

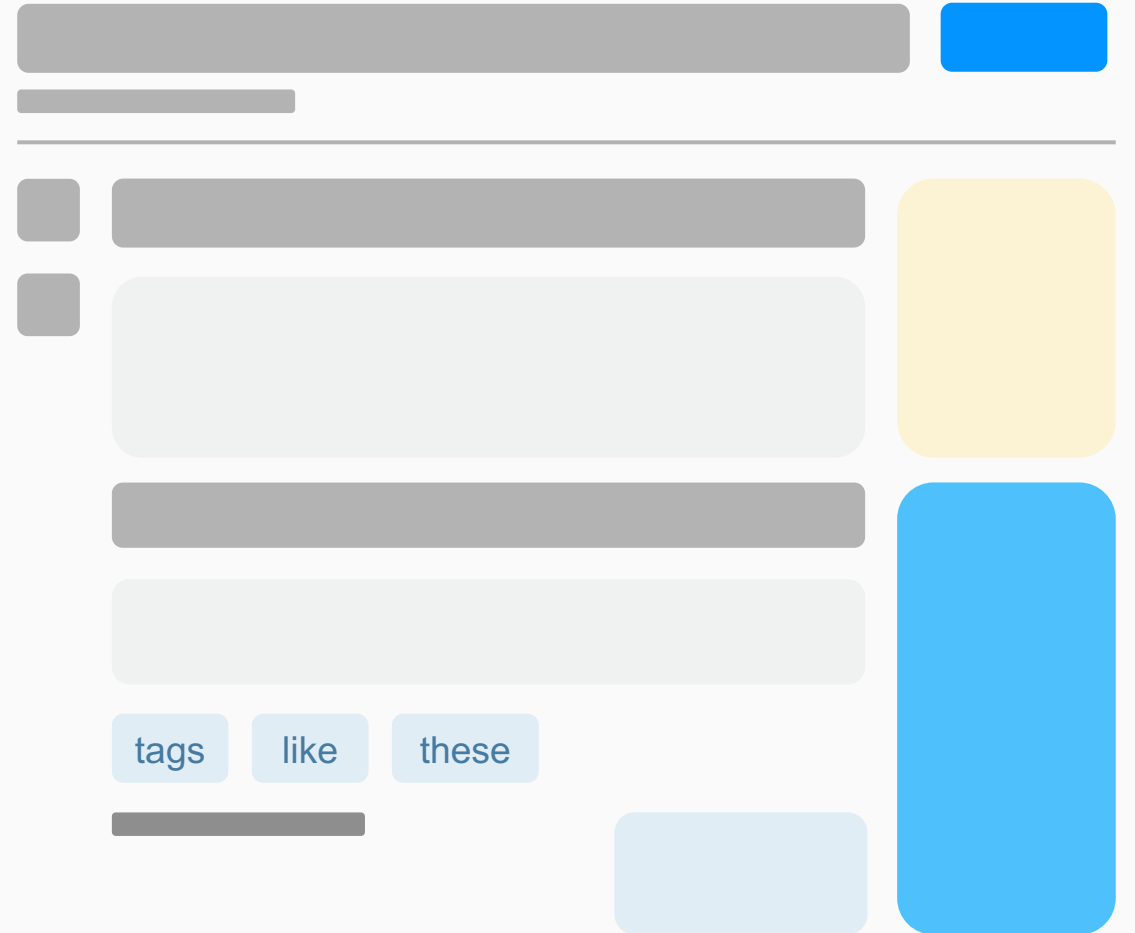
Stackoverflow Auto-Tagger

2020.6.18

김재우, 김태윤, 이준형, 전선영

Problem Statement

Design an algorithm that helps users of StackOverflow in the cumbersome process of manually attaching tags to their questions.



