



# Analyzing *Stack Overflow* Community Posts to Automate Knowledge Organization

---

U of T UnERD 2021

Arjun Sridharkumar, **Hamza Dugmag**,  
Iftexhar Ahmed, Shurui Zhou

# StackOverflow.com

The screenshot displays the Stack Overflow homepage. The top navigation bar includes the Stack Overflow logo, a 'Products' link, a search bar, and user profile icons. The left sidebar contains navigation links for Home, PUBLIC (Questions, Tags, Users), COLLECTIVES (Explore Collectives), FIND A JOB (Jobs, Companies), and TEAMS. A 'Stack Overflow for Teams' promotional box is also present.

The main content area is titled 'Top Questions' and features a filter bar with options: Interesting, 315 Bountied, Hot, Week, and Month. Below this, a list of questions is shown, each with its vote count, answer count, view count, title, tags, and the user who asked it.

**Questions listed:**

- Shuffle and unscramble a string (Lua)**: 0 votes, 0 answers, 3 views. Tags: string, random, lua, shuffle. Asked 1 min ago by Happy Coder 1.
- Printing without blank lines**: 0 votes, 0 answers, 5 views. Tags: python, python-3.x, beautifulsoup, line, space. Asked 1 min ago by GONZALO EMILIO CONDOR TASAYCO 9.
- How to make the type of a property dependent on the presence of another property?**: 0 votes, 0 answers, 3 views. Tags: reactjs, typescript. Asked 1 min ago by Flandre Scarlet 43.
- Problem setting dynamically the String text in a SpanLabel Constructor**: 1 vote, 1 answer, 12 views. Tags: label, codenameone. Answered 1 min ago by Shai Almog 50.1k.
- Encrypting and decrypting a folder**: 0 votes, 0 answers, 7 views. Tags: encryption, openssl, cryptography. Modified 1 min ago by Wang Zixiang 1.
- .htaccess select desktop or mobile folder**: 0 votes, 0 answers, 3 views. Tags: php, android, apache, htaccess. Asked 1 min ago by Chris Orea 1.
- How to use twig render variable inside another one**: 0 votes, 0 answers, 2 views. Tags: php, codeigniter, twig, twig-extension. Asked 2 mins ago by Erraco 75.
- Program Design - Userform and Dependent Lists/SQL Tables**: 0 votes, 0 answers, 3 views. Tags: sql, excel, vba, ms-access. Asked 2 mins ago by Tpeters 1.

The right sidebar contains sections for 'The Overflow Blog' (Podcast 367, Using stretch work assignments), 'Featured on Meta' (Outdated Answers, Don't be that account), 'Hot Meta Posts' (15 Is a small (but noteworthy) edit on several answers to one question acceptable?, 34 How does one distinguish typos from coding errors in questions?), 'Custom Filters' (Create a custom filter), and 'Watched Tags' (Watch a tag).



# 22M

# Questions

70% of which are answered as  
of July 2021<sup>1</sup>

<sup>1</sup> <https://stackexchange.com/sites?view=list#questions>





# 15M

# Users

As of July 2021 <sup>1</sup>

<sup>1</sup> <https://stackexchange.com/sites?view=list#questions>

## 2 How to horizontally center an element

Asked 12 years, 11 months ago Active 2 months ago Viewed 4.4m times

How can I horizontally center a `<div>` within another `<div>` using CSS?

4675

```
<div id="outer">
  <div id="inner">Foo foo</div>
</div>
```



1046

html css alignment centering

Share Edit Follow

edited Dec 28 '20 at 14:42

community wiki  
40 revs, 26 users 18%  
Peter Mortensen

30 Of those great answers, I just want to highlight that you must give `#inner` a `width`, or it will be `100%`, and you can't tell if it's already centered. – Jony Nov 7 '17 at 8:22

Add a comment

121 Answers

Active Oldest Votes

1 2 3 4 5 Next

You can apply this CSS to the inner `<div>`:

5133

```
#inner {
  width: 50%;
  margin: 0 auto;
}
```



Of course, you don't have to set the `width` to `50%`. Any width less than the containing `<div>` will work. The `margin: 0 auto` is what does the actual centering.

If you are targeting [Internet Explorer 8](#) (and later), it might be better to have this instead:

```
#inner {
  display: table;
  margin: 0 auto;
}
```

Question

Code  
snippet

Author

Answer

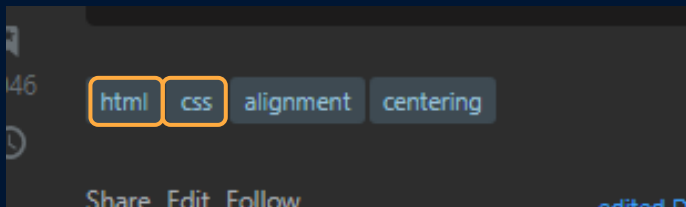
Upvotes

Tags

Comment

Accepted  
answer

# Tag Wiki



Similar to official documentations of programming technologies

<sup>3</sup> <https://stackoverflow.com/tags/html/info>

## About html

[Ask Question](#)

### Tag Info

Info	Newest	11 Bountied	Frequent	Votes	Active	Unanswered
------	--------	-------------	----------	-------	--------	------------

HTML (HyperText Markup Language) is the markup language for creating web pages and other information to be displayed in a web browser. Questions regarding HTML should include a minimal reproducible example and some idea of what you're trying to achieve. This tag is rarely used alone and is often paired with [CSS] and [javascript].

[HTML](#) (HyperText Markup Language) is the markup language used for structuring web pages and other information to be displayed in a web browser. HTML describes the structure of a web-page semantically along with cues for presentation, making it a markup language rather than a programming language. A browser 'renders' HTML in conjunction with CSS, which defines the 'style' (colors, fonts, layout, etc.) and JavaScript, which defines interactive and dynamic elements, adding style and behavior to the pages.

<https://html.spec.whatwg.org/multipage/> is the canonical HTML specification.

[HTML](#) (HyperText Markup Language) is the main markup language for creating web pages and other information to be displayed to humans in a web-browser.

It was invented by Sir [Tim Berners-Lee](#) while developing the first Web browser at [CERN](#) to enable researchers to share their findings and formally released in June 1993. The original "[HTML Tags](#)" were first publicly mentioned by Berners-Lee in 1991 and borrowed the syntax from CERN's [SGML-based documentation standard](#). The latest version for HTML is [HTML5.2](#).

HTML elements form the building blocks of all web-pages. HTML allows images and objects to be embedded in a page. It references styles and scripts and carries meta-data. It can be used to create interactive forms. It provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links, quotes and other items. It can embed scripts written in languages such as JavaScript, which affects the behavior of HTML web pages. Web pages written in different programming languages (PHP, JSP, VF, ASP.NET etc.) get rendered as HTML in a browser.

