

معرفی مسأله

در چند سال گذشته، انجمن‌های پاسخگویی به سؤالات آنلاین (Q&A) مانند Stack Overflow به مخازن ضروری دانش برای مهندسان نرم‌افزار تبدیل شده‌اند. از آنجایی که اطلاعات جمع‌آوری شده در پورتال به سؤالات حرفه‌ای که از متخصصان یا برنامه‌نویسان سرگرم‌کننده سرچشمه می‌گیرد، مرتبط است، این پست‌ها یک پشتیبانی مؤثر و عملاً کاربردی برای افراد حرفه‌ای مبتدی و با تجربه ارائه می‌کنند. توسعه دهندگان اغلب از سایت‌های پرسش و پاسخ اجتماعی مانند Stack Overflow برای حل چالش‌های برنامه نویسی استفاده می‌کنند. هر روز بیش از ۶۰۰۰ سوال جدید به Stack Overflow ارسال می‌شود و تقریباً ۱۰ میلیون کاربر. کاربران از مبتدی تا ماهر در تبادل دانش سازنده در این سایت شرکت می‌کنند و یک جامعه برنامه نویسی پویا را تشکیل می‌دهند. هرکسی می‌تواند برای رفع مشکلات خود درباره موضوعات مختلف سؤال بپرسد و سایر کاربران می‌توانند پاسخ دهند یا نظرات خود را در مورد آن ارائه دهند. برای اینکه این رویه کاربر پسندتر شود، Stack Overflow چندین گزینه فیلتر و ترجیحی مانند Bountied, Interesting, Watched List و Ignored Tags برای پیشنهاد موارد مناسب ارائه می‌دهد.

با این حال، حدود ۱۶ روز طول می‌کشد تا پاسخ دریافت شود در حالی که انحراف استاندارد تا ۱۱۳ روز متغیر است. علاوه بر این، ۳۰٪ از کل سؤالات بدون پاسخ باقی می‌مانند، که مانع از کارایی Stack Overflow می‌شود. بسیاری از محققان به این دغدغه کشیده شده‌اند و از زوایای بسیاری به آن پرداخته‌اند. این چالش هنوز حل نشده است و محققان اکنون با ارزیابی دشواری یک سوال برای حل این مشکل از منظر دیگری به آن نگاه می‌کنند. اکنون، با در نظر گرفتن ویژگی‌های متنی کاربران و پست‌ها، به تخمین دشواری پست‌های Stack Overflow نزدیک شدیم. هدف این پروژه پیشنهاد و ارزیابی مدل‌هایی برای طبقه‌بندی بر اساس سختی سوال است. موضوعات انجمن‌های زیر برای هر گروه قابل انتخاب است (ترجیحا موضوعاتی را انتخاب کنید که تعداد سوال کافی داشته باشند):

- ✓ جاوا
- ✓ سی شارپ
- ✓ پایتون
- ✓ اندروید
- ✓ iOS
- ✓ جاوا اسکریپت
- ✓ Html
- ✓ Php
- ✓ C++
- ✓ SQL
- ✓ CSS
- ✓ Excel

