

Data Mining Tutorial

Session 2: Stack Overflow data set

Erich Schubert, Eirini Ntoutsis

Ludwig-Maximilians-Universität München

2012-05-10 — KDD class tutorial

Stack overflow is a programming QA website:

<http://stackoverflow.com/>

- ▶ Users post programming questions
- ▶ Other users post answers
- ▶ Up- and Downvotes on questions and answers
- ▶ Awards for good answers, questions and active users
- ▶ Tags to organize questions
- ▶ Moderation by users with high reputation
- ▶ Size: 2.8m questions, 5.8m answers, 11m comments, 22m votes, 30k tags

(Yes, you could post your homework questions there. This is not recommended, as usually the teachers want you to solve the problems yourself to learn from the problem, not the solution. Plus, a good question there should already contain source code)

First (non-deleted) question on SO: When setting a form's opacity should I use a decimal or double?

127



13



I'm new to C# and I want to use a track-bar to change a form's opacity.

This is my code:

```
decimal trans = trackBar1.Value / 5000;
this.Opacity = trans;
```

When I try to build it, I get this error:

Cannot implicitly convert type 'decimal' to 'double'

I tried making `trans` a double, but then the control doesn't work. This code worked fine for me in VB.NET.

What do I need to do differently?

[c#](#) [winforms](#) [forms](#) [type-conversion](#) [opacity](#)

[link](#) | [edit](#) | [close](#) | [flag](#)

edited May 4 at 8:55

 [mattytommo](#)
3,775 ●3 ●8 ●18

asked Jul 31 '08 at 21:42

 [Eggs McLaren](#)
763 ●1 ●6 ●9

31 Oh Eggs....Every time I read this question I chuckle at your name. Your parents must be proud. – [jjinguy](#) ♦
Jan 5 '09 at 1:39

95 Ya know, this seems to be the first question on SO ever! – [cletus](#) Jan 31 '09 at 6:39

StackOverflow publishes data dumps:

<http://blog.stackoverflow.com/category/cc-wiki-dump/>

- ▶ A torrent download with about 5 GB. 7zip-compressed.
- ▶ About 4 GB for the main stackoverflow site.
- ▶ Main `.xml` file is 8 GB, post history is 11 GB.

So this means:

- ▶ That is pretty big!
- ▶ Maybe *not* everyone here should download it.
- ▶ I will not demo this live, but provide result data for you.
- ▶ You can not load the XML in your DOM parser.
- ▶ In fact, you might even be unable to decompress it (due to a 4 GB file size limit on many file systems).



Stack Overflow data set

A first peek inside the 7zip file.

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Introduction

Downloading

Preprocessing

Apriori FIM

Conclusions

```
> 7z l stackoverflow.com.7z.001
```

Date	Time	Size	Compressed	Name
2011-09-06	20:05:06	170594039	457479414	092011 Stack Overflow/badges.xml
2011-09-06	20:04:52	1916999879		092011 Stack Overflow/comments.xml
2011-09-06	21:10:00	10958639384	1841985260	092011 Stack Overflow/posthistory.xml
2011-09-06	19:57:53	7569879502	1454543990	092011 Stack Overflow/posts.xml
2011-09-06	20:00:50	193250161	132626278	092011 Stack Overflow/users.xml
2011-09-06	19:59:56	1346527241		092011 Stack Overflow/votes.xml
2011-06-13	09:26:12	1786		092011 Stack Overflow/license.txt
2011-09-06	19:41:44	4780		092011 Stack Overflow/readme.txt
2011-09-06	20:05:07	0	0	092011 Stack Overflow
		22155896772	3886634942	8 files, 1 folders

We are interested in the `posts.xml` file for our Apriori experiment.

So we will need to preprocess the data to get it into a “workable” size. Here is what we have to do:

- ▶ Decompress into a *stream*, not to the harddisk.
- ▶ Streaming XML processing using XML pull (to avoid processing the full XML file at once)
- ▶ Select interesting data only and dump these parts only

So, let us have a peek into the `posts.xml` file.

