Parsifer.com

A novel model for “Linguistic Fingerprinting” of Digital Content Using Semantics

荃者所以在鱼，得鱼而忘荃	Nets are for fish; Once you get the fish, you can forget the net.
言者所以在意，得意而忘言	Words are for meaning; Once you get the meaning, you can forget the words

庄子(Zhuangzi), Chapter 26

Introduction

The growth of Large Language Models (LLMs) in recent years has spurred a proliferation of web companies focused on Information Retrieval that far surpasses the common keyword query methods employed by Google and other Search vendors. With the advent of ‘prompt engineering’ and targeted Natural Language Processing (NLP) tools to mine information from increasingly complex models, users are given a more tailored response, but one that is fraught with dangers: bias, racism, plagiarism, and so forth.

It is the latter problem, plagiarism, where legal conflicts enmesh players in a tangled web of attribution, provenance, and usage fees. To wit, did the response given constitute a new work of art, uniquely crafted by the logic and content of the LLM or is it a regurgitation of training data? And more importantly, was the training data sourced ethically? Was it appropriated under presumption of a business use case that violates the Terms of Usage, License, or other agreement under which owners of the content put their content and IP out on the web?

From this new legal scenario, many issues are just now being teased out and addressed on the open web of ideas. Here in this paper, Parsifer.com is addressing only two of the myriad of concerns that swamp the industry: Where is my content? How is my content being used?

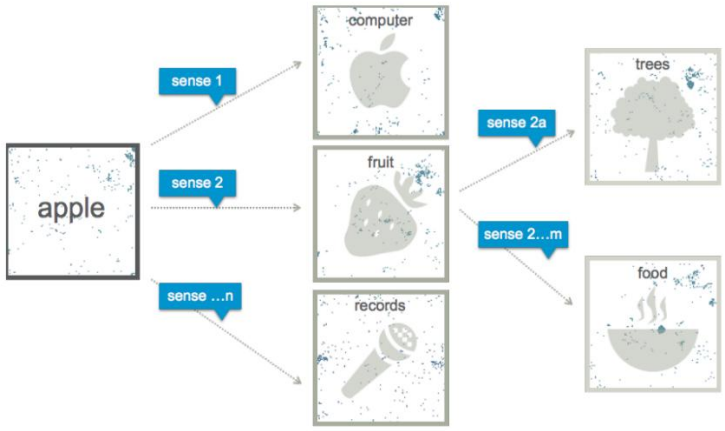
Background

Within the data sciences disciplines of parsing words, counting types of words (nouns, verbs, etc.) and determining their relationship to each other, statistics reigns as the methodology of choice when assigning importance to data. Collections of words (commonly referred to a 'bag of words') are ranked and organized into a view of the content and meaning of a document based on word density. Then other formulas are layered on top to give weighting, quality assessment, and ranking measures to the page. One the most prominent stylometric measures, Burrows's Delta, relies solely on statistics. Lost in the counting and measuring is the original context of the discourse. While statistics represent a significant step forward, statistics + semantics is far more powerful than statistics alone, offering a solution to restore context to the current tallying of nouns.

There currently exist standard toolsets to address this situation using semantics: Part of Speech (POS) tagging and distance (how many words between one noun and the other, such as Word2Vec) being the most popular. The Word2Vec method creates a basic two-dimensional representation of context. What is needed is a more n-dimensional representation to advance the art. In this proposal, I present a five-dimensional, layered method for creating a vector representation of content, or a 'linguistic fingerprint' (as sometimes referred to in research) using semantics instead of word counts.

What the industry calls a linguistic fingerprint is a vector representing the entity types (nouns) contained in a document. As such, semantic fingerprinting raises the analytics beyond the entity-level to the semantic, structural level overlaying the baseline context of entity usage. The method described herein combines the strength of Word2Vec and similar tooling, with the unique aspects of contextual semantics. This approach solves one piece of the foundational problem in Machine Learning (ML) today, namely that computational linguistics depends heavily on the statistics of word counts and virtually ignores the semantic aspects of grammar and structure.

With the method described herein, the corresponding semantic vector(s) produced are by orders of magnitude smaller than "bag-of-word" representations of the same content and, thus, more efficient to process. Utilizing entity-level information as a gateway to exploit sentence-level semantics, Semantic Fingerprinting is based on consecutive steps of understanding, depicted below graphically as a conceptual approach.