Technical Approach

Prepare corpus using StackOverflow API



Extract important words using tf-idf



Determine important sentences



Evaluate the performance against real tags



Provide suggestions from popular tags



Visualize the results

Corpus

id	title	link	content	tags	last_activity	answer_count
61573446	what's wrong in my	https://stacko	[<div class="post-text" itemprop="text"> Hey all I am new in android, and I have made a simple BMi calculator in this I have a <code>spinner</code> to select a value, and I have to make a calculator 	['android', 'kotlin', 'android-spinner']	1588506438	0
61573914	Flutter- how can i know i	https://stacko	[<div class="post-text" itemprop="text"> <p>i'm writing a flutter application and i want to know if the device is now connecte to a bluetooth device or no and i mean connected not associated Thank You </p>	['android', 'ios', 'swift', 'flutter', 'dart']	1588506437	0
61572274	Denormalization practice	https://stacko	[<div class="post-text" itemprop="text"> <p>I am creating a reactive application with Meteor (with MongoDB as a backend).</p> <p>I initially created a non-reactive-aware collection and denormalizers, eg.:</p> <pre><code>class DocCollection extends Mongo.Collection {	['mongodb', 'meteor', 'rxjs', 'observable']	1588506430	0
61571901	Delete Duplicate Data on	https://stacko	[<div class="post-text" itemprop="text"> <p>How to delete duplicate data on 1 table which have kind data like these, enter image description here	['postgresql', 'postgresql-9.3']	1588506427	0
61573913	ARQuickLookPreview for	https://stacko	[<div class="post-text" itemprop="text"> <p>I am trying to move the AR assets (.reality) files in my app to being On Demand. I created a resource tag and have created a manager to download them on demand. It seems like it is downloading the assets fine since I can see it in the XCode. I use a	['ios', 'swift', 'quicklook']	1588506426	0
61573912	(Python) deploying sqlalchemy	https://stacko	[<div class="post-text" itemprop="text"> <p>I was struggling with the following exception and have found a solution, but I don't really understand why the solution works. Anyway, I am posting my code and the solution in case it helps anyone else. If anyone understands exactly why this fixes the issue please share your	['python', 'sqlalchemy', 'cx-freeze']	1588506422	0
61573911	Order an array using swa	https://stacko	[<div class="post-text" itemprop="text"> <p>Given an initial array/list and a target array, the goal is to find the smallest list of movements that would transform the initial array into the target one. The movements that can be used are: </p>	['algorithm', 'sorting']	1588506422	0
61570612	CSS sidebar with flexbox	https://stacko	[<div class="post-text" itemprop="text"> <p>I'm trying to implement a sidebar with static width and flexbox. I have two issues:</p> When the sidebar is closed, there is still space, and I don't understand where it comes	['html', 'css']	1588506416	2

Extracting Important Words

TF-IDF (term frequency, inverse document frequency)

Measures the significance of a word in a document in terms of the amount of information it contains.

$$\text{tf-idf}(t, d) = \text{tf}(t, d) * \text{idf}(t)$$

where $\text{tf}(t, d) = \text{count of } t \text{ in } d / \text{number of words in } d$

$\text{df}(t) = \text{occurrence of } t \text{ in documents}$

$$\text{idf}(t) = \log(N / (\text{df}(t) + 1))$$

Measured the degree of grammatical dependency of a word in a sentence using ShiftReduceParser

Scored each word based on the above two criteria

Optionally, consider the length and position of sentences the words are contained in

Evaluating Performance

Compare predicted tags with actual tags with two parameters

STRICT: ignore tags that are different but have same lemma

e.g. 'python2' and 'python3'

INCLUSIVE: count tags that are part of another

e.g. 'react' and 'react-native'

Check and suggest if predictions are one of popular tags of StackOverflow:

('javascript', 2030742), ('java', 1684580), ('python', 1461591), ('c#', 1414574), ('php', 1356698), ('android', 1281286), ('html', 1000331), ('jquery', 989318), ('c++', 677591), ('css', 668868), ('ios', 632876), ('mysql', 599509), ('sql', 547382), ('asp.net', 356558), ('r', 350335), ('node.js', 333901), ('c', 333280), ('arrays', 332805), ('ruby-on-rails', 318341), ('.net', 297755), ('json', 293988), ('objective-c', 290115), ('sql-server', 286924), ('swift', 269448), ('angularjs', 260707), ('python-3.x', 235231), ('django', 234845), ('regex', 227935), ('excel', 224195), ('reactjs', 221150), ('iphone', 220455), ('angular', 220405) ...

An Example

Home

PUBLIC

Stack Overflow

Tags

Users

Jobs

TEAMS

What's this?

Free 30 Day Trial

How to include user_id (foreignKey) when posting an new record?

Asked 1 month ago · Active 1 month ago · Viewed 15 times

0

I'm new to python Django rest-framework, I'm facing this problem when creating a new address:

```
null value in column "user_id" violates not-null constraint
DETAIL:  Failing row contains (21, full name, 123456789, any, any, any, any, a
```

This is The address model:

```
from django.db import models
from django.contrib.auth.models import User

class UserAddress (models.Model):
    user = models.ForeignKey(
        User, related_name='addresses', on_delete=models.CASCADE)
    full_name = models.TextField(default='')
    phone = models.CharField(max_length=30, default='')
    city = models.TextField()
    province = models.TextField()
    street = models.TextField()
    description = models.TextField()
    postcode = models.CharField(max_length=20)
    country = models.CharField(max_length=100)
    is_default = models.BooleanField(default=True)

class Meta:
    db_table = 'user_addresses'
```