Inspect the file with: `7z x -so stackoverflow.com.7z.001`

"092011 Stack Overflow/posts.xml" | less

```
<?xml version="1.0" encoding="utf-8"?>
<posts>
<row Id="4" PostTypeId="1" AcceptedAnswerId="7"
CreationDate="2008-07-31T21:42:52.667" Score="83" ViewCount="7351"
Body="<p>I'm new to C#, and I want to use a track-bar to change a form's
opacity.</p>&#xA;&#xA;<p>This is my
code:</p>&#xA;&#xA;<p>pre<code>decimal trans = trackBar1.Value
/ 5000&#xA;this.Opacity =
trans&#xA;</code></pre>&#xA;&#xA;<p>When I try to build it, I
get this error:</p>&#xA;&#xA;<blockquote>&#xA;  <p>Cannot
implicitly convert type 'decimal' to
'double'</p>&#xA;</blockquote>&#xA;&#xA;<p>I tried making
<code>trans</code> a double, but then the control doesn't work.
This code worked fine for me in VB.NET. </p>&#xA;&#xA;<p>What do I
need to do differently?</p>&#xA;" OwnerUserId="8"
LastEditorUserId="140328" LastEditorDisplayName="Rich B"
LastEditDate="2011-08-20T23:22:24.213"
LastActivityDate="2011-08-31T19:42:18.077" Title="When setting a form's opacity
should I use a decimal or double?" Tags="<c#><winforms>"
AnswerCount="12" CommentCount="19" FavoriteCount="13" />
[...]
```

Looks worse than it is:

The `Tags=" . . . "` attribute is quite simple, it decodes to:
`<c#><winforms>`

- ▶ Stream via `7z x -so`
- ▶ Process one `<row . . ./>` element at a time
- ▶ Extract `Tags=" . . . "` attribute
- ▶ Extract tags by `<.*>` pattern
- ▶ Output tags for use in Apriori into an ascii file
- ▶ Do not include anything else: no ID, no title — we do not need them here
- ▶ Worst case size estimation: 2.8m questions $\times$ up to 5 tags $\times$ 20 characters = 280 MB
- ▶ Reality: 50 MB uncompressed, 15 MB gzip.



Stack Overflow data set

Preprocessing — posts.xml in python

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Introduction

Downloading

Preprocessing

Apriori FIM

Conclusions

```
#!/bin/python
import subprocess, sys, re
from xml.dom import pulldom

archive = "stackoverflow.com.7z.001"
xmlfile = "092011 Stack Overflow/posts.xml"

cmd = ["7z", "x", "-so", archive, xmlfile]

proc = subprocess.Popen(cmd, stdin=None,
                        stdout=subprocess.PIPE, shell=False)
events = pulldom.parse(proc.stdout)

for event, node in events:
    if event == pulldom.START_ELEMENT:
        if node.tagName == "row":
            events.expandNode(node)
            processRow(node) # NEXT SLIDE
            # print node.toxml()

proc.stdout.close()
```

And the main processing function:

```
tagre = re.compile("<([>]+)>")
```

```
def processRow(node):
    # Questions (type 1) only
    typ = node.getAttribute("PostTypeId")
    if typ != "1":
        return
    # Get Tags attribute
    tags = node.getAttribute("Tags")
    if not tags or len(tags) == 0:
        return
    # Remove the <> wrappers, separate by space.
    print " ".join(tagre.findall(tags))
```



Stack Overflow data set

Preprocessed data file — `all-tags.txt`

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Introduction

Downloading

Preprocessing

Apriori FIM

Conclusions

Resulting data set:

```
c# winforms
html css internet-explorer-7
c# conversion j#
c# datetime
c# .net datetime timespan
html browser time timezone
c# math
c# ling web-services .net-3.5
mysql database
performance algorithm language-agnostic unix pi
php
mysql database triggers
c++ c sockets mainframe zos
flex actionscript-3
sql-server datatable
c# .net vb.net timer
```

Now, let us do some Apriori on this data set!

Weka unfortunately does not scale up well to this data set size.

Plus, we first would need to convert it into the `.arff` file for Weka.

Why not just write Apriori ourselves?

Choosing appropriate `minsup` values might be tricky, too.

So we will just look at the top itemsets in each run.

With just 50 MB of uncompressed data, we *should* be able to keep all of them in memory!