As the illustration shows, a message with identical syntax has a different semantic meaning depending on what (1st dimension, words), structure (2nd dimension, grammar), where (3rd dimension of location), when (4th dimension of time), and under what circumstance (5th dimension, manner), this message was created.

Item	Dimension	Method
1	Word	Entity (nouns) extraction, NER
2	Grammar	Part of Speech tagging, parsers
3	Location	Geolocation (physical), URL (virtual), Index (archival)
4	Time	Horological (physical), Timestamp (virtual)
5	Manner	Semantic framing: Context: Topic identification and theme generation Situation: Connotation, conceptual models (CNN) Sentiment: Sentiment analysis (positive, negative, neutral)

The uniqueness of Parsifer's approach is in the mapping of structure and context (2nd and 5th dimensions) alongside more standard approaches of word counts, location, and timestamps. While most researchers focus solely on Entity-based approaches (1st dimension) to create what they refer to as a semantic fingerprint, the richness of Parsifer's approach is that it layers speech patterns and emotions on top of entities to create a more fulsome heatmap of authorship to identify the creator of the work.

There is no requirement for an underlying ontology to drive classifications to a particular domain.[1] The fingerprint is based on the unique aspects of the data alone. Once created, the fingerprint graph is stored as logarithmic hash-code.

[1] Note that for improved efficiency, an ontological sub-categorization process may be employed, but is not necessary. This would represent one more piece of metadata, but not a controlling feature.

Word

Starting with a baseline of entity extraction to determine the trends, themes, and topics in a textual asset, the classification process sets up categories of words (or bags) that often fall into Person, Organization, Product, Place/Location, and other interesting groupings. This enables the later identification of major topics and themes for a document. The topic of a document is often able to be identified in the first two and the final paragraph.

To cluster the terms simply by density (or number of times reused in a document) gives a false weight to frequently used terms. Instead, it is more important to identify the concepts those words represent, and calculate the tokens based on meaning rather than mere spelling. (Concepts are defined as components of human cognition in the cognitive science disciplines of linguistics, psychology and philosophy.) This creates a feature set of categories (sometimes referred to as classes) that provide a schema for the document.

Grammar

Grammar, or the rules that define language structure, is a key component in both family of language recognition (is it Russian? Is it English?) and stylistic signature of an author. The method a Part of Speech (POS) tool uses to parse sentences is not just word order and selection, but repeating patterns of punctuation, amongst other tasks. For example, in this paper, Lee Vaughan is able to distinguish the works of Doyle from Wells just using semicolons. Repeating the use of adverbs in front of a verb as opposed to after a verb is another common signature habit. Co-occurrence refers to the distance between words in a sentence, creating phrases or "n-grams". For example, the bi-gram "hell hound" in Doyle's writing has a distance of 0, in that there are no words between the two. However the tri-gram "hell born hound" has a distance of 1 in that there is a single word between "hell" and "hound" as the nouns of interest.

Notably, the same syntax does not have to mean the same semantics. This simple fact is why the additional dimensions of location (3rd), time (4th) and manner (5th) are required to round out the fingerprint as truly semantic and not just word counts or co-occurrence. The inclusion of such pragmatic features in a text provides a factual basis for the contextual meaning of a text to be determined more precisely and qualitatively.

Location

Location has two aspects, one literal and the other digital. Extracting geolocation data from text can be as simple as a place name. The author can write about what is "in front" or "behind" an object or location. There can be an actual GPS tag to locate a picture in the real world. On the other hand, the URL where a work is posted is a virtual or digital location. The supposition of virtual and real-world data is used in AR to create 3-dimensional space.

Each of these elements adds up to metadata for a location of origin and creates a trail of movement from the first moment of publication to the final attribution of data within referrals, reuse, and provenance. A timeline of locations creates the lifetime value chain for the asset.

Time

Like Location, Time as the fourth dimension establishes original ownership, and patterns of posting, linking, and popularity of the asset. This is not so much a PageRank in the Google sense, but a chronology or paper-trail history. Time assists in establishing causal relationships such as staleness of data when looking at the edit history of an online post. Time also creates a beginning point from which to track the history of a work.