HTML is a hierarchical (tree-structured) markup language. That is, an item might be a descendant of another item, which is its ancestor. However, if item2 is a descendant of item1, then they have an additional special relation: item2 is *inside* of item1, or item1 is *wrapped around* item2.

### Syntax

HTML is written in the form of elements consisting of tags (and their attributes) enclosed in angle brackets (e.g., `<html>`).

HTML tags most commonly come in pairs. The first is known as the *opening tag* and the second, which includes a forward slash, as the *closing tag* (e.g., `<div>` and `</div>`). Major types of content, such as text or

# Community Wiki Posts

community wiki  
16 revs, 12 users 22%  
AhmerMH

Multiple collaborators  
can contribute to a  
question, answer, or  
tag wiki

```
width:100%;  
display: flex;  
justify-content: center;  
}
```

```
<div id="outer">  
  <div id="inner">Foo foo</div>  
</div>
```

[Run code snippet](#)[Expand snippet](#)

To align the div vertically centered, use the property `align-items: center`.

Share Edit Follow

edited Jan 23 at 6:05

community wiki  
16 revs, 12 users 22%  
AhmerMH

17 For the vertical centering I usually use "line-height" (line-height == height). This is simple and nice but it's only working with a one line content text :) – Nicolas Guillaume Jun 23 '10 at 12:36



# However...

the Q&A format makes it difficult to  
retrieve desired knowledge via unknown-  
item searches [1].

[1] A. Diriye, M. L. Wilson, A. Blandford and A. Tombros,  
"Revisiting Exploratory Search from the HCI Perspective,"  
London, 2010.



## The Stack Overflow Regular Expressions FAQ

1058

See also a lot of general hints and useful links at the [regex tag details page](#).

### Online tutorials

- [RegexOne](#)
- [Regular Expressions Info](#)

### Quantifiers

- Zero-or-more: `*:greedy`, `*?:reluctant`, `++:possessive`
- One-or-more: `+:greedy`, `++?:reluctant`, `++:possessive`
- `?:optional (zero-or-one)`
- Min/max ranges (all inclusive): `{n,m}:between n & m`, `{n,}:n-or-more`, `{n}:exactly n`
- Differences between greedy, reluctant (a.k.a. "lazy", "ungreedy") and possessive quantifier:
  - [Greedy vs. Reluctant vs. Possessive Quantifiers](#)
  - [In-depth discussion on the differences between greedy versus non-greedy](#)
  - [What's the difference between {n} and {n}?](#)
  - [Can someone explain Possessive Quantifiers to me?](#) `php`, `perl`, `java`, `ruby`
  - [Emulating possessive quantifiers](#) `.net`
  - Non-Stack Overflow references: From [Oracle](#), [regular-expressions.info](#)

### Character Classes

- [What is the difference between square brackets and parentheses?](#)
- `[...]`: any one character, `[^...]`: negated/any character but
- `[^]` matches any one character including newlines `javascript`
- `[\w-[\d]] / [a-z-[qz]]`: set subtraction `.net`, `xml-schema`, `xpath`, JGSoft
- `[\w&&[\d]]`: set intersection `java`, `ruby` 1.9+
- `[[:alpha:]]`: POSIX character classes
- [Why do \[\d\] , \[\d-9\] , \[\d-9\] get different results in Java?](#) `java`
- Shorthand:

Headings

Links to  
other *STO*  
posts

# “Structured Posts”

Community posts which aim to organize the *Stack Overflow* website

<sup>4</sup> <https://stackoverflow.com/a/22944075/10757178>

# 5K/22M

---

questions are linked in  
community posts, indicating  
that *Stack Overflow* is largely  
unorganized<sup>5</sup>

<sup>5</sup> <https://data.stackexchange.com/stackoverflow/query/new>

# Research Project Goals

## Analysis

What are the characteristics of community and structured posts?

## Prototyping

How can we create a tooling method that automatically generates structured posts?

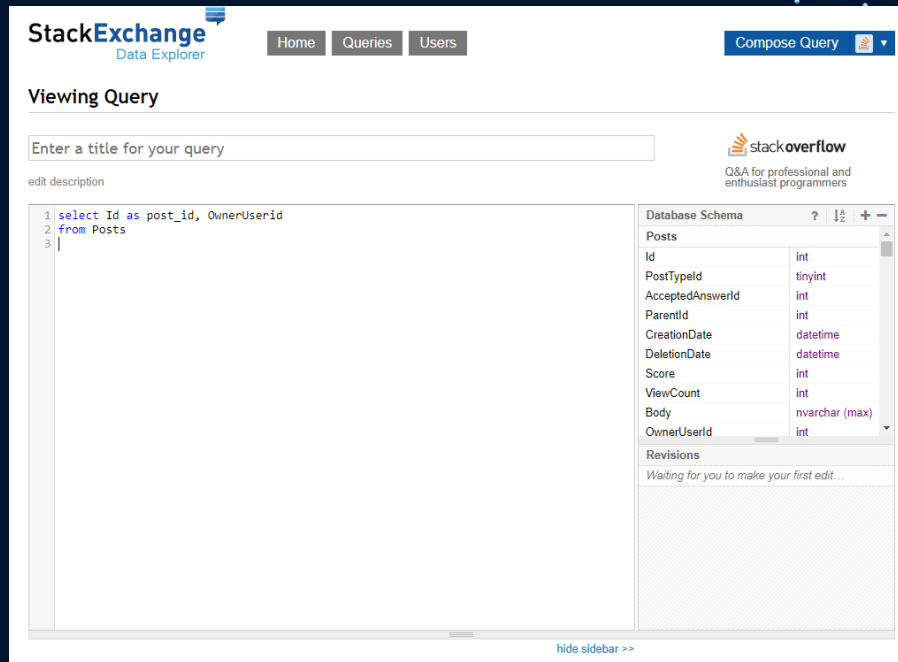
## Hypothesis

**It is expected that there  
are shared features  
among high-quality  
structured posts**

## Approach

**Unsupervised clustering  
groups similar posts  
together**

# Data Acquisition



The screenshot shows the StackExchange Data Explorer interface. At the top, there's a navigation bar with 'StackExchange Data Explorer' logo, 'Home', 'Queries', and 'Users' buttons, and a 'Compose Query' button. Below this, the 'Viewing Query' section has a text input for a title and an 'edit description' link. The main area displays a SQL query: 

```
1 select Id as post_id, OwnerUserId
2 from Posts
3 |
```

. To the right, the 'Database Schema' section shows the 'Posts' table with columns: Id (int), PostTypeId (tinyint), AcceptedAnswerId (int), ParentId (int), CreationDate (datetime), DeletionDate (datetime), Score (int), ViewCount (int), Body (nvarchar (max)), and OwnerUserId (int). Below the schema is a 'Revisions' section with the text 'Waiting for you to make your first edit...'. A 'hide sidebar >>' link is at the bottom right.

