Finch for Text

Turn Documents into Data Points

Finch for Text: Software that Reads

An entity extraction and disambiguation engine from Finch COMPUTING

> Structured data analytics can describe and explain what's happening, and unstructured data analytics can explain why it's happening. Together you get the whole picture. **Without both, you're half blind.**

Forbes, 03/05/15, Unstructured Data: The Other Side of Analytics

Introducing... Finch for Text

Table of Contents

Introduction

Data is being generated at record levels, yet the pace at which the world creates data will never be this slow again.

And much of this new data we're creating is unstructured, textual data. Emails. Word documents. News articles. Blogs. Reviews. Research reports…

Understanding what's in this text – and what isn't, and what matters – is critical to an organization's ability to understand the environments in which it operates. Its competitors. Its customers. Its weaknesses and its opportunities.

In recent years, a number of text analysis solutions have been developed and have come to market. But these solutions have forced customers to choose. Speed or scale. Accuracy or ease of use. Customizability or interoperability.

Finch for Text is an entity extraction and disambiguation solution that employs natural language processing, sophisticated statistical models and other heuristics to isolate and extract eight distinct entity types from unstructured text.

Lots of it. Very quickly. And very accurately.

Finch for Text identifies: people, places, organizations, cyber entities, IP addresses, phone numbers, currency values, and dates and times (including ranges). And, by taking in all of the relevant context surrounding these entities, it correctly distinguishes between same-named or similarly named entities.

Finch for Text draws on a rich IP portfolio, heavy in text analysis and knowledge discovery, to offer: extreme accuracy and precision; easy customization, installation and use; and the ability to write new applications utilizing its JSON outputs.

Most important, it gives organizations of any size and type greater awareness and insight into the mountains of information they collect and possess. It allows them to be certain of the value of that information. And it allows them to better and more fully capitalize on their unstructured information assets.

> "[Forrester] survey data shows that, on average, enterprises leverage about 35% of their structured data for insights and decision-making, but only 25% of their unstructured enterprise data."
>
> *Information Management: Expand Your Big Data Capabilities with Unstructured Text Analytics*

Finch COMPUTING

ex • trac • tion	Isolating an entity in text
type-match • ing	Determining an entity's type. (Is it a person, a place or a company, etc.?)
res • o • lu • tion	Correctly determining that multiple, differently spelled or differently represented references all refer to the same entity
en • rich • ment	Tagging an isolated entity with metadata, which is simply additional information about that entity, also known

dis • am • big • u • a • tion

Distinguishing between identically named entities of the same type via unique identifiers. This John Roberts or that one? The city of Alexandria in Virginia, Louisiana or Egypt?

Finch COMPUTING

Entity Disambiguation and Why It Matters

Understanding Finch for Text's power, and ultimately its value, begins with the concept of entity disambiguation. Not entity enrichment, entity type-matching, or entity resolution.

Disambiguation demands building algorithms that can – in an instant – process text and understand whether it is referring to George Washington the president, or George Washington the university … or the town of George, Washington … or the George Washington Bridge in New York City.

Disambiguation also means distinguishing between *identically named entities of the same type*. For example, John Roberts, the Chief Justice of the U.S. Supreme Court, or John Roberts, the Fox News correspondent. The images below provide an example.

Image 1

Image 2

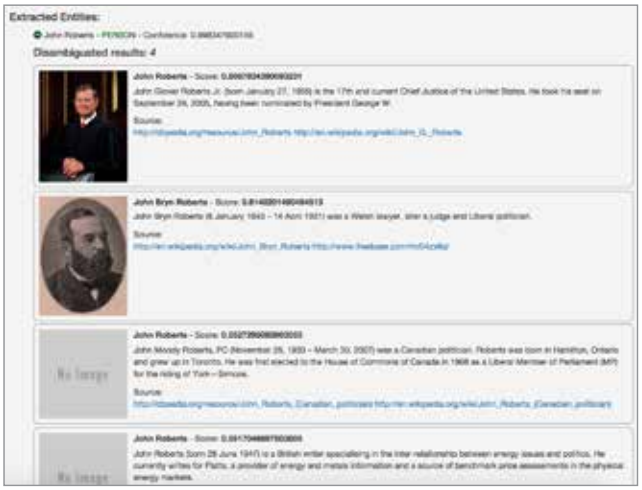


Image 3

- **(Image 1)** The input text on the left yields the tagged (highlighted) text on the right. Note the multiple entity types tagged; the implied references (northwest Indiana, Rehnquist); and how those references are tagged using context (e.g. The Supreme Court as an organization not a place).
- **(Image 2)** Hovering on a highlighted entity produces a snapshot of disambiguation candidates, ranked and assigned scores indicating accuracy. In this case, Chief Justice John Roberts, a person, is the correct entity match, hence the green entity-type indicator and the 99% confidence score at the top.
- **(Image 3)** A Google search yields the correct image of the John Roberts to which the text refers; but the first listing is for an insurance agent in Ohio.

