Journal of Data Analysis and Information Processing, 2026, 14(2), 247-253
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Full-Fidelity Semantic Aggregation: Navigating Datasets

Anthony Brian Mallgren

Independent Researcher, New York, NY, USA
Email: abmallgren@gmail.com

How to cite this paper: Mallgren, A.B. (2026) Full-Fidelity Semantic Aggregation: Navigating Datasets. *Journal of Data Analysis and Information Processing*, **14**, 247-253.
https://doi.org/10.4236/jdaip.2026.142012

Received: March 13, 2026
Accepted: May 9, 2026
Published: May 12, 2026

Abstract

Full-fidelity semantic aggregation presents significant advantages over utilizing publicly facing contemporary LLMs hosted by other organizations (e.g., cost, fewer required resources, interoperability, tail analysis, et cetera). Full-fidelity models can be trained with low-end/legacy hardware, become functional with virtually any dataset, and can be used in cross-dataset analysis. The model of truth approach that is generally made available to the general public seems to target a general audience by means of prescriptive heuristics, which involve issues related to dogma. It also lacks a more holistic familiarity through breadth in data. The approach outlined here exposes what the equivalent of neural network weights, or as it is labeled herein, conduciveness. It allows additional flexibility in assessments of heavily opinionated subjects. It allows analysis of divergence in variance and intensity. This method also opens a wide scope of innovation to occur in the academic world. The intent here is to provide a brief evaluation to prove the viability of the method.

Keywords

Data Analysis, Computational Semantics, Natural Language Processing, Unstructured Aggregation, Full-Fidelity Data

1. Introduction

Some technologists are focusing on utilizing models to produce what they term as artificial intelligence [1]. One method being utilized in an attempt to accomplish this is training large language models (LLMs). This requires substantial resources [2], which makes them heavily dependent upon what has thus far been corporate sponsorship. Additionally, this approach to data analysis consumes a significant amount of energy [3]. Some activity centers around constraining these models to make them perform as desired [4]. There are articles covering how these models may be utilized in data analysis [5].

There are a few assumptions that are being made. One, that data can be used primarily in a humanistic aspect, *i.e.*, that the primary goal is developmentally human centric. Two, that energy should be utilized as efficiently as possible to achieve this goal. And three, that democratizing data analysis of larger datasets, addressed more specifically herein, of unstructured datasets, is generally agreeable within and for society. These are the justifications behind the work presented here. Some advantages to the approach presented are that larger datasets can be wrangled and navigated with fewer resources, a more holistic understanding of a dataset is allowed, and dogmatic opinions formed by artificial intelligence companies are avoided.

What is presented herein is semantic aggregation. The examples presented are of linguistic datasets, though they could be utilized for other datasets.

2. Methods

2.1. Development

The resources required to set up a semantic aggregation data analysis environment are relatively low cost, or even free. In the most extreme data savings scenarios, a free environment could be used, such as those provided through https://ifastnet.com, or through provider partners such as https://infinityfree.com, https://aeonfree.com, etc. This allows free databases and web hosting. Due to some of the security restrictions and limited functionality in performing testing, a JavaScript agent was created to parse and process data. The intent of a security mechanism enforced in utilizing these services is that the web request is not cross-domain and is being executed in a browser which executes JavaScript. It is worth noting that batching is useful here, as the number of transactions is throttled. Also, occasionally, it may be necessary to file a ticket for review, as there are automated mechanisms that shut down services. If one can afford a few dollars a month (<$10), this opens access to providers that allow the use of expanded server-side execution, such as CRON jobs. This enables 24/7/365 processing and virtually unlimited sized datasets. If there is a larger budget, cloud providers can be utilized. Or, if private data suffices, and one has access to their own equipment, a local environment may be set up and utilized in developing aggregated maps.

Once an environment is settled upon, there is one method that has been found to be efficient in analyzing larger datasets; this can be termed the step-down method. A first table could simply keep the word count of the words occurring in a dataset and keep the count of occurrences in that dataset. For example, as partially shown in Table 1:

Table 1. Word occurrence count (conduciveness).

Word	Conduciveness
the	2,047,163
of	1,383,152

The next step-down table, as partially shown in Table 2:

Table 2. First expanded extension (may be dynamically named, e.g., "the").

Word	Following Word	Conduciveness	Stepdown Table
the	Secretary	125,050	the1
the	United	78,111	the2

And the next step-down table would be, as partially shown in Table 3:

Table 3. Second Expanded Extension (may be dynamically named with partition #, e.g., "the1").

