Stack Overflow Question Classification

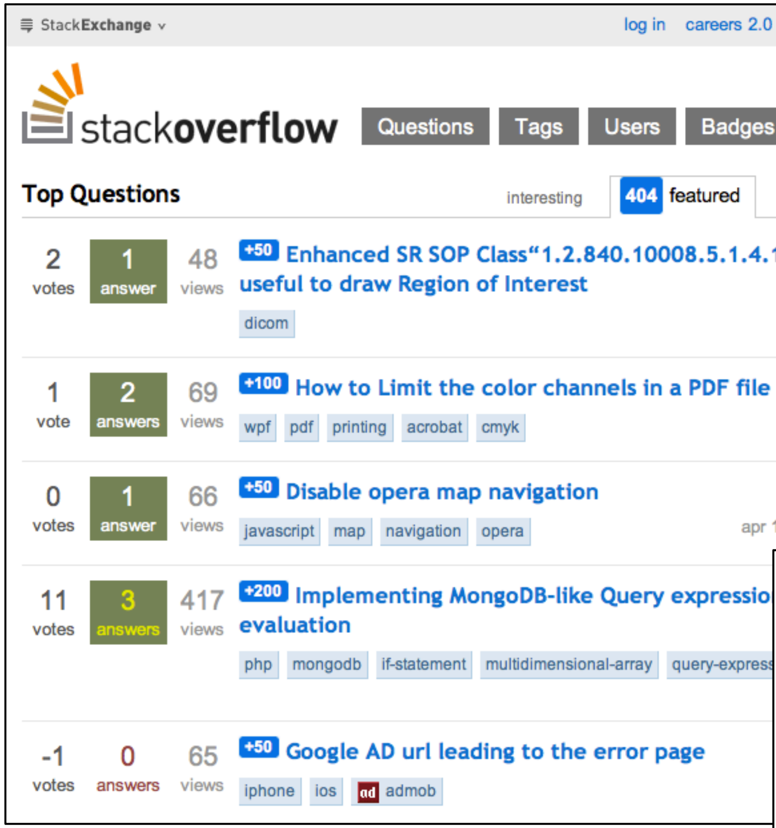
EEL709: Pattern Recognition

Arjun Attam, Dilpreet Singh Chahal and Vishnu Gupta

April 16, 2013

stackoverflow

- Community Q&A website on a wide range of topics in computer programming
- Encourages “practical, answerable questions based on actual problems” — chatty, open-ended questions are discouraged
- Bad questions are moderated and closed
- Multi-class classification: *open, non-constructive, off-topic, too localised, not a question*



Top Questions

2 votes · 1 answer · 48 views
+50 Enhanced SR SOP Class“1.2.840.10008.5.1.4.1.1.88.22” Is useful to draw Region of Interest
dicom
2d ago medPhys-pl 480

1 vote · 2 answers · 69 views
+100 How to Limit the color channels in a PDF file
wpf pdf printing acrobat cmyk
1h ago santa 419

