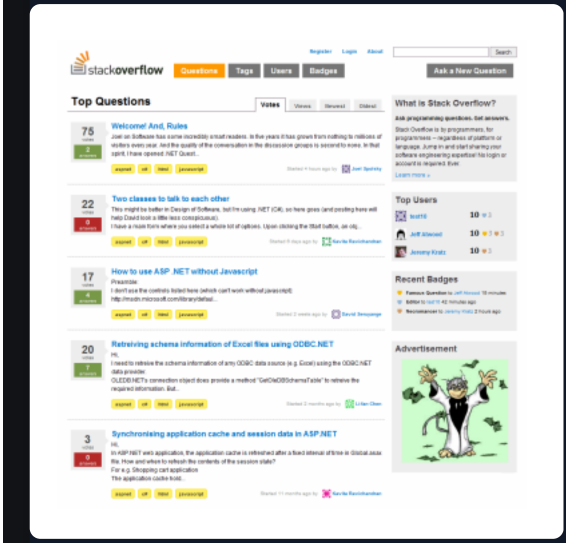
Stack Exchange Question Classifier

Points: 98.53 Rank: 21158

Problem | Submissions | Leaderboard | Editorial

Stack Exchange is an information powerhouse, built on the power of crowdsourcing. It has 105 different topics and each topic has a library of questions which have been asked and answered by knowledgeable members of the StackExchange community. The topics are as diverse as travel, cooking, programming, engineering and photography.



Top Questions

Votes Views Newest Oldest

75 votes 2 answers — Welcome! And, Rules
Joel on Software has some incredibly smart readers. In five years it has grown from nothing to millions of visitors every year. And the quality of the conversation in the discussion groups is second to none. In that spirit, I have opened .NET Quest...
aspnet c# html javascript — Started 4 hours ago by Joel Spolsky

22 votes 0 answers — Two classes to talk to each other
This might be better in Design of Software, but I'm using .NET (C#), so here goes (and posting here will help David look a little less conspicuous).
I have a main form where you select a whole lot of options. Upon clicking the Start button, an obj...
aspnet c# html javascript — Started 5 days ago by Kavita Ravichandran

17 votes 4 answers — How to use ASP .NET without Javascript
Preamble:
I don't use the controls listed here (which can't work without javascript)
http://msdn.microsoft.com/library/defaul...
aspnet c# html javascript — Started 2 weeks ago by David Senyange

20 votes 7 answers — Retreiving schema information of Excel files using ODBC.NET
Hi,
I need to retreive the schema information of any ODBC data source (e.g. Excel) using the ODBC.NET data provider.
OLEDB.NET's connection object does provide a method "GetOleDBSchemaTable" to retreive the required information. But...
aspnet c# html javascript — Started 2 months ago by Li Ian Chen

3 votes 0 answers — Synchronising application cache and session data in ASP.NET
Hi,
In ASP.NET web application, the application cache is refreshed after a fixed interval of time in Global.asax file. How and when to refresh the contents of the session state?
For e.g. Shopping cart application
The application cache hold...
aspnet c# html javascript — Started 11 months ago by Kavita Ravichandran

What is Stack Overflow?

Ask programming questions. Get answers.

Stack Overflow is by programmers, for programmers -- regardless of platform or language. Jump in and start sharing your software engineering expertise! No login or account is required. Ever.

Learn more »

Top Users

scott6	10 ● 3
Jeff Atwood	10 ● 3 ● 3
Jeremy Kratz	10 ● 3

Recent Badges

- Famous Question to Jeff Atwood 18 minutes
- Editor to [illegible] 42 minutes ago
- Necromancer to Jeremy Kratz 2 hours ago