سؤالات پروژه

پرسش ۱: کدام مدل برای تعریف سطح دشواری سؤال به خوبی عمل می‌کند؟

پرسش ۲: ویژگی‌های مختلف چگونه با سطح دشواری سؤال مرتبط هستند؟

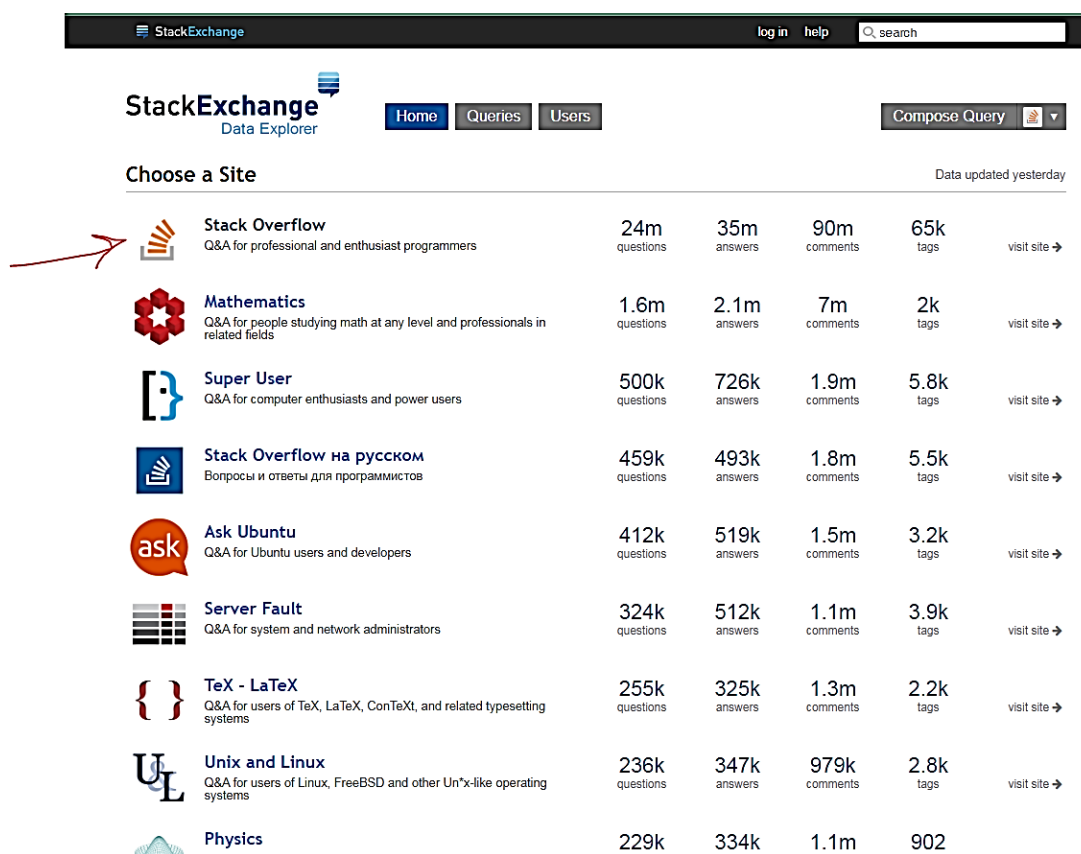
فاز اول










برای دستیابی به هدف خود مراحل زیر پیشنهاد می‌شود:

- ✓ ابتدا ۱۰۰۰ سؤال را که مربوط به سال ۲۰۱۸ تا ۲۰۲۲ باشند جمع‌آوری کنید (فاز اول).
- ✓ برچسب‌گذاری را به صورت سه کلاسه یعنی مقدماتی، متوسط و پیشرفته براساس روشی که در ادامه توضیح داده خواهد شد انجام دهید. سوالات به طور کامل توسط هر یک از اعضای گروه برچسب گذاری شوند و دلیل طبقه بندی پست در یک کلاس خاص مشخص باشد. سپس از روش رای گیری اکثریت (بین سه نفر) برای برچسب نهایی پیروی شود. نهایتاً اعتبار سنجی برچسب زدن دستی توسط استاد انجام می‌شود (فاز اول).
- ✓ پس از فرآیند برچسب‌گذاری، سه مجموعه ویژگی برای تکمیل مجموعه داده استخراج شود (فاز دوم).
- ✓ از سه مدل یادگیری تحت نظارت، Tf-Idf، Topic Modeling و Doc2Vec را با طبقه‌بندی‌کننده‌های مختلف شامل درخت تصمیم، نایو بیس، KNN، شبکه عصبی پیاده‌سازی کنید تا مناسب‌ترین مدل را برای داده‌های Stack Overflow پیدا نمایید و آنها را با معیارهای دقت، صحت، یادآوری و امتیاز f1 ارزیابی نمایید (فاز سوم).

