Large-Scale Indexing of Unstructured Social Media Texts to Find Interesting Conversations

Abstract

The primary reason for creating this paper is to make hopefully easy-to-understand documentation of how our algorithms work and share our discoveries. We hope that if you're reading this you'll gain a better understanding of the good work we are doing and then be able to inspire other ideas, optimizations and simplifications.

1. Design Goals

Our main goal is to provide high-quality results that the Discord community would love and find useful. This includes bringing together information and adding to the community to make it useful not only socially but intellectually and satisfy curiosity, providing another way to learn. Central to it is not only something that does its job well but something that helps our community find what they are looking for when they say “Get me information on Pokemon within this server”. Giving you information not only directly related to your search but also what you are likely to find helpful as well.

Our community should always feel like they’re getting a complete view of their data and never feel like they're missing something and so have to go back to the discord. We need to ensure they have the complete picture.

2. Evaluating existing ideas out there

Here we summarily discuss ideas that inspire our work. “The best navigation service should make it easy to find almost anything on the Web(given all the data)”

Discord Text Analysis Graphs:Github detailing statistical analyses methods on discord texts

Discord Text Analysis Graphs 2: Another visualizing link

Computer Networks and ISDN Systems | Journal | ScienceDirect.com by Elsevier: A peer-reviewed journal that covers the theory and practice of computer networks and integrated services digital network systems.

Web crawlers

Automated programs that are constantly in the background crawl across and index information. It's the reason why your Google search returns results nearly instantly.

Search Engine Watch

A website that provides in-depth resources and information on how search engines work and ranking algorithms. It’s the go-to source for understanding the mechanics of search engines.

World Wide Web Worm

One of the first search engines/crawlers to search for information. I believe studying how all these pre-existing systems worked would help us to better write out code. Includes Altavista which is the step up to search indexing efficiency, multimedia searching and natural language querying, from this we can possibly learn stuff about powerful searching.

WordNet

This is an extremely large lexical database for the English Language, it is designed to help machines (and humans) understand word relationships and meanings. It groups English words into sysnets. Synsnets are sets of synonyms that represent certain concepts. This could be a great resource for our search query augmentation- Understanding what the user is looking for(not just the keywords) -> returning useful information.

D-Wikis Algorithm 10/31/2024

SploreBot

Google Page Rank Algorithm:

The PageRank algorithm works as follows:

1. Suppose Page A has links pointing to it from pages $(T_1, T_2, \ldots, T_n)$.
2. Let (C(T)) represent the number of outbound links on page (T).
3. (d\) is the damping factor, typically set to around 0.85.

The PageRank formula for Page A is:

$$PR(A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right)$$

Since PageRanks represent a probability distribution across all webpages, the sum of all PageRanks is equal to 1.

This calculation corresponds to the principal eigenvector of the web's normalized link matrix.

More importantly, PageRank was a way to objectively measure a webpage’s importance via citations. Could use to rank conversations in order order importance.

3. Known problems and what current solutions don’t do so well

We try to avoid low-quality results at all costs. To do that we can identify issues that might cause such a problem and identify ways we can resolve them.

Spelling Mistakes:

A major problem is spelling mistakes…(Expand later)
We’re also keen on ensuring we have fast-crawling technology(Expand later)

Storage Indexing:

Given the potential user base and scale of information, we have to think about the storage.
space to efficiently store indices and running fast responses to queries.

Fast and Relevant Response to Queries:

(Expand Later)

Dealing with Junkwords:

Discord messages have a large amount of “lols” and “lmao”, “let’s goo” and other slang that add noise to the conversation--- “Junk Results”. Here we discuss ways to adequately id and take care of junk.

On average, people using search engines look at the first 10 results. Borrowing from Google’s thoughts, it’s a lot more important to return aa small amount of high-quality information was more important than simply returning a large amount of information Figuring out that there are certain things users are willing to do and unwilling to do is important

How to deal with shorthands and abbrevs for information since in social media, people are deadass trying to use the minimum range of their linguistic ability

4. Benchmarks and Quality Control

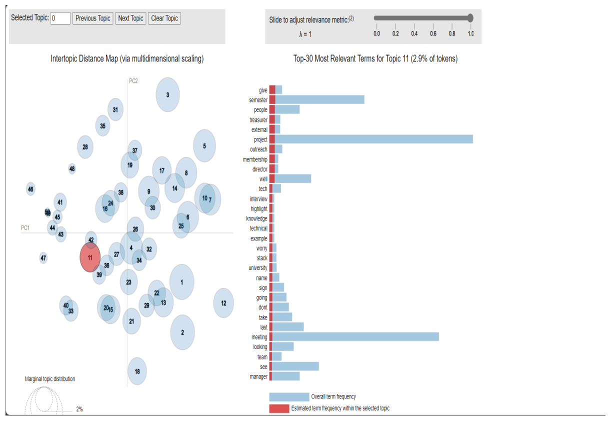
For x million texts, how many key interesting vectors can it find
I hate the fact that the algorithm might miss something interesting so how to make sure we are not missing anything

5. Structural Implementations and Ideas

Jiankun Wang Kyubin Min Ella Happel Shubhan Chari Nathan Amankwah

There are many ways and great discussions we're having about how to go about implementing all this.