Manner

Manner, the way in which intent and context are conveyed, is comprised of three elements that semantically frame the subject matter:

> **Context:** Topic identification and theme generation are the background elements that set the stage. The domain of knowledge, industry, and cultural background all add to the context. These are critical to knowing "what is being addressed." Is it Mercury the God, Planet, or Car? Is it Apple the Fruit, Company, or Record Label?
> **Situation:** Connotation, conceptual models (CNN), and the like are able to derive the situation being discussed or pictured. Circumstances shift the semantic payload, changing the intent of communication.
> **Sentiment:** Sentiment analysis extracts indicators of emotion, which reveal a lot about a subject and the creator's attitude toward their topic. Beyond the three axis points of Positive, Negative, Neutral, there are a range of psychological states that can be identified: Happy, Sad, Angry, Confused, Joyful, and so on.

Methodology

A frequent assumption about machine learning is that, in a machine executing a sufficiently complex computational algorithm, intelligence would emerge and manifest itself by generating output indistinguishable from that of humans (a Turning Test as it were.) The process of binding a symbol, like a written or spoken word, to a conceivable meaning represents the fundamental semantic grounding of language.

To bridge the gap between ML and NLP, there needs to be an inference of intent. The theory goes: Given a significantly large enough mass of inputs (training data) the combinatorics of selective, predictive "next word" logic can produce an infinite number of outputs. Machines will

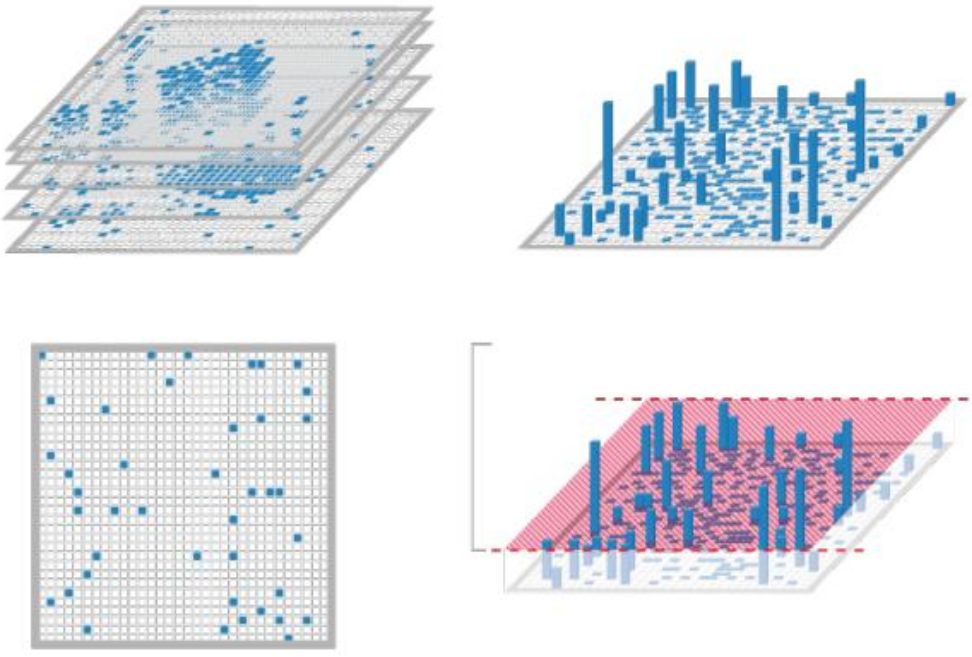
thereby be able to mimic natural language. In this formulation is an assumption of mind reading: What does the recipient really want to know? This is where Intent must be distilled from the query or prompt.

The inclusion of pragmatics (those 5 elements listed above) in the semantic and syntactic analysis of texts allows the contextual meaning of a text to be determined more precisely and qualitatively, therefore the base content can now also be generalized. This leads to Intent as a feature of the document that can be quantified.

Semantic Fingerprinting

The four steps of the semantic fingerprinting process are as follows:

1) Map the semantic space according to the five dimensions mentioned above
2) Calculate the classification for each concept and map to a 2-dimensional grid
3) Create vectors via layering of the 2-dimensional grids to derive a composite heatmap of the elements. This graph is essentially the semantic fingerprint of a single document. The fingerprint graph is stored as logarithmic hash-code.
4) Layer each document-level fingerprint into a heat map to derive the author-level composite fingerprint over time.