Word	Following Word	Second Following Word	Conduciveness	Stepdown Table
the	Secretary	of	41,234	the_Secretary1
the	Secretary	shall	30,236	the_Secretary1

The stepdown method allows efficient indexing, allows partitioning, et cetera. The exact allocation and optimization methods for data restructuring will depend on the size of the dataset, the goals, et cetera.

This method allows a better understanding of larger datasets and enables aggregate navigation. A comparison method between datasets can be utilized. This can involve comparing heterogeneous datasets or time slices of homogeneous datasets. For example, a report may be, as partially shown in Table 4:

Table 4. Dataset comparison result.

Word	Following Word	Position in Dataset 1	Position in Dataset 2
Government	of	1	1
Government	or	2	4
Government	and	3	0
Governmental	Affairs	4	2
Government	to	5	5
Governments	and	6	6
Governmental	organizations	7	7
Government	in	8	0
Government	Act	9	10
Government	Accountability	10	8

Additional information can be added to the dataset. For example, parts of speech. This can be done with higher accuracy using a parts-of-speech tagger, or a dictionary approach can be used to label the possible parts of speech. This allows more targeted queries. Also, it allows heuristic approaches to data analysis. For

example, word substitution analysis. For example, government [noun], or government [verb]. This can be expanded and distilled into larger models. Comparisons can also be made. For example, the President [verb] versus the Secretary [verb], and compared as previously done with heterogeneous or homogeneous datasets.

2.2. Analysis

The analysis that was performed was simple. As previously mentioned, free platforms were assessed for their usefulness in getting started with such data analysis. It seems that anyone with access to a library may begin partaking in this type of analysis, though an active session is required to process data. Cell phones may be used to process data after authoring the scripts, but the tab generally goes inactive, and settings may need to be modified for continuous processing. As previously mentioned, requests are throttled, and as was implied, space is limited. Therefore, it may be seen as a way to get started and bide one's time, though not a means to an end (unless the datasets are relatively small).

The next level up, *i.e.*, low-priced shared hosting, which is typically affordable even on the lowest tier of government benefits, offers 24/7/365 processing possibilities and virtually unlimited-sized datasets. It is noteworthy that, as seems to be the nature with information technology systems, progression tracking and robust exception handling may make the experience more agreeable.

Cloud providers were not utilized in the analysis, though the same approach for progression tracking and exception handling seems relevant, as other systems seem to err occasionally during data collection. However, it is assumed that data processing is generally less error-prone.

2.3. Processing Performance

Processing the data takes some time. The typical tasks take up to a few days to complete. For example, parsing the United States Security Exchange Commission data for the 4th quarter of 2025 (financial filings downloaded from EDGAR on the SEC website; about 3.5 gigabytes of extracted data; please note, this does not include the extraction of the data; only processing the extracted data, which was narrowed down to business, risk factors, legal proceedings, mdna, market risk, and controls procedures in 10-K and 10-Q documents; 5785 files) took a few days on the iFastNet shared hosting platform (2.91 GB RAM, with 3 cores) to build a two-step down table system (2-word table; 3,646,416 rows with 246,093,330 occurrences, and 3-word table; 13,932,144 rows with 358,450,849 occurrences). Parsing the bills for a session of the United States Congress takes under a day (downloaded from the GovInfo website; https://www.govinfo.gov/bulkdata/BILLS/; 330 MB) for a two-step down table system (117th, Session 1, 12,648 files: 2-word table; 983,413 rows with 24,939,993 occurrences, and 3-word; 2,867,280 rows with 21,799,646 occurrences). Note that files were split into sections based upon a period or a newline, cleaned of punctuation, leaving letters, numbers, and hyphens, then split by spaces. No other func-

tions were performed. Adding in parts of speech data to a congressional session bill dataset via the dictionary method takes a few days, although some improvements are needed to the dictionary that was used (Moby Part of Speech List by Grady Ward). Additional abstractions can take a longer amount of time depending upon the complexity of the processing.

2.4. Usefulness

Even the simplest abstractions produce a significant amount of value and provide a foundation upon which more advanced functionality can be built. At the lowest level, the data provides accurate summarization in navigating datasets, allowing one to perform a full assessment of the dataset as one progresses through it. The conduciveness measures help to understand word and phrase density. Also, anomalies can be identified and readily assessed. Repeating data can also be found with relative ease. A simple webpage was used to load full results and typically returned in under a minute.

The comparison functionality is very useful. For homogeneous datasets, trends between different historical datasets may be identified. As shown above, positionality may be distinguished between two or more datasets. Conduciveness was included in the testing and gives a sense of the relative density of occurrences, which helps. Parts of speech can be included to ensure better filtering. For example, prepositions and conjunctions may be eliminated from the comparison above. Associating the parts of speech also allows entity comparison. For example, here is an example of verb sets based on a query for president, as partially shown in Table 5:

Table 5. Example query results.