And this is the serializer:

```
from rest_framework import serializers
from user_action.models.address import UserAddress

class UserAddressSerializer(serializers.ModelSerializer):
    id = serializers.IntegerField(read_only=True)
    class Meta:
        model = UserAddress
        fields = ['id', 'full_name', 'phone', 'city', 'province',
                  'street', 'description', 'postcode', 'country', 'is_default']
```

And the POST method:

```
@api_view(['POST'])
@permission_classes((IsAuthenticated,))
def createUserAddress(request):
    user = request.user
    if request.method == 'POST':
        serializer = UserAddressSerializer(data=request.data)
        if serializer.is_valid():
            newAddress = serializer.save()
        else:
            return Response(serializer.errors)
    return Response(serializer.data)
```

Thanks in advance.

django · django-models · django-rest-framework · django-serializer

The Overflow Blog

Talking TypeScript with the engineer who leads the team

Podcast 244: Dropping some knowledge on Drupal with Dries

Featured on Meta

We're switching to CommonMark

New post lock available on meta sites: Policy Lock

What more can be done to prevent questions with just external links/images?

Now that the Edit Question button for closed questions is more prominent,...

Looking for a job?

Full-Stack Engineer

Octane AI · No office location

REMOTE

python · javascript

Senior Android Developer (Kotlin)

Mindvalley · Kuala Lumpur, Malaysia

\$30K - \$45K · REMOTE

android · kotlin

Senior Android/Kotlin Developer (Remote)

X-Team · No office location

REMOTE

android · ionic-framework

iOS Software Engineer

Zuhike Engineering Ltd · London, UK

£45K - £85K

android · swift

High response rate

Machine Learning Engineer - Remote

Numbrs · No office location

REMOTE

java · scala

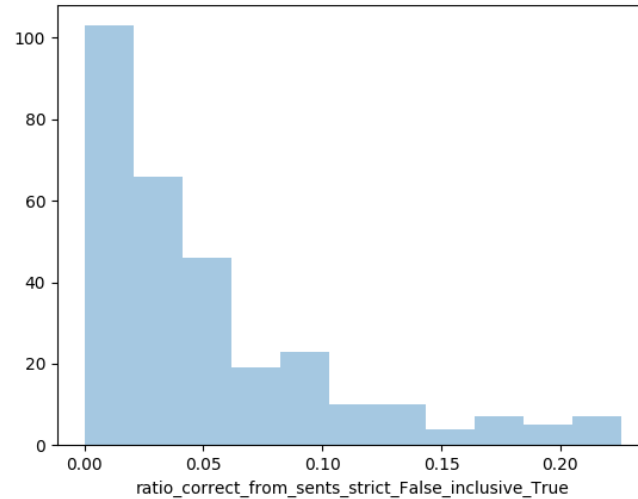
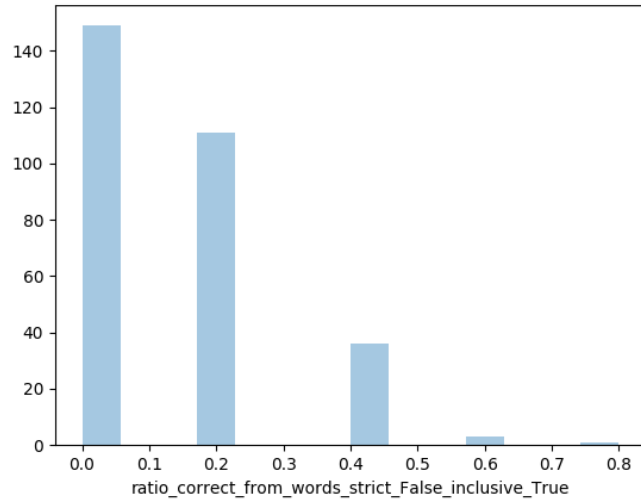
Real tags:

django, django-models, django-rest-framework, django-serializer

Predicted tags:

django, serialiser, serializer, address, framework

Results



From words	Strict	Not Strict	From sentences	Strict	Not Strict
Inclusive	0.109	0.131	Inclusive	0.031	0.049
Non-inclusive	0.077	0.090	Non-inclusive	0.018	0.030

Results



Conclusion

Considering important sentences had no apparent benefit in finding tag words.

Predicted tags that match actual tags are usually proper nouns, such as SW/HW products and packages etc.

e.g.) mongodb, django, pytorch, android, python, etc.

Predicted tags that do not match with actual tags are usually more abstract words that are used widely across various fields.

e.g.) duplicate, iterate, selection, dataframe, function, etc

This algorithm is most useful when you want to add such abstract tags.

Thank you

github.com/tykimseoul/CS372_NLP_Project