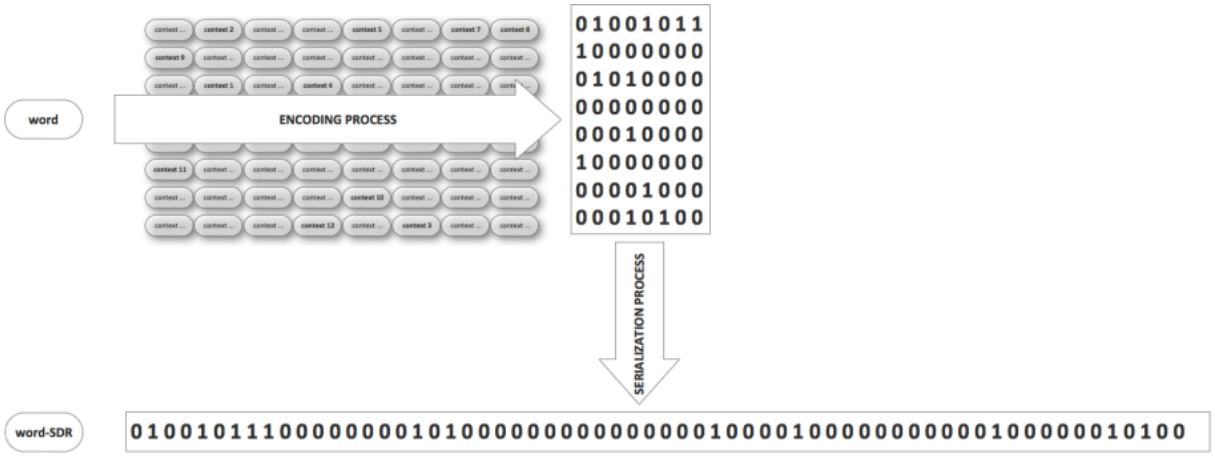


After capturing a given semantic universe from a reference set of documents for an author by means of the fully unsupervised learning mechanism described above, the

resulting semantic space is linked to each and every word-representation vector, not just nouns, but including adjectives as properties of nouns, verbs connected to adverbs as properties of verbs, and so forth accounting for each POS. These vectors are large, sparsely filled binary vectors. Every feature bit in this vector not only corresponds to but also equals a specific semantic feature of the document and is therefore semantically grounded.

Conversion to a Binary String

How to convert language from its symbolic representation (text) into an explicit, semantically grounded representation that can be generically processed by Hierarchical Temporal Memory (HTM theory of Jeff Hawkins) networks? Take those 2-dimensional tables produced in step 2 above and the resulting heat maps from step 3, with their resulting word vectors, and using Sparse Distributed Representation[2] (SDR), convert the semantic representations into valid SDR format. This is computed using standard methodologies.



The main advantage of using the SDR format is that it allows any data items to be directly compared. In fact, it turns out that by applying Boolean operators and a similarity function, such as Euclidian Distance, many NLP operations can be implemented in a very elegant and efficient way.

Many practical problems of statistical NLP systems, like the high cost of computation, the fundamental incongruity of precision and recall, the complex tuning procedures etc., can be overcome by applying semantic fingerprinting.

[2] https://discourse.numenta.org/t/sparse-distributed-representations/2150

The set of all possible word-SDRs corresponds to a word-vector-space. By applying a distance metric (like Euclidian distance) that represents the semantic closeness of two words, the word-SDR space satisfies the requirements of a metric space:

- Distances between words are always non-negative: Words that occur in similar contexts tend to have similar meanings. This link between similarity in how words are distributed and similarity distributional in what they mean is called the distributional hypothesis.
- If the distance between two words is 0 then the two words are semantically identical (perfect synonyms).
- If two words A and B are in a distance d from each other, d(A,B) = d(B,A). (Symmetry).
- For three distinct words A, B, C we have d(A,C) <= d(A,B) + d(B,C). (Triangle inequality).

There are two different semantic aspects that can be detected while comparing two Semantic Fingerprints:

- The absolute number of bits that overlap between two fingerprints describes the semantic closeness of the expressed concepts.
- By looking at the topological position where the overlap happens, the shared contexts can be explicitly determined.