StackExchange Data Explorer

Home Queries Users

Compose Query

Viewing Query

Enter a title for your query

edit description

```
1 select Id as post_id, OwnerUserId
2 from Posts
3 |
```

Database Schema

Posts

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

Revisions

Waiting for you to make your first edit...

hide sidebar >>

Stack Overflow data dump since 2008 <sup>5</sup>

<sup>5</sup> <https://data.stackexchange.com/stackoverflow/query/new>





# 126K

community posts created since 2008

# Post Properties and Features



## Length

Text Length  
No. of *STO* Links  
No. of External Links  
Total No. of Links  
Ratio of *STO* Links



## Activity

No. of Edits  
Creation Date  
Last Update Date  
Inactivity Time  
No. of Contributors



## Popularity

No. of Upvotes



## Structure

No. of Headings  
Avg. No. of *STO* Links /Heading

# Raw Data

	post_id	Text Length	No. of SO Links	No. of ALL Links	ALL - SO	SO / ALL	No. of Revisions	Creation Date	Last Update Time	No. of Contributors	Creation - Updated	No. of Upvotes	No. of Headings	Avg. No. of SO Links / Heading
0	1587.0	59.0	0.0	3.0	3.0	0.0	1.0	4760.0	4760.0	1.0	0.0	14.0	0.0	0.0
1	1660.0	7.0	0.0	8.0	8.0	0.0	5.0	4760.0	4728.0	2.0	31.0	91.0	3.0	0.0
2	1752.0	107.0	0.0	0.0	0.0	0.0	5.0	4756.0	4756.0	1.0	0.0	15.0	0.0	0.0
3	2376.0	40.0	0.0	0.0	0.0	0.0	2.0	4756.0	4756.0	2.0	0.0	0.0	0.0	0.0
4	2420.0	97.0	0.0	3.0	3.0	0.0	5.0	4756.0	4352.0	4.0	404.0	35.0	1.0	0.0

First five unscaled data points used after data analysis

	Text Length	No. of SO Links	No. of ALL Links	ALL - SO	SO / ALL	No. of Revisions	Creation Date	Last Update Time	No. of Contributors	Creation - Updated	No. of Upvotes	No. of Headings	Avg. No. of SO Links / Heading
0	0.010269	0.0	0.009583	0.013947	0.0	0.003731	1.000000	1.000000	0.008926	0.000212	0.002283	0.000000	0.0
1	0.001219	0.0	0.025558	0.037201	0.0	0.018646	1.000000	0.993652	0.017853	0.006779	0.005367	0.048401	0.0
2	0.018616	0.0	0.000000	0.000000	0.0	0.018646	0.999023	0.999023	0.008926	0.000212	0.002323	0.000000	0.0
3	0.006962	0.0	0.000000	0.000000	0.0	0.007462	0.999023	0.999023	0.017853	0.000212	0.001722	0.000000	0.0
4	0.016876	0.0	0.009583	0.013947	0.0	0.018646	0.999023	0.914062	0.035706	0.085754	0.003124	0.016129	0.0

First five normalized data points used in clustering

Note: the following data is from a random sample of 6K posts since the script is very slow and we have yet to migrate all the data to our server

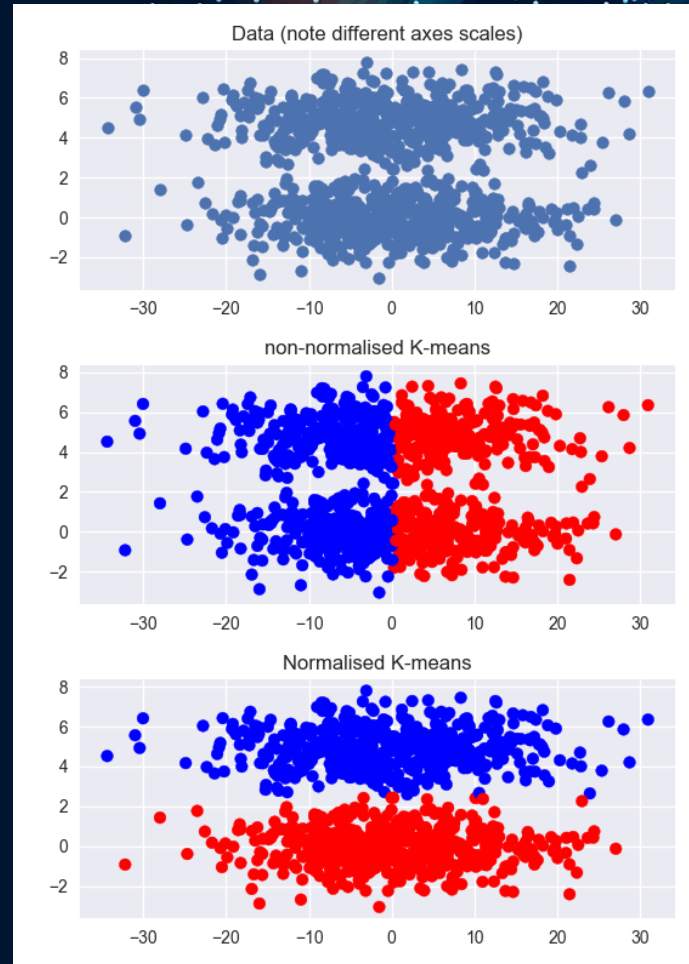
# Data Normalization

Normalizing data equally weighs all features [2].

Don't want distances to be a factor

[2] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft computing*, vol. 97, 2020.

<sup>6</sup> <https://stats.stackexchange.com/a/283941>



# Visualizing High-Dimensional Data with T-Distributed Stochastic Neighbor Embedding (TSNE)

TSNE represents high-dimensional data in 2D by creating a probability distribution of how close data points are to each other [3].

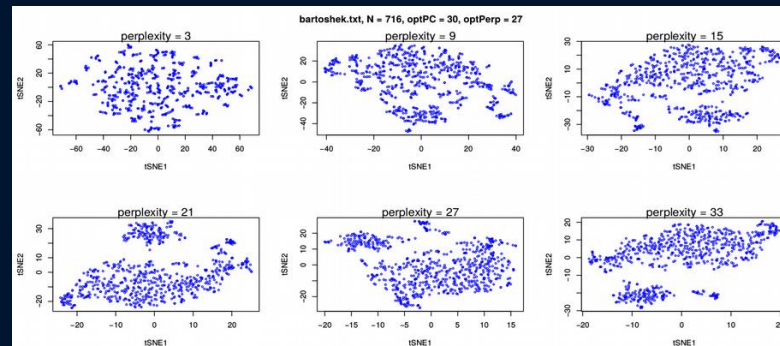
## ***Hyperparameter: Perplexity***

Attention between local and global aspects of the data

Perplexity  $\sim N^{0.5}$  [4].