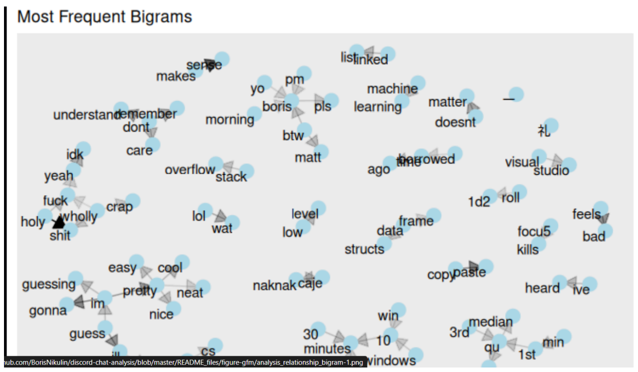
Samarth's Solve



Using the LDA unsupervised approach, we'll sort words into a bucket of 50 topic IDs. The words should roughly be representative of the topic and ranked in someway

D-Wikis Algorithm 10/31/2024

Characteristic Words. More than just tf-idf
How likely are two words related to each other, will help with bag of words algorithm



Nathan’s Solve:

General Structure::
1. Focus on a particular message M_i. Clean it by remove stopwords, extracting the nouns from it and augmenting the nouns in it
2. Extend your area of data to 5 texts before and then 5 texts after, selecting only messages which match the bag of nouns. Repeat 1. Again, further augmenting and adding to the bag of nouns.
3. Repeat at the fringes and continue this process.
4. Now we have the messages that are relevant to some topic idea or query search. Further processing.

This is simply a general idea for how the algorithm could work, extra things need to be done to sort nouns in order of topical relevance, eg: a custom tf-IDF prioritizing more topical nouns. Good use could be made of the English Lexicon in this which is the complete set of words in the English Language.

Cultivation: This refers to growing text/focusing analysis based on the idea of a zoom in feature. The first pass of the algorithm will likely have a very generalized view of the data but if the user wants to further explore a concept, they could specify as such and the algorithm will get more information related to that. A huge challenge to this algorithm will be figuring out ways to minimize noise and augment quality. One way we can do this already mentions is by augmenting the data in various ways including weighting various nouns and texts based on the user or length of message. Standardized quality tests need to be developed to validate and compare any versions of the algorithm for quality output. This training data could be found in public discord servers.

An edge case of a foreseeable problem however is data voids. This happens when the user wants info on something in the server but it is simply to sparse for quality results on it.
The link below is a helpful resource for data augumentation via synonyms.
https://conceptnet.io/

Various preprocessing techniques that could be used include Stemming and stop word removal. Word clouds could be use to visualize and verify the removal of stop words

The algorithm needs to become more sophisticated for zoom-in features. This includes subindexing, and recrawling. Different servers naturally have different qualities in the kind of data present and as such a more dynamic algorithm might be necessary to adjust parameters to suit the kind of information.

Extension: The algorithm could potentially be extended to reddit threads to create a sort of Reddit Wikis. Allowing us to reduce the mess and turn people's opinions into some kind of agreed consensus as a source similar to how Wikipedia functions.

For the wiki we needs someway to sort the importance of a topic and I beleive this can be done similarly to google's pageRank but applied to users where users who contribute more in the community have more authoritative 'webpage' rank in their comments

Parameters::
Energy of Algorithm - Parameter involving how deeply it continues to search on trees/threads of information when it finds more quality bits of information.

For deeper inspiration this is a video about the knowledge graph, google's more intelligent search that integrates information together:
https://www.youtube.com/watch?v=mmQl6VGvX-c

How we deal with multi-modal data.