With Semantic Fingerprints, there is no need to train the standard classifiers. The only mechanism needed is a reference fingerprint that specifies an explicit set of semantic features describing a class. This reference set or semantic class skeleton can be obtained either through direct description by enumerating a small number of generic class features and creating a Semantic Fingerprint of this list (for example the three words “mammal” + “mammals” + “mammalian”), or by formulating an expression. By computing the expression: “tiger” AND “lion” AND “panther”, a Semantic Fingerprint is created that specifies **big cat** features.

The development of these features and their fingerprints also may be applied to documents of an unknown or artificial origin. By comparing the semantic fingerprint of an unknown document to a document by a known author that has been analyzed, the attribution or provenance of the unknown work can be reestablished with a high degree of certainty.

The invention can be applied to a number of problem spaces: piracy, plagiarism, authorship attribution, IP protection, Brand and Trademark protection, Code authorship attribution, and evidence authentication. The applications for Publishing, Education, Legal, Law Enforcement, and Software are numerous.

Advantages:

Simplicity

1. No Natural Language Processing skills are needed.
2. Training of the system is fully unsupervised (no human work needed).
3. Tuning of the system is purely data-driven and only requires domain expert review by a human for quality control and no specialized technical staff.
4. Ability to create an API that is very simple and intuitive to utilize.
5. The technology can be easily integrated into larger systems by incorporating the API over REST or by the inclusion of a plain Java library with no external dependencies for local (Cloudera/Apache Spark) deployments.

Quality

1. Rich semantic feature set allows a fine-grained representation of concepts.
2. All semantic features are self-learned, thus reducing semantic bias in the language model used.
3. The descriptive features are explicit and semantically grounded and can be inspected for the interpretation of any generated results.
4. By drastically reducing the vocabulary mismatch, far less false positive results are generated.
5. If aligned concept semantic spaces for different languages are used, the resulting fingerprints become language-independent. This means that an English fingerprint can be directly matched with an Arabic fingerprint. When filtering text sources, the filter criterion can be designed in English while being directly applied to all other languages. An application can be developed using the English assets while being deployed to a Chinese one.

Speed

1. Encoding the semantics in binary form (instead of the usual floating-point matrices) provides orders of magnitudes of speed improvement over traditional methods.
2. All Semantic Fingerprints have the same size, which allows for an optimal processing pipeline implementation.
3. The semantic representations are pre-calculated and therefore do not affect the query response time from LLMs or other systems.

4. Once in binary format, the algorithms only apply independent calculations (no corpus relative computation) and therefore easily scale to any performance needed.
5. Potential to further improve search results by increased performance gains. In a document search context, providing a stable query response time of < 5 microseconds, independently of the size of the searched collection, reduces compute overhead and reduces energy consumption.

References

Sparse Representation, Modeling and Learning in Visual Recognition: Theory, Algorithms and Applications Hong Cheng, University of Electronic Science and Technology of China Chengdu, Sichuan, China (2012)

Evolutionary Cognitive Neuroscience Edited by Steven M. Platek, Julian Paul Keenan, and Todd K. Shackelford, The MIT Press, Cambridge, Massachusetts, London, England (2006)

Content-Addressable Memories T. Kohonen 2nd Edition, Springer Series in Information Sciences, Editors: Thomas S. Huang Manfred R. Schroeder (1989)

Foundations of Language - Brain, Meaning, Grammar Ray Jackendoff, Evolution-Oxford University Press, USA (2003)

Connections and Symbols (Cognition Special Issue) Steven Pinker, The MIT Press; 1st MIT Press edition (1988)

The Language Instinct - How the Mind Creates Language Steven Pinker, Perennial (HarperCollins) (1995)

Linguistics as a Window to Understanding the Brain, Steven Pinker, Harvard University, 2012, retrieved from https://youtu.be/Q-B_ONJIEcE

Sparse Distributed Representations (SDRs), Subutai Ahmad, Numenta, 2017, retrieved from https://discourse.numenta.org/t/sparse-distributed-representations/2150

Modeling Data Streams Using Sparse Distributed Representations, Jeff Hawkins, Numenta, 2012, retrieved from https://youtu.be/iNMbsvK8Q8Y