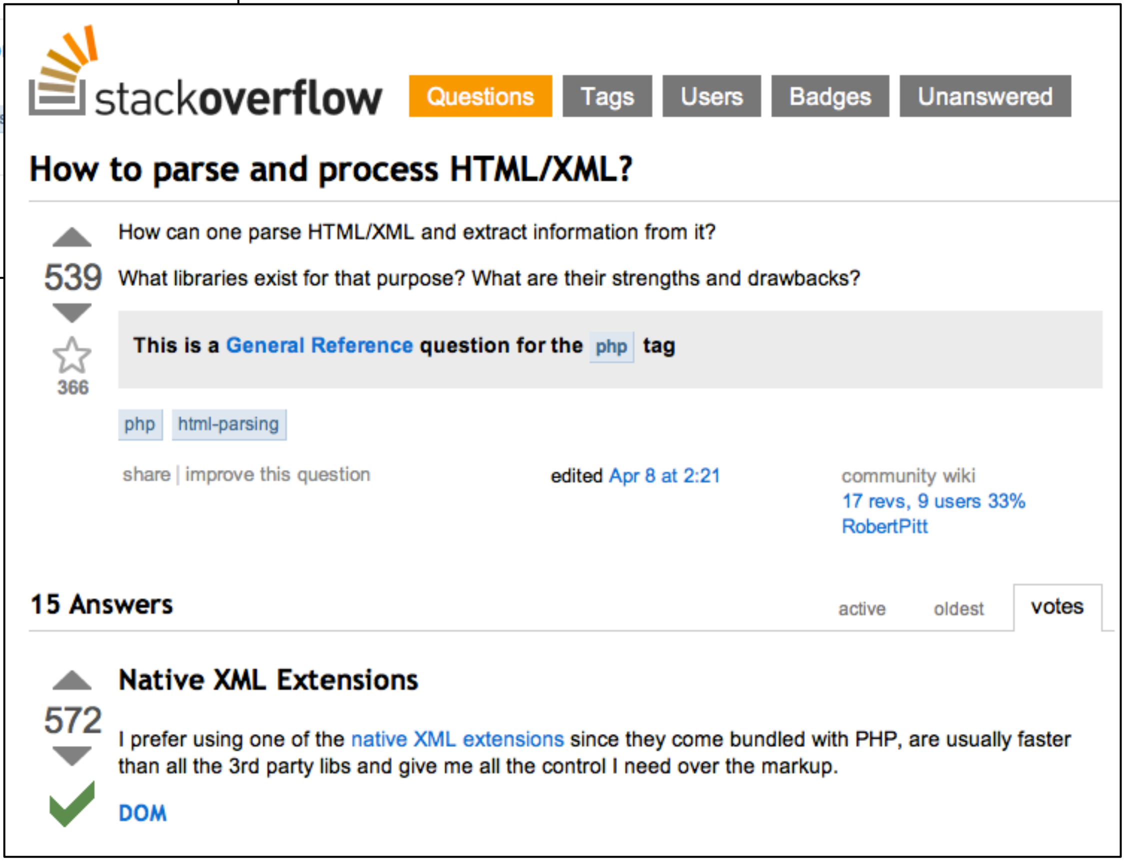
0 votes · 1 answer · 66 views
+50 Disable opera map navigation
javascript map navigation opera
apr 12 at 5:14 Pim Schaaf 86

11 votes · 3 answers · 417 views
+200 Implementing MongoDB-like Query expressio evaluation
php mongodb if-statement multidimensional-array query-express

-1 votes · 0 answers · 65 views
+50 Google AD url leading to the error page
iphone ios ad admob

Stack Overflow homepage

One of the 4M questions



How to parse and process HTML/XML?

539

366

How can one parse HTML/XML and extract information from it?

What libraries exist for that purpose? What are their strengths and drawbacks?

This is a General Reference question for the php tag

php html-parsing

share | improve this question

edited Apr 8 at 2:21

community wiki
17 revs, 9 users 33%
RobertPitt

15 Answers

572

Native XML Extensions

I prefer using one of the native XML extensions since they come bundled with PHP, are usually faster than all the 3rd party libs and give me all the control I need over the markup.

DOM

Dataset

- 3.4M questions from Stack Overflow till July 31, 2013
  - *PostId, PostCreationDate*
  - *OwnerUserId, OwnerCreationDate*
  - *ReputationAtPostCreation*
  - *OwnerUndeletedAnswerCountAtPostTime*
  - *Title, BodyMarkdown*
  - *Tag1, Tag2, Tag3, Tag4, Tag5*
  - *PostClosedDate, OpenStatus*

First attempt

- Logistic regression with bag of words (1-grams)