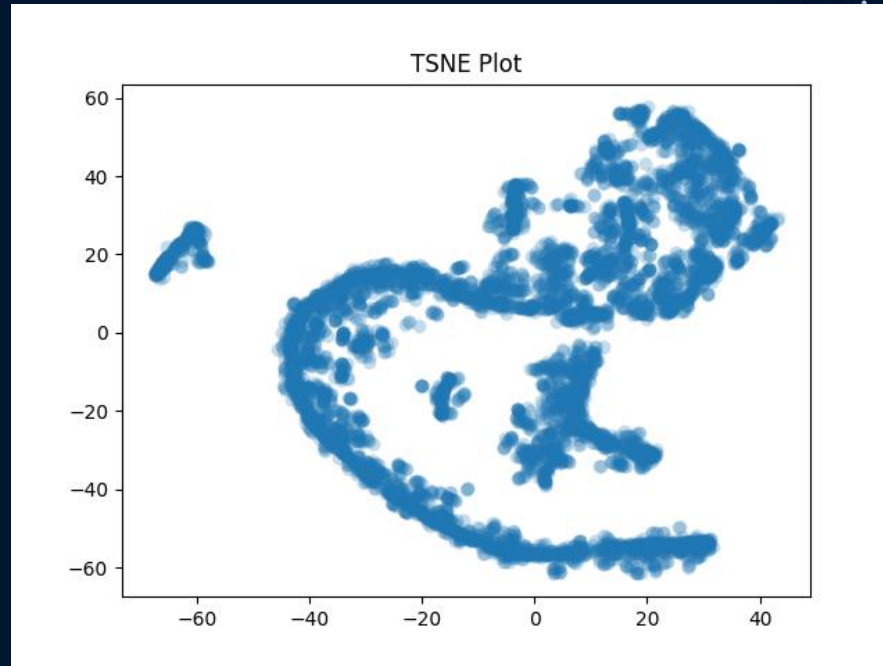
[3] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

[4] N. Oskolkov, "How to tune hyperparameters of tSNE," Towards Data Science, 18 July 2019. [Online]. Available: <https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868>.





# TSNE Plot



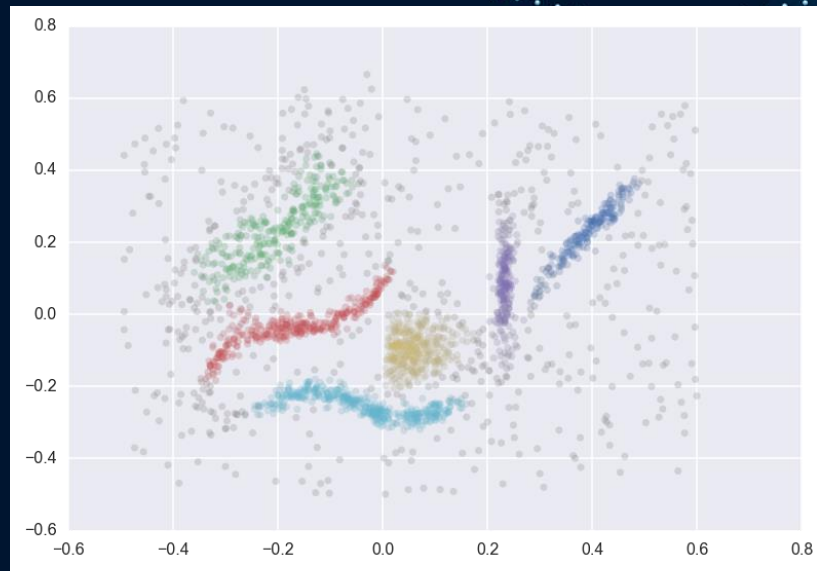
Unclustered, scaled, projected data

# Hierarchical Density Based Spatial Clustering of Applications w/ Noise (HDBSCAN)

1. Generates sets of nested, related clusters
2. We don't know how many clusters we have
3. We don't know how big the clusters are
4. Indifferent to cluster shapes
5. Detects outliers
6. Easy and optimized parameter selection

[5]

[5] L. Yang et al., "Semi-Supervised Log-Based Anomaly Detection via Probabilistic Label Estimation," 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), 2021, pp. 1448-1460, doi: 10.1109/ICSE43902.2021.00130.



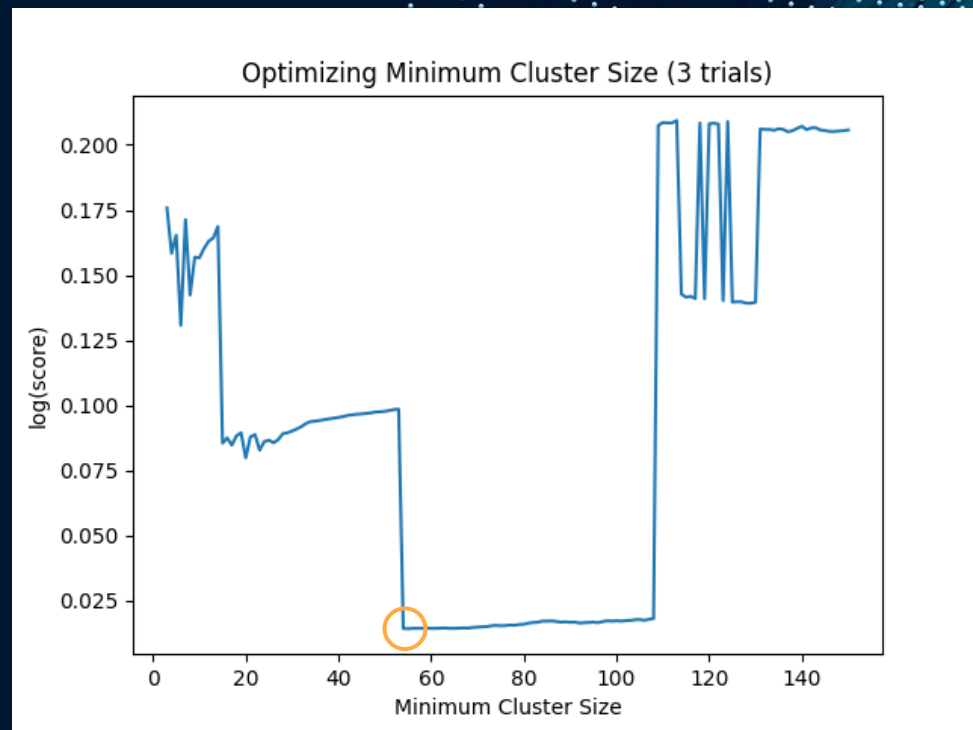
<sup>7</sup> [https://hdbscan.readthedocs.io/en/latest/advanced\\_hdbscan.html](https://hdbscan.readthedocs.io/en/latest/advanced_hdbscan.html)

