集合知プログラミング
6.9～6.10

発表者：林華

改善策

- テキストの代わりにエントリを丸ごと利用
- タイトル、要約などの単語の集合も利用
- 大文字と小文字を区別して扱う
- メタ情報をもっと捕捉する
- URLと数を完全な状態で保存する

エントリを丸ごと利用

テキスト利用の場合：

```
fulltext='%s¥n%s¥n%s' %
        (entry['title'],entry['publisher'],entry['summary'])
print 'Guess: '+str(classifier.classify(fulltext))
cl=raw_input('Enter category: ')
classifier.train(fulltext,cl)
```

エントリ丸ごと利用の場合：

```
print 'Guess: '+str(classifier.classify(entry))
cl=raw_input('Enter category: ')
classifier.train(entry,cl)
```

タイトル、要約の単語集合の利用

- # Extract the title words and annotate

```
titlewords=[s.lower() for s in
      splitter.split(entry['title'])
       if len(s)>2 and len(s)<20]
for w in titlewords: f['Title:'+w]=1
```

- # Extract the summary words

```
summarywords=[s.lower() for s in
      splitter.split(entry['summary'])
       if len(s)>2 and len(s)<20]
```

python_search.xml

<item><title>My new baby boy!</title><link>http://blog.myspace.com/index.cfm?fuseaction=blog.view&friendID=43128005&blogID=184109011</link><description>THis is my new baby, Anthem. He is a 3 and half month old ball python, orange shaded normal pattern. I have held him about 5 time since I brought him home tonight at 8:00pm...</description><dc:publisher>Shetan Noir, the zombie belly dancer! - MySpace Blog</dc:publisher><dc:creator>unknown</dc:creator><dc:date>2006-10-24T02:35:50Z</dc:date></item>

<item><title>If you need a laugh...</title><link>http://greeneyedmermaid.spaces.live.com/Blog/cns!9B007E8A1F8FED56!161.entry</link><description>Even does 'funny walks' from Monty Python. He talks about all the ol' ladies that are after him. He teases me about my horror obsession. He attempts suicide. And best of all, he talks about poo. Who doesn't think poo is funny???!</description><dc:publisher>Kate's space</dc:publisher><dc:creator>unknown</dc:creator><dc:date>2006-10-24T02:17:29Z</dc:date></item>

大文字と小文字を区別して扱う

```
# Count uppercase words
  uc=0
  for i in range(len(summarywords)):
    w=summarywords[i]
    f[w]=1
    if w.isupper(): uc+=1
```

```
IDLE 2.6.2
>>> import re
>>> splitter=re.compile('\\W*')
>>> line='This are test string strING STRING'
>>> t=[s.lower() for s in splitter.split(line) if len(s)>2]
>>> t
['this', 'are', 'test', 'string', 'string', 'string']
>>> for i in range(len(t)):
        w=t[i]
        if w.isupper(): print w


>>> t=[s for s in splitter.split(line) if len(s)>2]
>>> t
['This', 'are', 'test', 'string', 'strING', 'STRING']
>>> for i in range(len(t)):

KeyboardInterrupt
>>> for i in range(len(t)):
        w=t[i]
        if w.isupper(): print w


STRING
>>> for i in range(len(t)):
        w=t[i]
        if w.isupper() or w.istitle(): print w


This
STRING
>>>
```