روش برچسب‌گذاری

از ابزار Data Explorer در سایت Stack Exchange برای دریافت داده‌ها استفاده کنید.



StackExchange Data Explorer		log in	help	search
StackExchange Data Explorer		Home	Queries	Users
Choose a Site		Data updated yesterday		
	Stack Overflow Q&A for professional and enthusiast programmers	24m questions	35m answers	90m comments
	Mathematics Q&A for people studying math at any level and professionals in related fields	1.6m questions	2.1m answers	7m comments
	Super User Q&A for computer enthusiasts and power users	500k questions	726k answers	1.9m comments
	Stack Overflow на русском Вопросы и ответы для программистов	459k questions	493k answers	1.8m comments
	Ask Ubuntu Q&A for Ubuntu users and developers	412k questions	519k answers	1.5m comments
	Server Fault Q&A for system and network administrators	324k questions	512k answers	1.1m comments
	TeX - LaTeX Q&A for users of TeX, LaTeX, ConTeXt, and related typesetting systems	255k questions	325k answers	1.3m comments
	Unix and Linux Q&A for users of Linux, FreeBSD and other Un*x-like operating systems	236k questions	347k answers	979k comments
	Physics Q&A for practicing researchers, academics and students of physics	229k questions	334k answers	1.1m comments

برای دریافت موارد زیر کوئری لازم براساس موضوع گروه خود را بنویسید. نمونه ای از کوئری برای استخراج برچسب های پرتعداد در تصویر زیر آورده شده است.

- ✓ انتخاب ۱۰۰۰ ردیف دارای شناسه پست، عنوان پست، متن پست و تگ های پست برای برچسب گذاری.

- ✓ پس از اینکه اولین مجموعه برای برچسب‌گذاری دستی ساخته شد، با تجزیه و تحلیل بدنه پست، داده‌ها را به صورت دستی برچسب‌گذاری کنید.
- دسته‌هایی که باید در هنگام برچسب‌گذاری در نظر گرفته شوند، پایه، متوسط و پیشرفته هستند. و کل فرآیند برچسب زدن توسط هر عضو گروه به صورت جدا انجام شود (هر نفر یک فایل اکسل)
- ✓ برای برچسب‌گذاری از مجموعه قوانین جدول زیر استفاده کنید که در آن هر یک از قوانین را به قطعات تقسیم کردیم، یعنی قوانین زیربخش هر کدام به سؤالاتی مرتبط هستند که در مجموعه داده گنجانده شده است.

Enter a title for your query

stackoverflow

Q&A for professional and enthusiast programmers

edit description

```

1  -- Most popular StackOverflow tags in May 2010
2
3  select
4      num.TagName as Tag,
5      row_number() over (order by rate.Rate desc) as MayRank,
6      row_number() over (order by num.Num desc) as TotalRank,
7      rate.Rate as QuestionsInMay,
8      num.Num as QuestionsTotal
9
10 from
11
12 (select count(PostId) as Rate, TagName
13  from
14      Tags, PostTags, Posts
15  where Tags.Id = PostTags.TagId and Posts.Id = PostId
16  and Posts.CreationDate < '2020-06-01'
17  and Posts.CreationDate > '2019-05-01'
18  group by TagName) as rate
19
20 INNER JOIN
21
22 (select count(PostId) as Num, TagName
23  from
24      Tags, PostTags, Posts
25  where Tags.Id = PostTags.TagId and Posts.Id = PostId
26  group by TagName
27  having count(PostId) > 800)
28 as num ON rate.TagName = num.TagName
29 order by rate.rate desc
30 :

```

Database Schema

Posts

Id	int
PostTypeId	tinyint
AcceptedAnswerId	int
ParentId	int
CreationDate	datetime
DeletionDate	datetime
Score	int
ViewCount	int
Body	nvarchar (max)

Revisions

Waiting for you to make your first edit...