# Optimizing Min\_Cluster\_Size

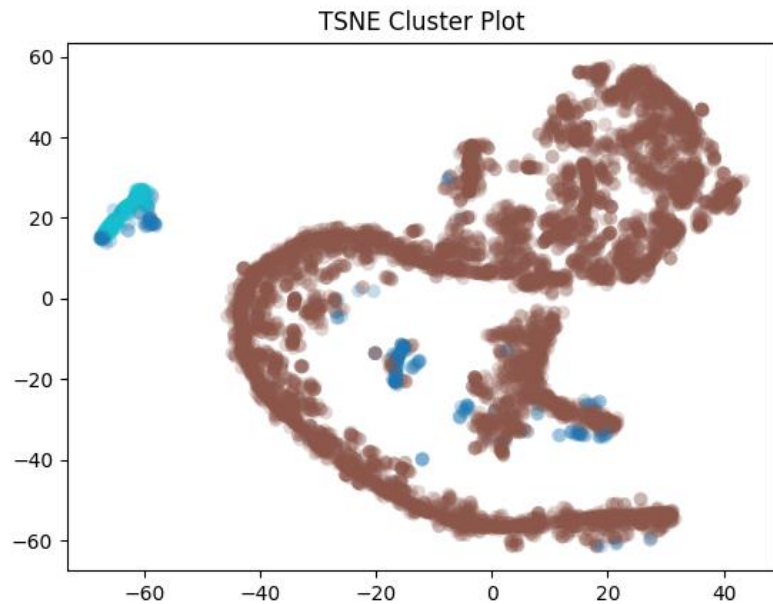
Want to maximize confidence that data points belong in their clusters

Find Min\_Cluster\_Size that minimizes number of points with <5% confidence [6]

[6] N. Oskolkov, "How to cluster in High Dimensions," Towards Data Science, 23 July 2019. [Online]. Available: <https://towardsdatascience.com/how-to-cluster-in-high-dimensions-4ef693bacc6>.