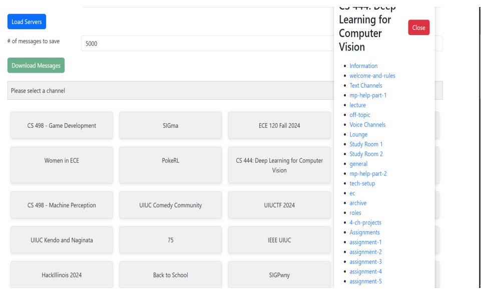
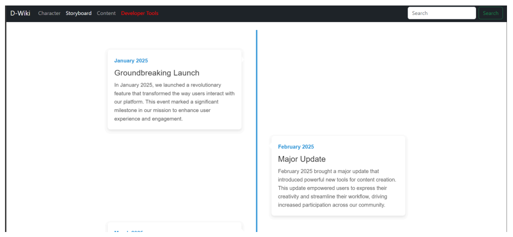
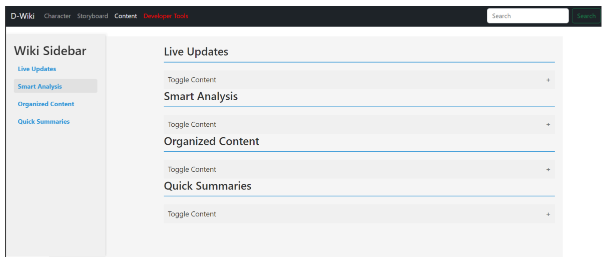
We can use the hyper-summary below image and videos text

To be expanded:

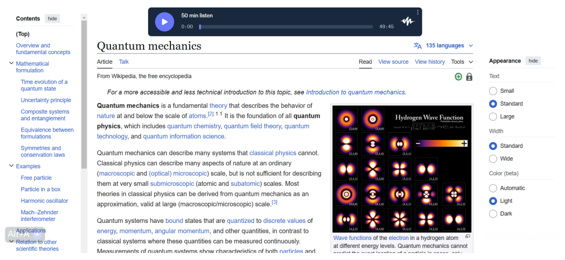
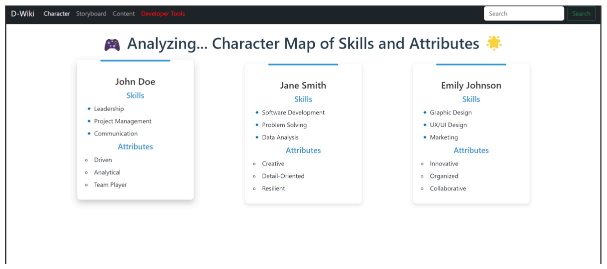
…

D-Wikis Algorithm 10/31/2024

Visuals Implementation

Visuals.Theme	Kyubin Min Ella Happel
Visuals.Servers	 Load Servers # of messages to save 5000 Download Messages Please select a channel CS 498 - Game Development · SIGma · ECE 120 Fall 2024 Women in ECE · PokeRL · CS 444: Deep Learning for Computer Vision CS 498 - Machine Perception · UIUC Comedy Community · UIUCTF 2024 UIUC Kendo and Naginata · 75 · IEEE UIUC HackIllinois 2024 · Back to School · SIGPwny Learning for Computer Vision · Close • Information • welcome-and-rules • Text Channels • mp-help-part-1 • lecture • off-topic • Voice Channels • Lounge • Study Room 1 • Study Room 2 • general • mp-help-part-2 • tech-setup • ec • archive • roles • 4-ch-projects • Assignments • assignment-1 • assignment-2 • assignment-3 • assignment-4 • assignment-5
Visuals.StoryBoard	 D-Wiki Character Storyboard Content Developer Tools Search Search January 2025 Groundbreaking Launch In January 2025, we launched a revolutionary feature that transformed the way users interact with our platform. This event marked a significant milestone in our mission to enhance user experience and engagement. February 2025 Major Update February 2025 brought a major update that introduced powerful new tools for content creation. This update empowered users to express their creativity and streamline their workflow, driving increased participation across our community.
Visuals.Content	 D-Wiki Character Storyboard Content Developer Tools Search Search Wiki Sidebar Live Updates Smart Analysis Organized Content Quick Summaries Live Updates Toggle Content + Smart Analysis Toggle Content + Organized Content Toggle Content + Quick Summaries Toggle Content + Reference Quality:

D-Wikis Algorithm 10/31/2024

	
Visuals.CharacterMap | | |—— (Needs improvement)	

Contents hide

(Top)

Overview and fundamental concepts

Mathematical formulation

Time evolution of a quantum state

Uncertainty principle

Composite systems and entanglement

Equivalence between formulations

Symmetries and conservation laws

Examples

Free particle

Particle in a box

Harmonic oscillator

Mach–Zehnder interferometer

Applications

Relation to other scientific theories

50 min listen 0:00 49:45

Quantum mechanics

135 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

For a more accessible and less technical introduction to this topic, see Introduction to quantum mechanics.