Run Query

Cancel

Options:

☐ Text-only results
 ☐ Include execution plan

hide sidebar >>

Difficulty Class	General Rule set	Granular Breakdown
Basic	Questions on simple built-in functions/API documentation/beginner level knowledge	1 Simple Built-in-function
		2 Simple Operator
		3 API documentation
		4 Beginner level Theory Question
		5 Basic OOP problem
		6 Simple Program Understanding
	Questions related to comparison between functions of various languages	7 Analysis of various languages' functions
		8 Beginner level query difference
		9 Simple problem solving
	Questions with simple problem-solving	10 Simple query problem solving
		11 Simple functionality related
	Questions with simple exception, error and other problem	12 Solve for nullpointer exception
		13 Simple Error Handling
		14 Simple configuration problem

Intermediate	Questions demanding deeper understanding of the programming language to answer	15	Built-in function deep understanding
		16	Need more knowledge about the algorithms
		17	Multiple questions
		18	Need knowledge on Advanced Programming topics
		19	Difference between two packages
	Questions stating the answer of the problem but still inquires about more efficient answer	20	Looking for Appropriate way
		21	Analyzing different alternatives
	Questions about a system's computational cost, space utilization, or other resource usages	22	Efficient way
		23	Performance, optimization, accuracy
		24	Memory related
	Questions requiring conceptual thinking in response to any programming structure, API, or design principle	25	Reverse programming
		26	Underlying philosophy of any programming construction
		27	Design pattern
		28	Feasibility study
		29	Question about built-in documentation in details
	Required Testing Related Knowledge	30	Automated testing related problem
		31	Requires knowledge of Testing
Advanced	Questions about critical challenges that require in-depth technical expertise or logical reasoning to solve	32	Critical problems where solution needs in-depth programming knowledge or conceptual thinking.
		33	Multiple question, Solution needs in-depth programming knowledge or logical thinking.
		34	Multiple question and in depth knowledge needed
	Questions that require advanced in-depth knowledge of internal language structure	35	In-depth knowledge of internal language structure.
		36	In-depth knowledge of packages
	Questions that deals with infrequently/rarely used framework/API	37	Deals with infrequently/rarely used framework
		38	Deals with deprecated framework
	Related to real life scenario	39	Efficiency related Question in real life scenario
		40	Optimization in Real life scenario
	Other Rules	41	Mentioned having the answer, asking for suggestion
		42	In-depth knowledge on Garbage Collection Algorithm
		43	In-depth testing and security knowledge
		44	Need in-depth knowledge multiple topics
		45	Large amount of study already known
		46	Need deep knowledge about design architecture , SW maintenance, and SDLC,new plugin
		47	Works on large dataset, artificial intelligence

✓ هر نفر برای فایل اکسل خود توزیع آماری سه کلاس را مشخص کند (هر کلاس چند ردیف دارد؟)

فاز دوم

ویژگی‌های نهایی شامل موارد موجود در جدول زیر را با استفاده از کوئری‌های مناسب یا نوشتن کدهای لازم ایجاد نمایید.

Feature Name	Defination	Included in
Processed Body	Post full textual body excluding code snippets and the html tags like <p>, <code>,<href>	Semantic Features,Pre-hoc,Post-hoc
Tags	Post tags, decided at the time of posting, e.g. <java> <oop><multithread>	Semantic Features,Pre-hoc,Post-hoc
Title	Post title, decided by questioner	Semantic Features,Pre-hoc,Post-hoc
Question Length	Length of the whole Processed Body	Semantic Features,Pre-hoc,Post-hoc
Url+Image_Count	Number links the post	Semantic Features,Pre-hoc,Post-hoc
LOC	Line of Code, counting only physical lines of source code in snippet extracted from post body Summing up all the LOCs from a certain post	Semantic Features,Pre-hoc,Post-hoc
User Reputation	User Reputation Point given by Stack Overflow activities like answering, questioning	Pre-hoc,Post-hoc
User_Bronze_Badge	Number of awards for basic use of the site	Pre-hoc,Post-hoc
User_Gold_Badge	Number of awards for important contributions from members of the community	Pre-hoc,Post-hoc
User_Silver_Badge	Number of awards for being experienced users who regularly use Stack Overflow	Pre-hoc,Post-hoc
Accept Rate	The percentage of answers accepted based on the questions asked by the user.	Pre-hoc,Post-hoc
View Count	Number of time viewed by users	Post-hoc
Favorite_Count	Number of times save as favorite	Post-hoc
Up_vote_Count	Number of up votes for being useful and appropriate	Post-hoc
Answer Count	Number of answers in a question thread	Post-hoc
Question_Score	The total number of upvotes it received minus the total number of downvotes it received	Post-hoc
First_Answer_Interval	Interval in days between question creation date to first answer creation	Post-hoc
Accepted_Answer_Interval	Interval in days between question creation date to accepted answer creation	Post-hoc