# The Clusters



Cluster 0

Cluster 1

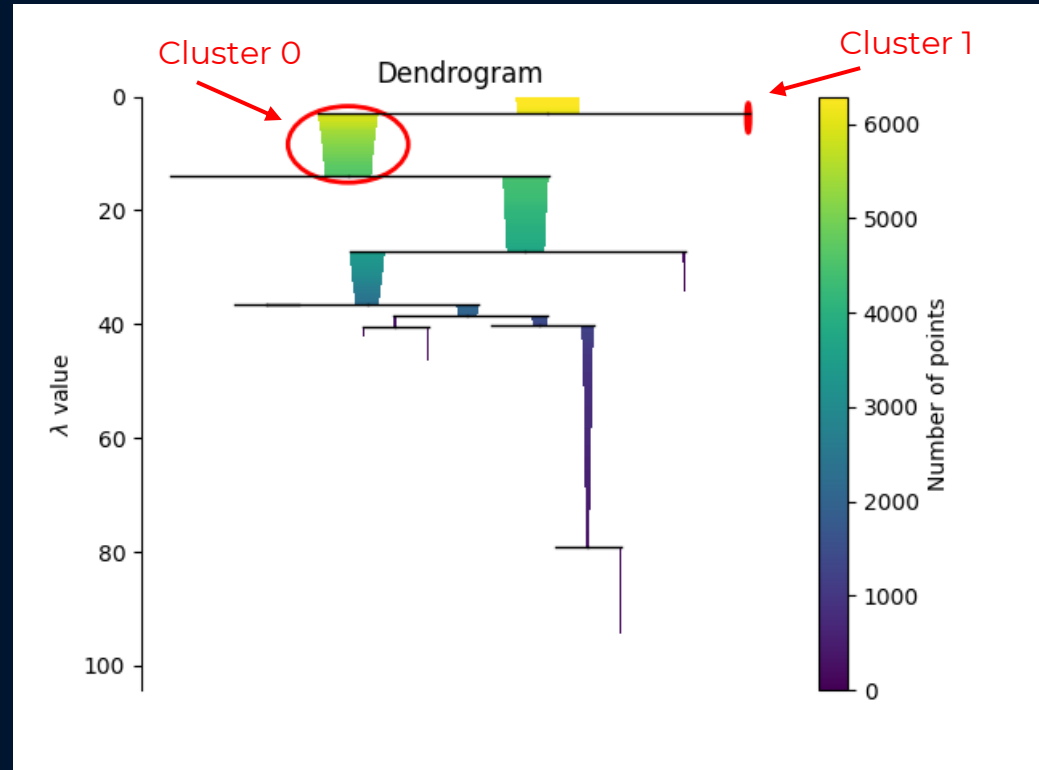
Noise

HDBSCAN TSNE plot with min\_cluster\_size = 55

# Dendrogram Analysis

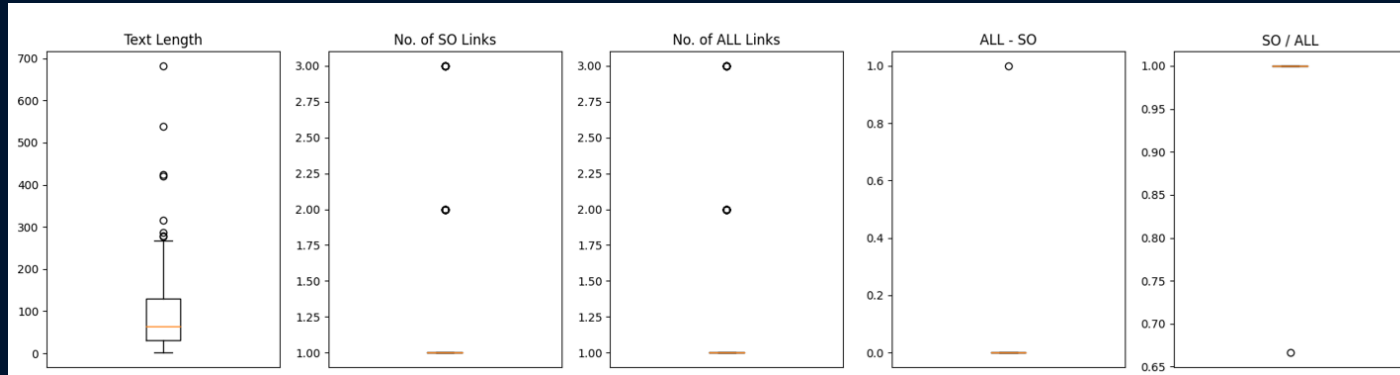
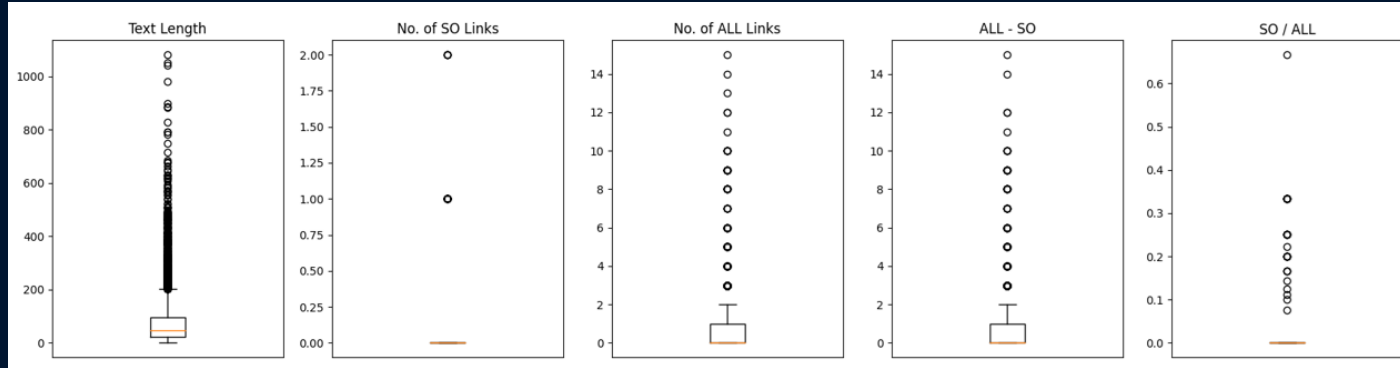
HDBSCAN uses cluster hierarchy to maximize the clusters' lifespans without introducing too many artifacts [7].

[7] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering In: Journal of Open Source Software, The Open Journal, volume 2, number 11. 2017

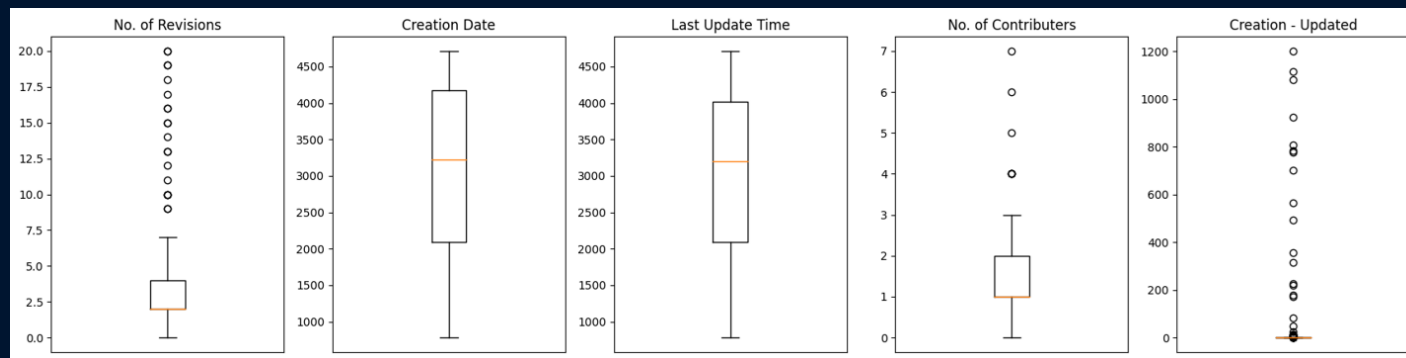
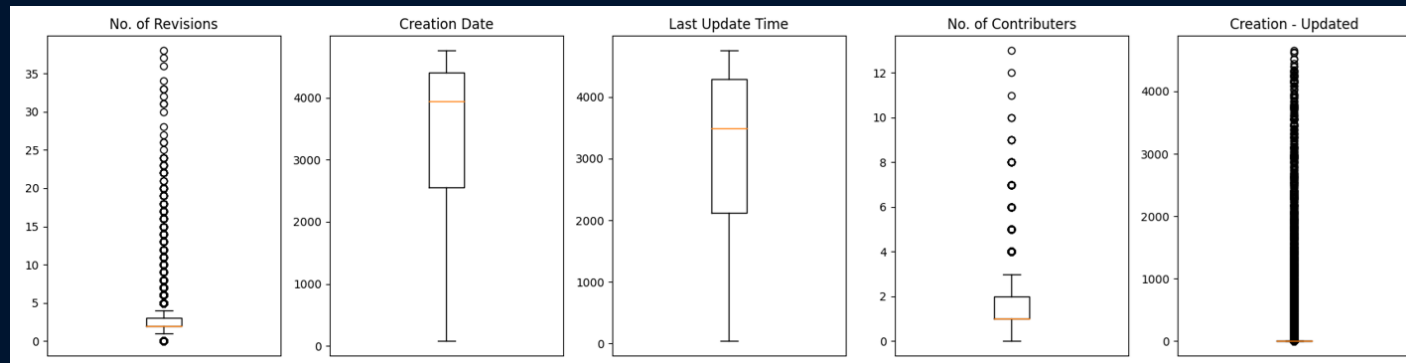




