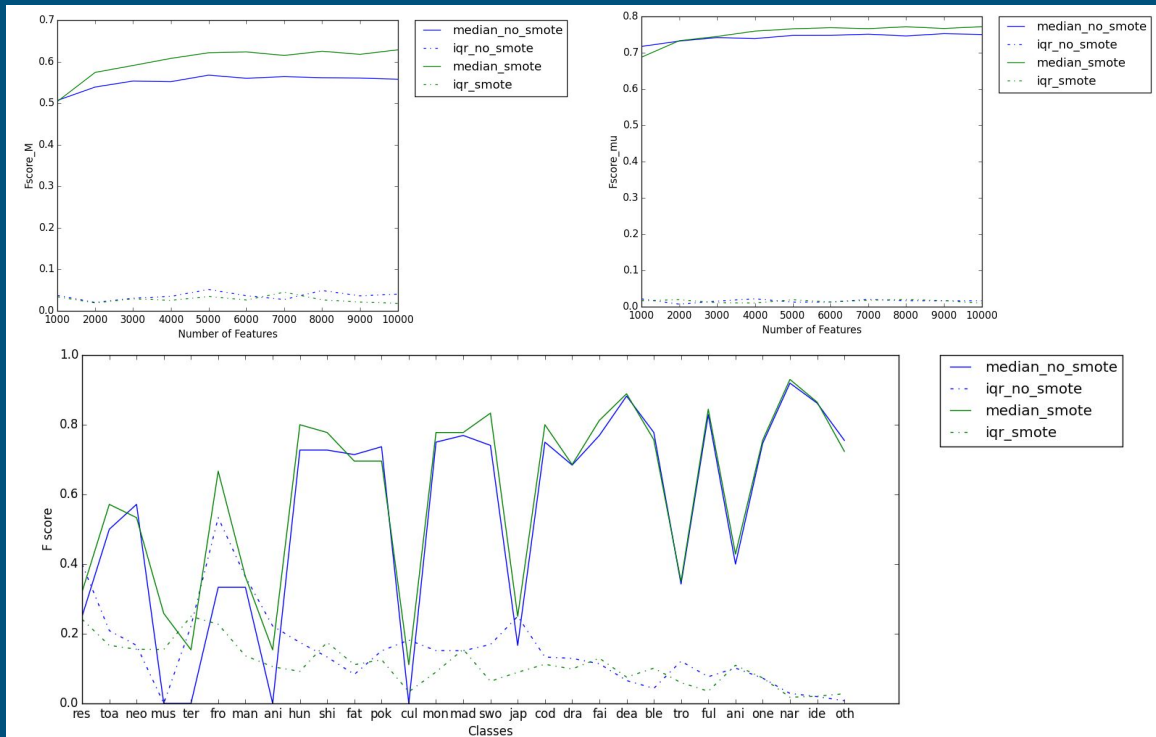


After:

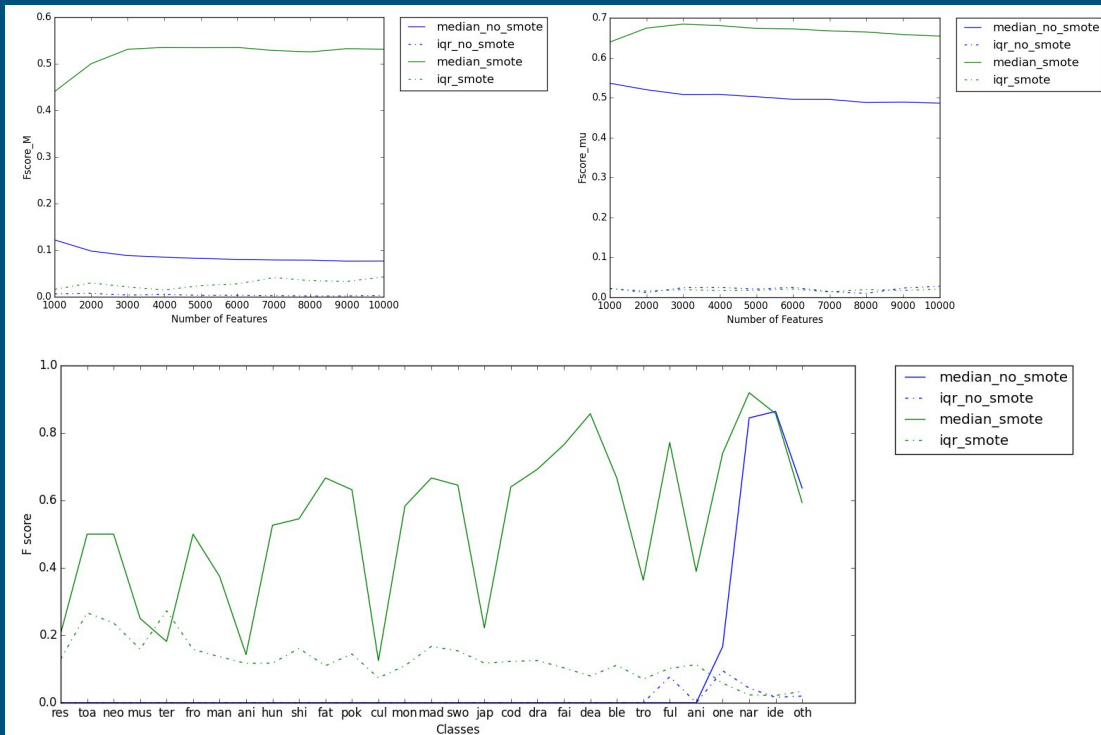




# SVM (Linear)



# Naive Bayes (Multinomial)



# Decision Tree (CART)