Python is for lazy people. Loading text data is easy:

```
#!/bin/python
import gzip

db= [ ]

for line in gzip.open("all-tags.txt.gz"):
    db.append(line.strip().split(" "))

print "Database size:", len(db)
```

Output:

Database size: 2012348

Class to represent an itemset:

```
class itemset():
    def __init__(self, tokens, support=0):
        self.tokens = list(tokens)
        self.support = support

    def tokenstr(self):
        return "+".join(self.tokens)

    def __str__(self):
        return self.tokenstr()+": "+str(self.support)

    def __cmp__(self, other):
        return cmp(self.tokenstr(), other.tokenstr())

    def __hash__(self):
        return hash(self.tokenstr())
```

```
oneitems = dict()
for rec in db:
    for tag in rec:
        item = itemset([tag])
        item = oneitems.setdefault(item, item)
        item.support += 1
oneitems = list(oneitems.keys())

# Inspect:
oneitems.sort(lambda a,b: cmp(b.support, a.support))
print len(oneitems), map(str, oneitems[:10])
print str(oneitems[100]), str(oneitems[200])
```

Output:

```
29551 ['c#: 211338', 'java: 153561', 'php: 142125',
'javascript: 126296', 'jquery: 109129', 'iphone: 96748',
'android: 93247', '.net: 89646', 'asp.net: 88938',
'c++: 84777']
asp.net-mvc-2: 7219 c#-4.0: 4055
```

So we will try with minsupport= 1000:

```
minsupport = 1000
oneitems = filter(
    lambda x: x.support >= minsupport,
    oneitems)
itemsets = [oneitems]

print len(oneitems)
```

Output:

777

Generating candidates:

```
def apriorigen(curitems):  
    curitems.sort() # by tags  
    for i in range(0, len(curitems) - 1):  
        toka = curitems[i].tokens  
        for j in range(i + 1, len(curitems)):  
            tokb = curitems[j].tokens  
            # Prefix test:  
            if not toka[:-1] == tokb[:-1]: break  
            cand = toka + tokb[-1:] # Extend with last  
            # Pruning test:  
            ok = True  
            for i in range(len(cand) - 2):  
                t = cand[:i] + cand[i+1:] # without i  
                if not contains(curitems, itemset(t)):  
                    ok = False  
                    break  
            if ok: yield itemset(cand) # generate itemset
```

Main loop for FIM:

```
while True:
    size = len(itemsets) + 1
    cand = dict()
    for c in apriorigen(itemsets[-1]): cand[c] = c
    if len(cand) == 0: break
    for rec in db:
        for subset in itertools.combinations(rec, size):
            subset = cand.get(itemset(subset))
            if subset: subset.support += 1
    itemsets.append(filter(
        lambda i: i.support >= minsupport, cand))
    itemsets[-1].sort(lambda a,b: cmp(b.support, a.support))
    print size, map(str, itemsets[-1][:5])
```

Output:

```
['ruby+ruby-on-rails: 12266', 'c#+winforms: 12236',
 'android+java: 11252', 'c#+wpf: 11143', 'ios+iphone: 10431']
['c#+wpf+xaml: 1600', 'mysql+php+sql: 1064']
```

Setting minsupport = 500 finds more itemsets:

Output:

```
['ruby+ruby-on-rails: 12266', 'c#+winforms: 12236', 'android+java: 11252',
'c#+wpf: 11143', 'ios+iphone: 10431', 'cocoa-touch+iphone: 10371',
'c#+linq: 8184', 'c+c++: 8131', 'ajax+javascript: 7951', 'java+swing: 7296',
'mysql+sql: 7018', 'cocoa+objective-c: 6670', 'cocoa-touch+objective-c: 6598',
'jquery+php: 6564', 'ipad+iphone: 6311', 'asp.net+javascript: 5761',
'sql-server+tsql: 5728', 'jquery+jquery-ui: 5728', 'eclipse+java: 5537',
'hibernate+java: 5336']
```

```
['c#+wpf+xaml: 1600', 'mysql+php+sql: 1064', 'cocoa-touch+ios+iphone: 935',
'cocoa-touch+iphone+uikit: 879', 'hibernate+java+orm: 838',
'activerecord+ruby+ruby-on-rails: 783', 'c#+databinding+wpf: 754',
'gui+java+swing: 730', 'oracle+plsql+sql: 698',
'cocoa-touch+iphone+uitableview: 693', 'cocoa+cocoa-touch+iphone: 652',
'cocoa+cocoa-touch+objective-c: 570', 'database+mysql+sql: 567',
'database+database-design+mysql: 532', 'ios+iphone+uitableview: 514',
'c#+silverlight+windows-phone-7: 513', 'cocoa-touch+ipad+iphone: 500']
```

- ▶ The results were okay, but not very surprising (in fact, most results are very obvious!)
- ▶ The data contains redundant tags (mysql, sql)
- ▶ 5 tag limit affects output
- ▶ Data mining does not guarantee new results, unfortunately