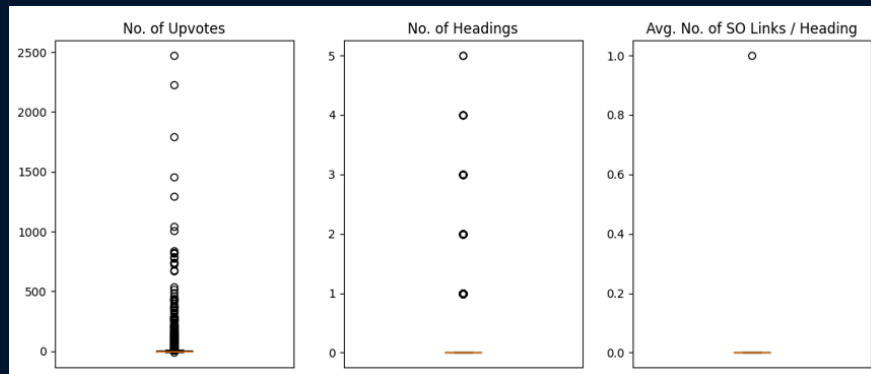
# Length Comparison



# Activity Comparison

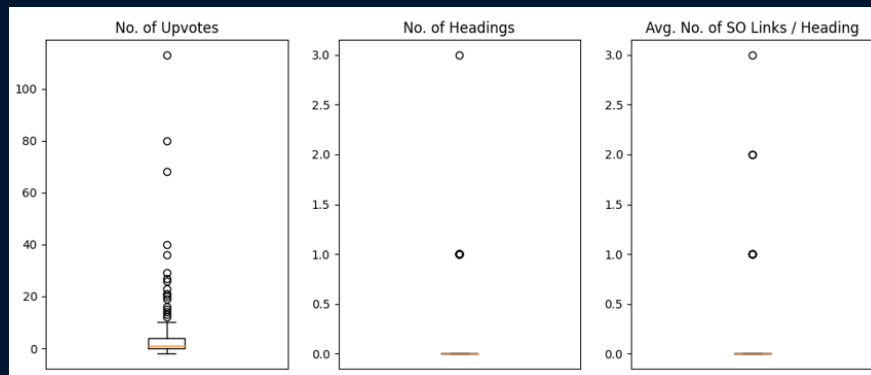


# Popularity and Structure Comparison



Cluster 0

Lower median upvotes  
and *STO* links per  
heading



Cluster 1

# Stratified Sampling: Cluster 0

Pretty long, but is not a collaborative “structured post” that serves to organize *Stack Overflow*

<sup>8</sup> <https://stackoverflow.com/questions/40648561>



6



Heres a complete function for adding and removing parameters based on this question and this github gist: <https://gist.github.com/excalq/2961415>

```
var updateQueryStringParam = function (key, value) {
  var baseUrl = [location.protocol, '//' , location.host, location.pathname].join('');
  urlQueryString = document.location.search;
  newParam = key + '=' + value;
  params = '?' + newParam;

  // If the "search" string exists, then build params from it
  if (urlQueryString) {
    updateRegex = new RegExp('[\?&]' + key + '[^&]*');
    removeRegex = new RegExp('[\?&]' + key + '[^&]*[&]?');

    if( typeof value == 'undefined' || value == null || value == '' ) { // Remove
      params = urlQueryString.replace(removeRegex, "$1");
      params = params.replace( /[&]$/, "" );
    } else if (urlQueryString.match(updateRegex) != null) { // If param exists a
      params = urlQueryString.replace(updateRegex, "$1" + newParam);
    } else { // Otherwise, add it to end of query string
      params = urlQueryString + '&' + newParam;
    }
  }
  window.history.replaceState({}, "", baseUrl + params);
};
```

You can add parameters like this:

```
updateQueryStringParam( 'myparam', 'true' );
```

And remove it like this:

```
updateQueryStringParam( 'myparam', null );
```

In this thread many said that the regex is probably not the best/stable solution ... so im not 100% sure if this thing has some flaws but as far as i tested it it works pretty fine.

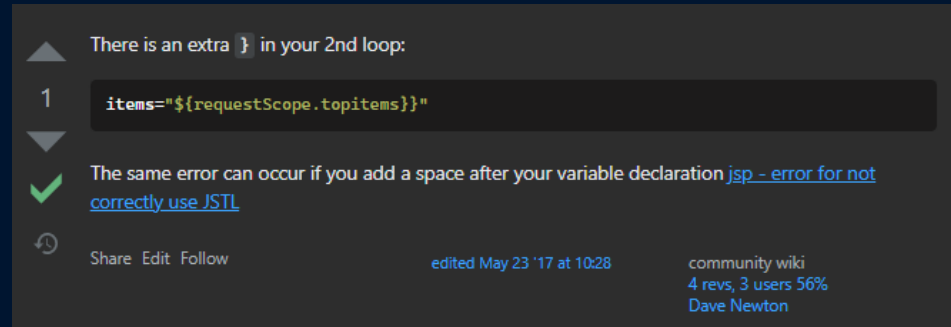
Share Edit Follow

answered Nov 17 '16 at 6:59

community wiki  
GDY

# Stratified Sampling: Cluster 1

Too short to be considered  
“structured,” but is a collaborative  
effort and facilitates finding other  
information on *Stack Overflow*



The screenshot shows a Stack Overflow interface. At the top, a grey bar contains the text "There is an extra `}` in your 2nd loop:". Below this, a question is listed with the number "1" and a code snippet: `items="${requestScope.topitems}}"`. A green checkmark icon indicates an accepted answer. The answer text reads: "The same error can occur if you add a space after your variable declaration [jsp - error for not correctly use JSTL](#)". At the bottom of the answer, it says "edited May 23 '17 at 10:28", "community wiki", "4 revs, 3 users 56%", and "Dave Newton".

<sup>9</sup> <https://stackoverflow.com/questions/24247862>



# Stratified Sampling: Noise

1  
0One word answer: **asynchronicity**.

649



## Forewords



This topic has been iterated at least a couple of thousands of times, here, in Stack Overflow. Hence, first off I'd like to point out some extremely useful resources:

+400



- [@Felix Kling's answer to "How do I return the response from an asynchronous call?"](#). See his excellent answer explaining synchronous and asynchronous flows, as well as the "Restructure code" section. [@Benjamin Gruenbaum](#) has also put a lot of effort explaining asynchronicity in the same thread.
- [@Matt Esch's answer to "Get data from fs.readFile"](#) also explains asynchronicity extremely well in a simple manner.

## The answer to the question at hand

Let's trace the common behavior first. In all examples, the `outerScopeVar` is modified inside of a *function*. That function is clearly not executed immediately, it is being assigned or passed as an argument. That is what we call a **callback**.

Now the question is, when is that callback called?

It depends on the case. Let's try to trace some common behavior again:

- `img.onload` may be called *sometime in the future*, when (and if) the image has successfully loaded.
- `setTimeout` may be called *sometime in the future*, after the delay has expired and the timeout hasn't been canceled by `clearTimeout`. Note: even when using `0` as delay, all browsers have a minimum timeout delay cap (specified to be 4ms in the HTML5 spec).
- jQuery `$.post`'s callback may be called *sometime in the future*, when (and if) the Ajax request has been completed successfully.
- Nodejs's `fs.readFile` may be called *sometime in the future*, when the file has been read successfully or thrown an error.

In all cases, we have a callback which may run *sometime in the future*. This "sometime in the future" is what we refer to as **asynchronous flow**.