メタ情報をもっと捕捉する

```
# Keep creator and publisher whole
  f['Publisher:'+entry['publisher']]=1
```

```
IDLE 2.6.2
>>> import docclass
>>> import feedfilter
>>> cl=docclass.fisherclassifier(feedfilter.entryfeatures)
>>> cl.setdb('python_feed.db')
>>> feedfilter.read('python_search.xml',cl)

-----
Title:     My new baby boy!
Publisher: Shetan Noir, the zombie belly dancer! - MySpace Blog

THis is my new baby, Anthem. He is a 3 and half month old ball <b>python</b>, orange shaded normal pattern. I have held him about 5 time since I brought him hom
ht at 8:00pm...
Guess: None
Enter category: snake

-----
Title:     If you need a laugh...
Publisher: Kate&#39;s space

Even does 'funny walks' from Monty <b>Python</b>. He talks about all the ol' ladies that are after him. He teases me about my horror obsession. He attempts suic
d best of all, he talks about poo. Who doesn't think poo is funny???!
Guess: snake
Enter category: monty

-----
Title:     And another one checked off the list..New pix comment ppl
Publisher: Python Guru - MySpace Blog

Now the one of a kind NERD bred Carplot male is in our possesion. His name is Broken (not because he is sterile) lol But check out the pic and leave one bitches
.......
Guess: snake
Enter category: snake
```

```
Title:     And another one checked off the list..New pix comment ppl
Publisher: Python Guru - MySpace Blog

Now the one of a kind NERD bred Carplot male is in our possesion. His name is Broken (not because he is sterile) lol But check out the pic and leave one bitches
.......
Guess: snake
Enter category: snake

-----
Title:     <b>Python</b> vs Java - It欹椸 not only the language, it欹椸 the tools
Publisher: A Drop In The Stream

Well, I欹椷e done a bit of <b>Python</b> coding by now in my new job and here欹椸 my take on the <b>Python</b> vs. Java question. <b>Python</b> is concise and ju
s for doing tricky, complicated things using simple semantics. I just learned of a chunk of <b>...</b>
Guess: snake
Enter category: python

-----
Title:     BF 2142 ads successfully hacked - add your own and/or block the <b>...</b>
Publisher: Aaron Tiensivu&#39;s Blog

One <b>Python</b> script exports the graphics and a different one imports the graphics. I'm not posting the source code/scripts because I don't know who to cred
he source, and I don't want any weird DMCA lawsuits sent my way for posting <b>...</b>
Guess: snake
Enter category: python

-----
Title:     Ruby, <b>Python</b>, JavaScript, Perl, C++
Publisher: 祜禊芯谐禊邪屑 啸萤谐卸爷爷谢禅亵 孝械亵薪懈泻, 孝械亵薪芯谢芯谐褊褊写

Ruby (String), <b>Python</b> (str), JavaScript (String), Perl, C++  (std::string). s = "abc", s = "abc", s = "abc", $s = "abc", string s = "abc". s = x + y, s =
s = x + y, $s = $x . $y, s = x + y*1 <b>...</b>
Guess: python
Enter category: python
```

Akismetを利用する

```
Python 2.6.2 (r262:71605, Apr 14 2009, 22:40:02) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.

    ****************************************************************
    Personal firewall software may warn about the connection IDLE
    makes to its subprocess using this computer's internal loopback
    interface.  This connection is not visible on any external
    interface and no data is sent to or received from the Internet.
    ****************************************************************

IDLE 2.6.2
>>> import akismettest
>>> msg='Make money fast! Online Casino!'
>>> akismettest.isspam(msg,'spammer@spam.com','127.0.0.1')

Traceback (most recent call last):
  File "<pyshell#2>", line 1, in <module>
    akismettest.isspam(msg,'spammer@spam.com','127.0.0.1')
  File "C:\Python26\akismettest.py", line 8, in isspam
    valid = akismet.verify_key(apikey,pageurl)
AttributeError: 'module' object has no attribute 'verify_key'
>>>
```

akismettest.py

```
import akismet
defaultkey = "SOME KEY"
pageurl="http://linhuarinka.wordpress.com/wp-
   admin/"
defaultagent="Mozilla/5.0 (Windows; U; Windows
   NT 5.1; euc-jp; rv:1.8.0.7) "
defaultagent+="Gecko/20060909 Firefox/3.0.10"
def
   isspam(comment,author,ipaddress,agent=defaul
   tagent,apikey=defaultkey):
```