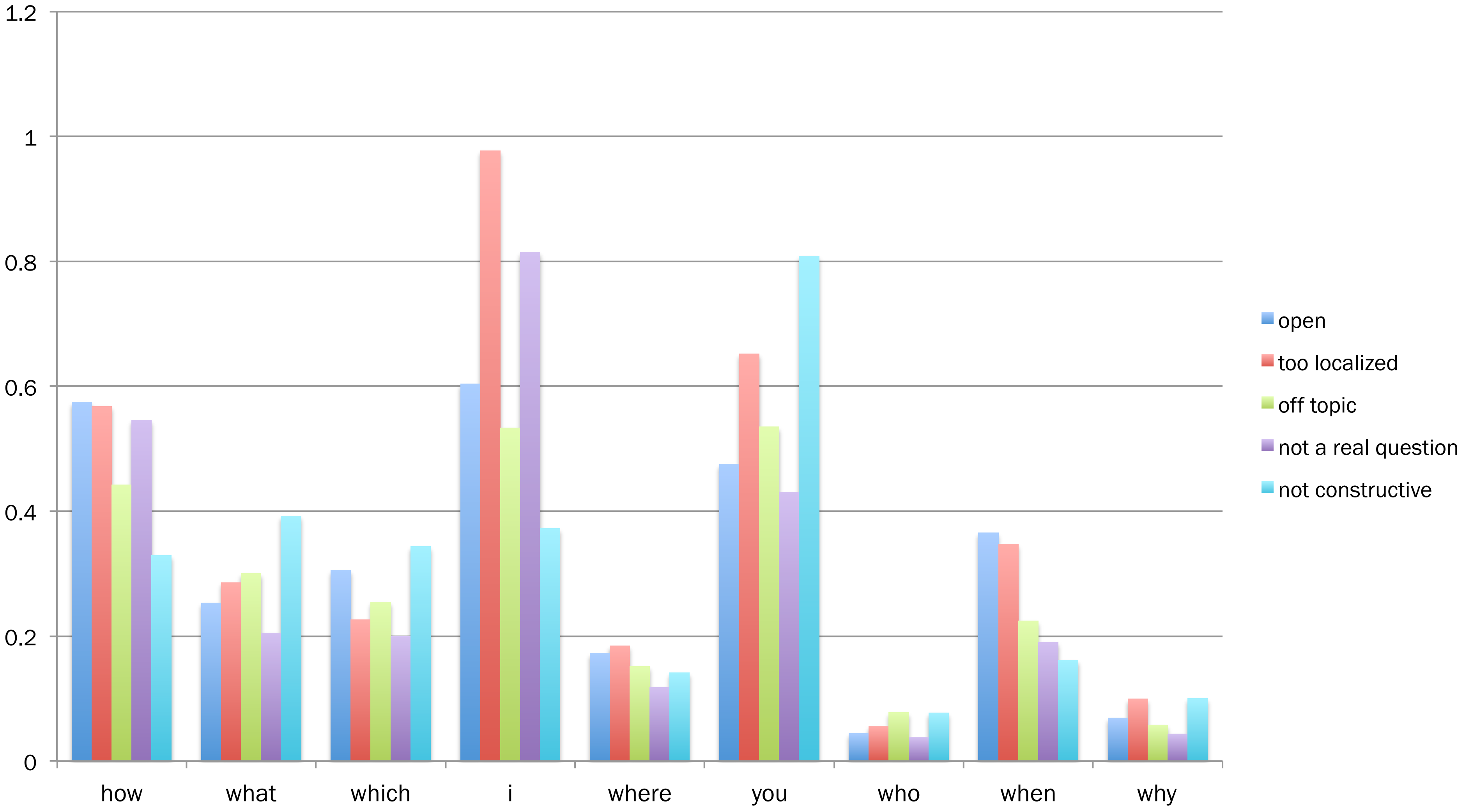
- Poor results: 24% accuracy

- Coming up: a better look at the dataset

Features

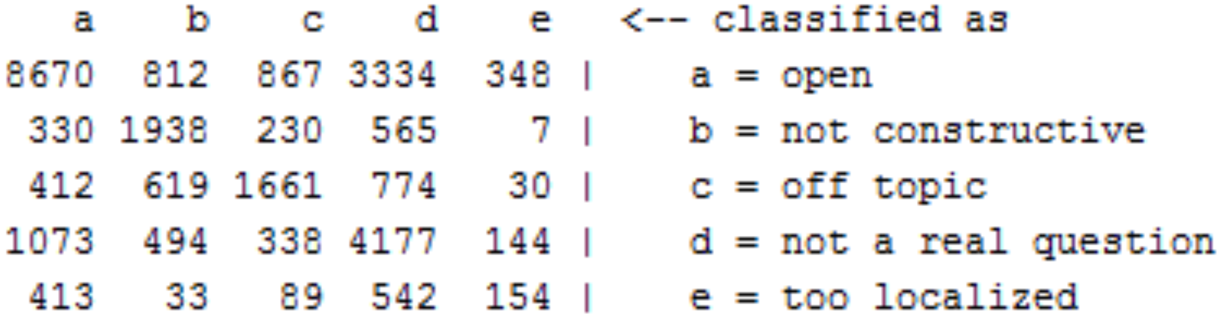
- Questions words: what, who, when, how, which, where, why
- Pronouns: I, you
- Positive features:
  - Presence of a code sample
  - High user reputation