Asynchronous execution is pushed out of the synchronous flow. That is, the asynchronous code will **never** execute while the synchronous code stack is executing. This is the meaning of JavaScript

```
// call the callback passing the result as argument
callback('Nya');
}, Math.random() * 2000);
}
```

[Run code snippet](#)
[Expand snippet](#)

Most often in real use cases, the DOM API and most libraries already provide the callback functionality (the `helloCatAsync` implementation in this demonstrative example). You only need to pass the callback function and understand that it will execute out of the synchronous flow, and restructure your code to accommodate for that.

You will also notice that due to the asynchronous nature, it is impossible to `return` a value from an asynchronous flow back to the synchronous flow where the callback was defined, as the asynchronous callbacks are executed long after the synchronous code has already finished executing.

Instead of `return`ing a value from an asynchronous callback, you will have to make use of the callback pattern, or... Promises.

## Promises

Although there are ways to keep the [callback hell](#) at bay with vanilla JS, promises are growing in popularity and are currently being standardized in ES6 (see [Promise - MDN](#)).

Promises (a.k.a. Futures) provide a more linear, and thus pleasant, reading of the asynchronous code, but explaining their entire functionality is out of the scope of this question. Instead, I'll leave these excellent resources for the interested:

- [JavaScript Promises - HTML5 Rocks](#)
- [You're Missing the Point of Promises - domenic.me](#)

## More reading material about JavaScript asynchronicity

- [The Art of Node - Callbacks](#) explains asynchronous code and callbacks very well with vanilla JS examples and Nodejs code as well.

A structured post! However, noise is inconsistent.

10

<https://stackoverflow.com/questions/23667087>

# Inference

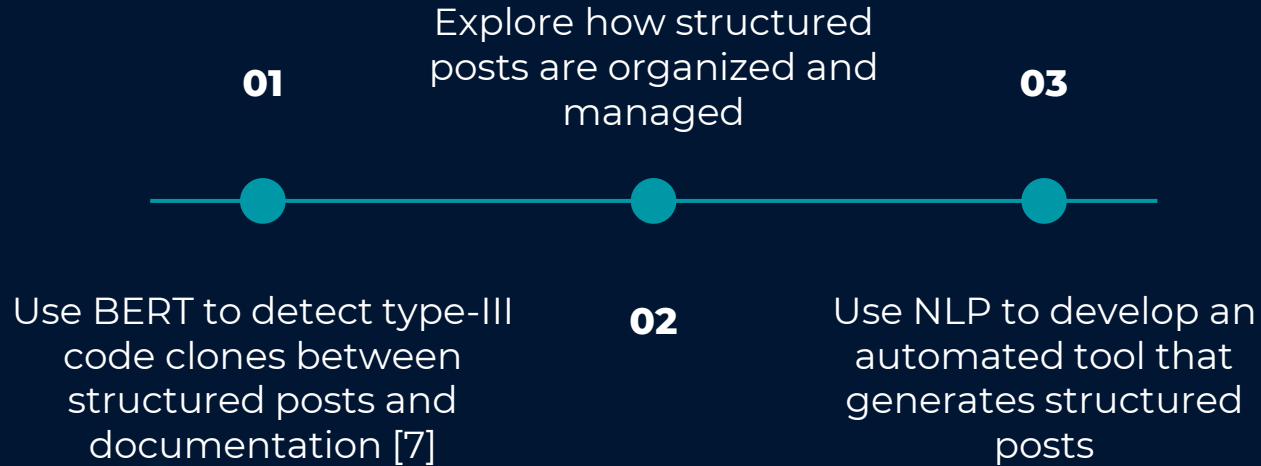
Structured posts are very rare:

- Time consuming to make
- Difficult to find (ironic!)

How can we better reveal structured posts?

- Cluster by property rather than all properties at the same time
- Introduce more features (e.g., no. of comments)
- Cluster latent variables

# Future Work



[7] Saini, Vaibhav, et al. "Oreo: Detection of clones in the twilight zone." Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE), 2018.

# Research Impact



## Collaboration

Encourages  
innovation and idea-  
sharing among  
developers



## Education

Increases  
accessibility to  
information for new  
and experienced  
developers



## Efficiency

Saves developers'  
time to increase  
innovation

---

# References

- <sup>1</sup> <https://stackexchange.com/sites?view=list#questions>
- <sup>2</sup> <https://stackoverflow.com/questions/114543/how-to-horizontally-center-an-element>
- <sup>3</sup> <https://stackoverflow.com/tags/html/info>
- <sup>4</sup> <https://stackoverflow.com/a/22944075/10757178>
- <sup>5</sup> <https://data.stackexchange.com/stackoverflow/query/new>
- <sup>6</sup> <https://stats.stackexchange.com/a/283941>
- <sup>7</sup> [https://hdbscan.readthedocs.io/en/latest/advanced\\_hdbscan.html](https://hdbscan.readthedocs.io/en/latest/advanced_hdbscan.html)
- <sup>8</sup> <https://stackoverflow.com/questions/40648561>
- <sup>9</sup> <https://stackoverflow.com/questions/24247862>
- <sup>10</sup> <https://stackoverflow.com/questions/23667087>



---

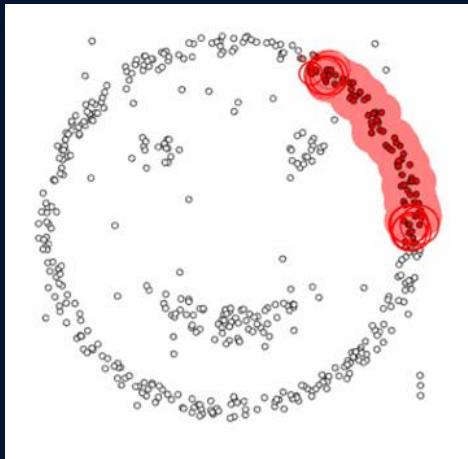
# References

- [1] A. Diriye, M. L. Wilson, A. Blandford and A. Tombros, "Revisiting Exploratory Search from the HCI Perspective," London, 2010.
- [2] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft computing*, vol. 97, 2020.
- [3] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.
- [4] N. Oskolkov, "How to tune hyperparameters of tSNE," Towards Data Science, 18 July 2019. [Online]. Available: <https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868>.
- [5] L. Yang et al., "Semi-Supervised Log-Based Anomaly Detection via Probabilistic Label Estimation," 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), 2021, pp. 1448-1460, doi: 10.1109/ICSE43902.2021.00130.
- [6] N. Oskolkov, "How to cluster in High Dimensions," Towards Data Science, 23 July 2019. [Online]. Available: <https://towardsdatascience.com/how-to-cluster-in-high-dimensions-4ef693bacc6>.
- [7] Saini, Vaibhav, et al. "Oreo: Detection of clones in the twilight zone." Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE). 2018.



# Thanks for listening!

---



[6]