Query	First Following Word	Second Following Word
President	shall	submit
President	or	tempore
President	shall	impose
President	shall	provide
President	shall	appoint
President	shall	establish
President	shall	not
President	shall	exercise
President	shall	transmit
President	shall	designate

As this example demonstrates, it helps to assess the actions that the president shall take. Based on this, word substitutability may be established. For example, submit is similar to impose, provide, appoint, et cetera.

This provides a basis that may be taken quite a bit further.

3. Discussion

The method outlined here presents a way to analyze and navigate through dataset aggregates and serves as the foundation for more complex tasks. This provides a low barrier to entry into data analysis of larger datasets. Momentum can be captured and extended into more advanced functionality. A database perspective has been presented here, but other abstractions, such as web-based, can be used for further enhancements. Return sets are compressed in a way that allows efficient navigation, even in a server/client model.

While others utilize data synthesis techniques which require new hardware, this method accommodates existing hardware and offerings. An approach such as this may be more appropriate for many scenarios. Rather than having tools serve as an author based upon the work of others, this method rather allows work to be explored, from which insights may be gained and used in true authoring. It seems that vector searches which pull along only one line of logic, this approach enables the user to gain a more thorough understanding of data, empowering them both in their development and creation.

4. Conclusion

Here it has been demonstrated how almost anyone may begin participating in semantic aggregation analysis and begin building a foundation for more complex models via a cost-effective method. Utilizing the full-fidelity approach enables more value to be extracted from datasets, lowers the amount of data needed to make a dataset analytically significant, requires less resources to perform the analysis, and provides the foundation for exciting innovation to occur. The democratization of such a practice could yield significant benefits in the field. The more people partaking in this field of study, the more advancements may be produced.

Future Work

There is much opportunity for innovation, including additional algorithms for data abstractions, domain-specific models, taxonomical standardization, data induction methods, versioned abstraction, integration, security, visualizations, graphing techniques, et cetera. While all of these areas of potential innovation are quite exciting, visualizations may become especially important in multidimensional analytical comparative progressions of datasets; for example, comparing a timelapse of two different datasets for specific models (e.g., weighted analysis of "president" [verb] [word] over the timespan of 1990-2020 for both the United States Security Exchange Commission and the United States Congress to understand how the tasks of presidents compared with each other, and to find likely candidates for future elections). With innovations in graphical rendering and interaction, analysis could become quite rich.

It was found that the technologies and basis for performing this very basic work were a start, but subpar. Some of the simplest things, like basic words such as the, or, and, et cetera, were mislabeled. Accurate dictionary databases are difficult to

come by in the public domain. Academic projects are going up and are being taken down quite frequently. Some of the intellectual property is proprietary. There are also a lot of opportunities for improvement even in the foundational aspects of this type of data analysis.

It was found that the technologies and basis for performing this very basic work were a start, but subpar. Some of the simplest things, like basic words such as the, or, and, et cetera, were mislabeled. Accurate dictionary databases are difficult to come by in the public domain. Academic projects are going up and are being taken down quite frequently. Some of the intellectual property is proprietary. There are also a lot of opportunities for improvement even in the foundational aspects of this type of data analysis.

Acknowledgement

Sincere thanks to the members of JDAIP for their professional performance, especially to editorial assistant *Delia Zhu* for achieving collaborative and facilitative excellence.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Copertari, L.F. (2019) On Natural and Artificial Intelligence. *OALib*, **6**, 1-9. https://doi.org/10.4236/oalib.1105221

[2] Duan, J., Zhang, S., Wang, Z., Jiang, L., Qu, W., *et al.* (2024) Efficient Training of Large Language Models on Distributed Infra-Structures: A Survey. arXiv:2407.20018.

[3] Jegham, N., Abdelatti, M., Koh, C.Y., Elmoubarki, L. and Hendawi, A. (2025) How Hungry Is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference. arXiv:2505.09598.

[4] Liu, B., Weng, Z. and Ayinhu (2026) The Advantages and Constraints of LLM-Augmented Knowledge Graphs in University Innovation and Entrepreneurship Education. *Advances in Applied Sociology*, **16**, 38-47. https://doi.org/10.4236/aasoci.2026.161003

[5] Wang, X., Tan, Y., Yang, T., Yuan, M., Wang, S., Chen, M., *et al.* (2024) Efficient Large Language Model Application Development: A Case Study of Knowledge Base, API, and Deep Web Search Integration. *Journal of Computer and Communications*, **12**, 171-200. https://doi.org/10.4236/jcc.2024.1212011