- Bad tags
  - Directly related to *off-topic* class
- New sampling
  - Stratification to independently sample subpopulations


1.2
1
0.8
0.6
0.4
0.2
0
how
what
which
i
where
you
who
when
why
open
too localized
off topic
not a real question
not constructive

Naïve Bayes

- Bag of words
  - Numeric
    - Tries to fit Gaussian model; calculates variance
    - Computationally very expensive
  - Binary
    - Bernoulli model
    - 59.17% accuracy



```
   a    b    c    d    e   <-- classified as
8670  812  867 3334  348 |    a = open
 330 1938  230  565    7 |    b = not constructive
 412  619 1661  774   30 |    c = off topic
1073  494  338 4177  144 |    d = not a real question
 413   33   89  542  154 |    e = too localized
```

Incremental Naïve Bayes

- Bag of words
  - Numeric
    - Tries to fit Gaussian model; assumes variance of 0.1
    - Accuracy of 37.12%
  - Binary
    - Bernoulli model
    - Accuracy of 53.91%

- N-grams
  - Binary: Roughly the same accuracy (56.52%)

Multinomial Naïve Bayes

- Recommended for unbalanced text classification problems

$$P(c|d) = \frac{P(c)\prod_{w\in d} P(w|c)^{n_{wd}}}{P(d)}, \qquad \alpha \times \frac{n_{wd}}{\sum_{w'}\sum_{d\in D_c} n_{w'd}},$$

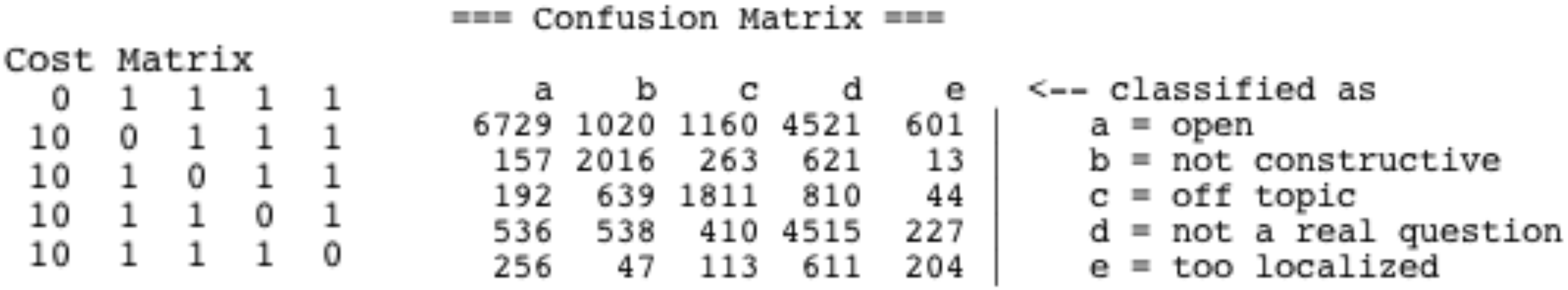
- Replace exponent: normalize the word counts in each class so that the total size of the classes is the same for both classes after normalization

- Accuracy of 58.44%

- Ignores non-string features

Cost Matrix

- Penalize assigning closed questions to *open* class
  - Accuracy: 54.34%



```
                          === Confusion Matrix ===
Cost Matrix
  0   1   1   1   1            a    b    c    d    e   <-- classified as
 10   0   1   1   1         6729 1020 1160 4521  601 |    a = open
 10   1   0   1   1          157 2016  263  621   13 |    b = not constructive
 10   1   1   0   1          192  639 1811  810   44 |    c = off topic
 10   1   1   1   0          536  538  410 4515  227 |    d = not a real question
                             256   47  113  611  204 |    e = too localized
```

- Results worsen with more cost (50)

To do

- Diversity metrics—Yule’s Q statistic: lower Q values indicate greater diversity
- Minimizing expected cost using the cost matrix
- Other Q&A websites of the Stack Exchange network

Thank you

Questions?