Quantum mechanics is a fundamental theory that describes the behavior of nature at and below the scale of atoms.[2]: 1.1 It is the foundation of all **quantum physics**, which includes quantum chemistry, quantum field theory, quantum technology, and quantum information science.

Quantum mechanics can describe many systems that classical physics cannot. Classical physics can describe many aspects of nature at an ordinary (macroscopic and (optical) microscopic) scale, but is not sufficient for describing them at very small submicroscopic (atomic and subatomic) scales. Most theories in classical physics can be derived from quantum mechanics as an approximation, valid at large (macroscopic/microscopic) scale.[3]

Quantum systems have bound states that are quantized to discrete values of energy, momentum, angular momentum, and other quantities, in contrast to classical systems where these quantities can be measured continuously. Measurements of quantum systems show characteristics of both particles and

Wave functions of the electron in a hydrogen atom at different energy levels. Quantum mechanics cannot

Appearance hide

Text: Small, [✓ Standard], Large

Width: [✓ Standard], Wide

Color (beta): Automatic, [✓ Light], Dark

D-Wiki Character Storyboard Content Developer Tools Search Search

🎮 Analyzing... Character Map of Skills and Attributes 🌟

John Doe

Skills
- Leadership
- Project Management
- Communication

Attributes
- Driven
- Analytical
- Team Player

Jane Smith

Skills
- Software Development
- Problem Solving
- Data Analysis

Attributes
- Creative
- Detail-Oriented
- Resilient

Emily Johnson

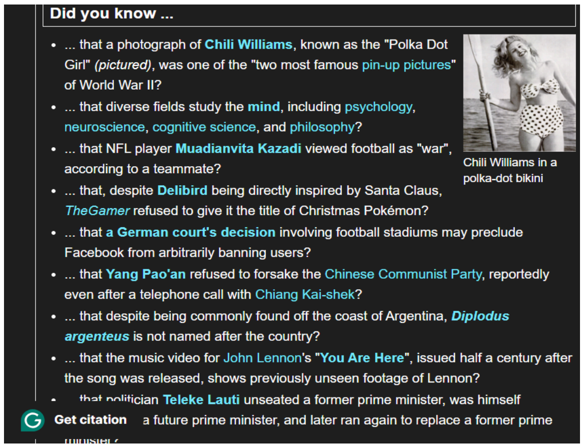
Skills
- Graphic Design
- UX/UI Design
- Marketing

Attributes
- Innovative
- Organized
- Collaborative

Features.	
Features.Animations	ChatGPT-like thinking ability to visually assemble webpage dynamically
Features.Dynamism:	Should feel dynamic and changing like discord but with little distracting noise. Eg: Live information updates
Features.Community-Feel:	The website should augment the community experience of discord. That is our audience we are developing for. It should feel like its separating the two.

D-Wikis Algorithm 10/31/2024

Cool Features To Add

Insight.Fun-Facts	 **Did you know ...** • ... that a photograph of **Chili Williams**, known as the "Polka Dot Girl" *(pictured)*, was one of the "two most famous pin-up pictures" of World War II? • ... that diverse fields study the **mind**, including psychology, neuroscience, cognitive science, and philosophy? • ... that NFL player **Muadianvita Kazadi** viewed football as "war", according to a teammate? • ... that, despite **Delibird** being directly inspired by Santa Claus, *TheGamer* refused to give it the title of Christmas Pokémon? • ... that **a German court's decision** involving football stadiums may preclude Facebook from arbitrarily banning users? • ... that **Yang Pao'an** refused to forsake the Chinese Communist Party, reportedly even after a telephone call with Chiang Kai-shek? • ... that despite being commonly found off the coast of Argentina, ***Diplodus argenteus*** is not named after the country? • ... that the music video for John Lennon's "**You Are Here**", issued half a century after the song was released, shows previously unseen footage of Lennon? • ... that politician **Teleke Lauti** unseated a former prime minister, was himself a future prime minister, and later ran again to replace a former prime Chili Williams in a polka-dot bikini
Addition.GhostLink	Ghost Links: Links to access websites our algorithm create. Would need means of temporary storage.
Addition.UserFeedback	Developing a good way to get user feedback on results to finetune system

Questions to Answer

Discord has a direct query search engine. What are some ways we can use it?	

D-Wikis Algorithm 10/31/2024

How can compression assist our algorithm? Instead of using raw words can we compress by using word ID’s?	
How can we speed things up with threading?	
Thinking about compression?	
What do we miss by treating this discord problem solely as search engine problem?	
Are there places where we can make use of hash based access? Especially indexing for storage and memory.	
How could knowledge graphs be applied to this?	