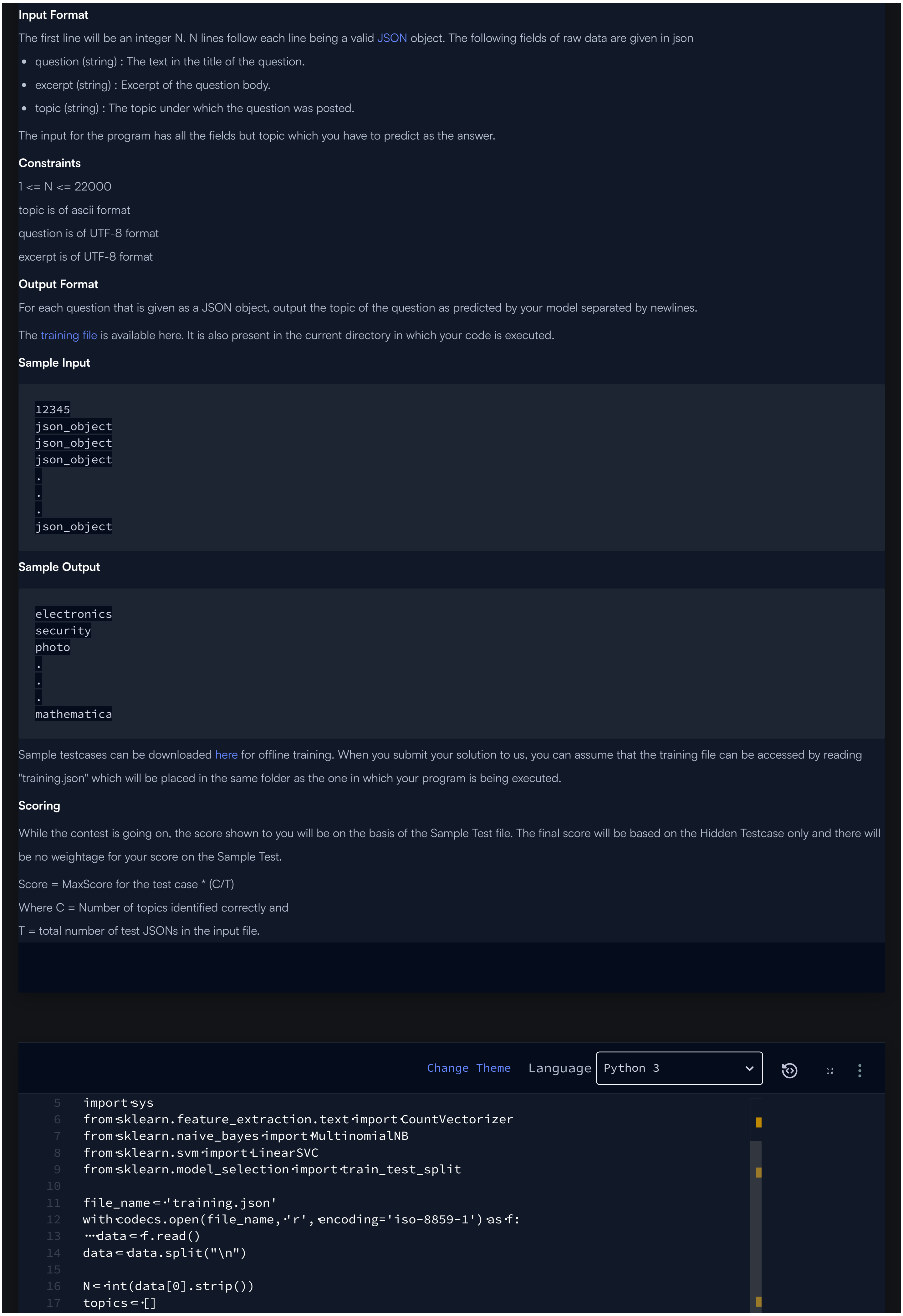
Advertisement

We have hand-picked ten different topics (such as Electronics, Mathematics, Photography etc.) from Stack Exchange, and we provide you with a set of questions from these topics.

Given a question and an excerpt, your task is to identify which among the 10 topics it belongs to.

Getting started with text classification

For those getting started with this fascinating domain of text classification, here's a wonderful Youtube video of Professor Dan Jurafsky from Stanford, explaining the Naive Bayes classification algorithm, which you could consider using as a starting point.



Input Format

The first line will be an integer N. N lines follow each line being a valid JSON object. The following fields of raw data are given in json

- question (string) : The text in the title of the question.
- excerpt (string) : Excerpt of the question body.
- topic (string) : The topic under which the question was posted.

The input for the program has all the fields but topic which you have to predict as the answer.

Constraints

1 <= N <= 22000

topic is of ascii format

question is of UTF-8 format

excerpt is of UTF-8 format

Output Format

For each question that is given as a JSON object, output the topic of the question as predicted by your model separated by newlines.

The training file is available here. It is also present in the current directory in which your code is executed.

Sample Input

```
12345
json_object
json_object
json_object
.
.
.
json_object
```

Sample Output

```
electronics
security
photo
.
.
.
mathematica
```

Sample testcases can be downloaded here for offline training. When you submit your solution to us, you can assume that the training file can be accessed by reading "training.json" which will be placed in the same folder as the one in which your program is being executed.

Scoring

While the contest is going on, the score shown to you will be on the basis of the Sample Test file. The final score will be based on the Hidden Testcase only and there will be no weightage for your score on the Sample Test.

Score = MaxScore for the test case * (C/T)