ستون آخر جدول نشان دهنده گروه یا نوع ویژگی است.

فاز سوم (ساخت مدل)

- برای ویژگی‌های متن Title و Body و Tags باید متن‌ها را به ویژگی تبدیل کنید که برای این کار باید ابتدا این سه مورد را با هم یکی کرده یعنی آن‌ها را به هم بچسبانید، سپس مراحل زیر را انجام دهید:
- پیش پردازش متن شامل:
 - حذف stop word ها
 - ریشه یابی
 - تصحیح نگارشی
- سپس متن‌ها را به بردارهای عددی تبدیل کنید. می‌توانید از Vectorizer های مختلف موجود در پایتون استفاده کنید. همچنین ویژگی‌های n-gram و tfidf را نیز به مجموعه داده اضافه کنید.
- مجموعه داده را به مجموعه‌های آموزش و اعتبار سنجی با استفاده از متد موجود train-test-split تقسیم کنید.
- در انتهای این فاز مدل‌های یادگیری ماشین زیر را روی مجموعه train آموزش داده سپس روی مجموعه داده ارزیابی آزمایش کنید.
 - Naïve Bayes
 - SVC
 - CART
 - Logistic Regression
 - MLP
 - XGBoost
- دقت، صحت، معیار F1، ماتریس Confusion و نمودارهای لازم برای مقایسه مدل‌ها را آماده کرده و در فاز بعد در تهیه مستندات از آن‌ها استفاده کنید.

فاز چهارم (مستندسازی)

نوشتن داکيومنت با فرمت گزارش پایانی کارشناسی.

نمره اضافی (بهبود)

- در حالی که چارچوب فوق را می‌توان برای مسائل طبقه بندی متن اعمال کرد، اما برای دستیابی به یک دقت خوب می‌توان بهبودهایی را در چارچوب کلی انجام داد. به عنوان مثال، نکات زیر برای بهبود عملکرد مدل‌های طبقه بندی متن و این چارچوب ارائه شده است.
- Hstacking Text و ویژگی‌های NLP با بردارهای ویژگی متن: در بخش مهندسی ویژگی، تعدادی بردار ویژگی‌های مختلف تولید کردیم که ترکیب آنها با هم می‌تواند به بهبود دقت طبقه بندی کننده کمک کند.
 - تنظیم Hyperparamter در مدل‌سازی: تنظیم پارامترها یک مرحله مهم است، تعدادی از پارامترها مانند طول درخت، برگ‌ها، پارامترهای شبکه و غیره را می‌توان به خوبی تنظیم کرد تا بهترین مدل متناسب را به دست آورد.
 - مدل‌های گروهی: چیدمان مدل‌های مختلف و ترکیب خروجی‌های آنها می‌تواند به بهبود بیشتر نتایج کمک کند. اطلاعات بیشتر در مورد مدل‌های مجموعه را در لینک زیر بخوانید:

<https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning>

پروژه مشابه

می‌توانید از پروژه <https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand->

<https://github.com/jhaber-zz/text-classify-2021> یا <https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand->

بگیرید.

هرگونه ابهام یا سوال را حضوری یا آنلاین بپرسید.