Where C = Number of topics identified correctly and

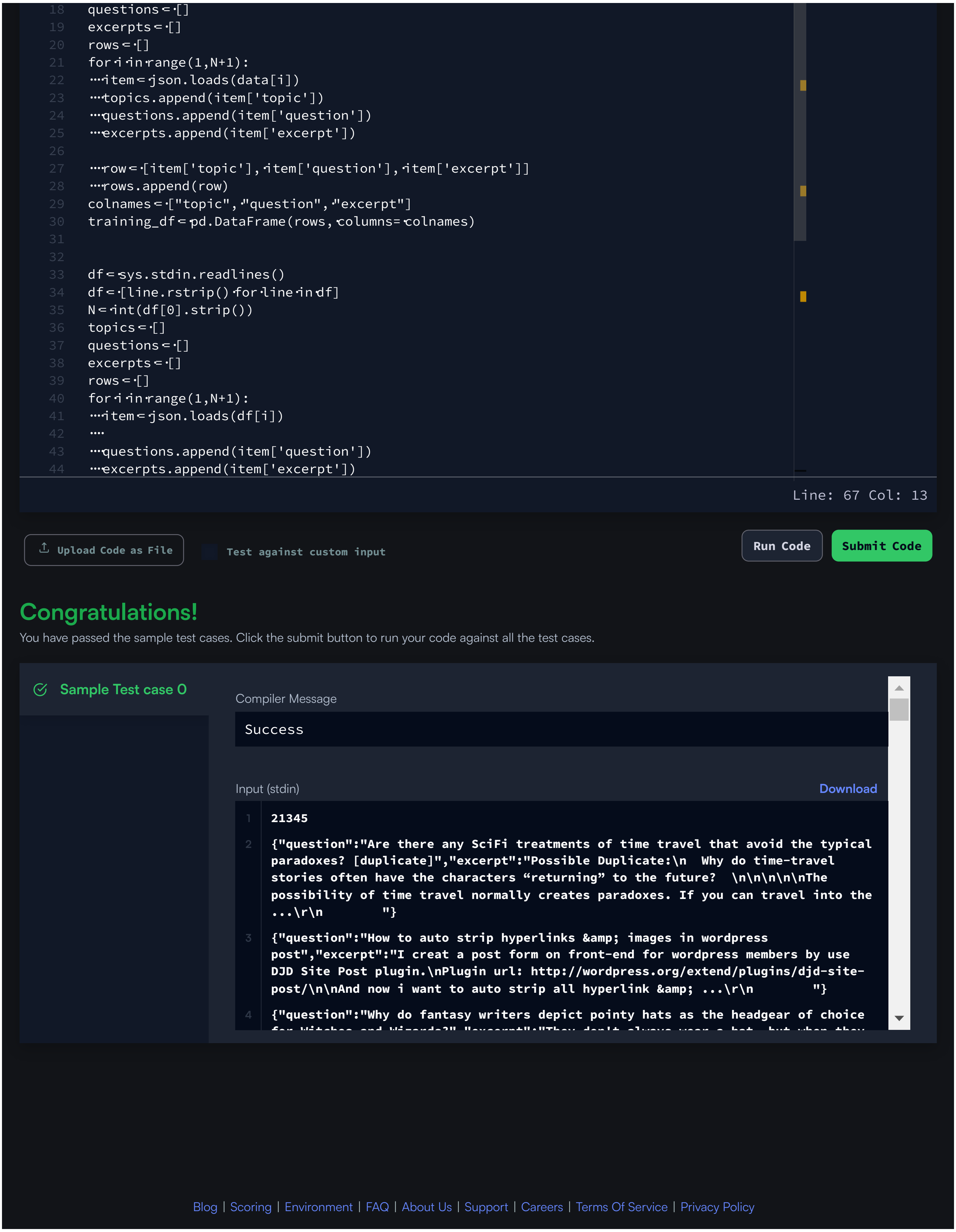
T = total number of test JSONs in the input file.

Language Python 3

```
import sys
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn.model_selection import train_test_split

file_name = 'training.json'
with codecs.open(file_name, 'r', encoding='iso-8859-1') as f:
    data = f.read()
data = data.split("\n")

N = int(data[0].strip())
topics = []
```



```
questions = []
excerpts = []
rows = []
for i in range(1,N+1):
    item = json.loads(data[i])
    topics.append(item['topic'])
    questions.append(item['question'])
    excerpts.append(item['excerpt'])

    row = [item['topic'], item['question'], item['excerpt']]
    rows.append(row)
colnames = ["topic", "question", "excerpt"]
training_df = pd.DataFrame(rows, columns= colnames)


df = sys.stdin.readlines()
df = [line.rstrip() for line in df]
N = int(df[0].strip())
topics = []
questions = []
excerpts = []
rows = []
for i in range(1,N+1):
    item = json.loads(df[i])
    
    questions.append(item['question'])
    excerpts.append(item['excerpt'])
```

Line: 67 Col: 13

Congratulations!

You have passed the sample test cases. Click the submit button to run your code against all the test cases.

Compiler Message

```
Success
```

Input (stdin)

```
21345
{"question":"Are there any SciFi treatments of time travel that avoid the typical paradoxes? [duplicate]","excerpt":"Possible Duplicate:\n  Why do time-travel stories often have the characters “returning” to the future?  \n\n\n\n\nThe possibility of time travel normally creates paradoxes. If you can travel into the ...\r\n        "}
{"question":"How to auto strip hyperlinks &amp; images in wordpress post","excerpt":"I creat a post form on front-end for wordpress members by use DJD Site Post plugin.\nPlugin url: http://wordpress.org/extend/plugins/djd-site-post/\n\nAnd now i want to auto strip all hyperlink &amp; ...\r\n        "}
{"question":"Why do fantasy writers depict pointy hats as the headgear of choice
```

Blog | Scoring | Environment | FAQ | About Us | Support | Careers | Terms Of Service | Privacy Policy