Finch COMPUTING

Our Unique Approach to Entity Disambiguation

Of course, quality entity disambiguation depends on quality extraction, and Finch for Text excels there as well. We take a big data approach to both extraction and disambiguation – and by that we mean that we take into account all of the surrounding context around an entity to correctly disambiguate it.

We leverage our high-throughput, in-memory computing platform, FinchDB, to enable the processing of massive amounts of text – on the order of 233 disambiguation queries performed per second, *on a four-node cluster of standard servers in the cloud.*

Using this powerful platform, we were able to tell our engineers and modelers – those developing the algorithms necessary to take in multiple factors to disambiguate an entity – to remove speed from the equation completely. Not to worry about it. Instead, we told them to write the longest, most complex, most correct code possible. And that we'd put the code in-memory to produce the kind of speed and accuracy customers want.

Massive Amounts of Data

Sophisticated Algorithms

\+ In-Memory Computing Platform

Disambiguation without Tradeoffs

This approach allows us to perform true knowledge discovery on the part of our customers. And it means that our models can get better the more often they're run. Our knowledge bases can get bigger the more content we process.

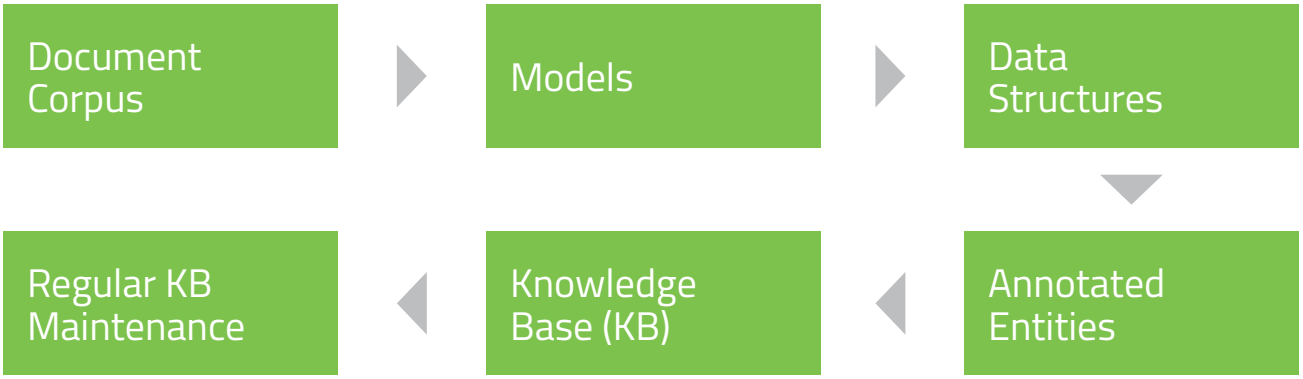
Additionally, we're able to do in-field customizations for our customers. Custom knowledge bases, custom models – tuned and tailored to their industries and lexicons. And these can also get better and better over time.

Finch COMPUTING

To build the kind of long and complex algorithms that enable Finch for Text to correctly determine which unique entity is being mentioned, we had to first build giant knowledge bases.

Doing that requires starting with text. Lots and lots of text. This is called a document corpus. Once Finch for Text reads that corpus, our engineers develop models that teach it how to understand nuances in language that provide clues about an entity's correct identity. As in the John Roberts example on page 3, does the text mention judges, The Supreme Court and laws? Or does it mention television news, reporting and cable television? These features, describing a particular entity, help to distinguish it from others with which it shares a name or part of a name.

Once those models are developed, related data structures are created. This enables Finch for Text to identify patterns in text that offer clues about an entity's type (is it a person, place, organization, etc.) and its resolved identity. The resulting outputs are a set of annotated entities that comprise a knowledge base – adept at reading the specific type of content in the document corpus. Regular additions and maintenance make the knowledge base more complete and more accurate over time.



As an example of this process, to correctly disambiguate people mentioned in a streaming corpus of news documents, our team produced a person knowledgebase nearly 500 million features describing 3 million people. This specific knowledge base was "trained" on a news corpus. Other knowledge base creation efforts can employ the same process, the same sophisticated algorithms and the same training to deliver the kind of accuracy and precision that users demand.

> "Algorithms and natural language generators have been around for a while, but they're getting better and faster as the demand spurs investment and innovation. The sheer volume and complexity of the Big Data we generate, too much for mere mortals to tackle, calls for artificial rather than human intelligence to derive meaning from it all."
>
> *New York Times, 03/07/15: If An Algorithm Wrote This, How Would You Even Know?*

Finch COMPUTING

Finch for Text: An Overview

Finch for Text was developed based on years working with customers who needed to understand geographic references in massive amounts of streaming text. And, before that, on our technical leadership team's experience with a major global information service provider.

As described earlier in this document, Finch for Text is a reflection of this experience – combined with new intellectual property and an entirely different approach to understanding nuance and context in language. Paired with speed and industry-leading accuracy, as detailed later in this document, Finch for Text extracts and disambiguates multiple entity types, allowing our customers to interact with information in entirely new ways.

These entity types include:

People

Finch for Text's person-based extraction approach takes into account attributes such as gender, first names, middle names and last names; titles (like "Judge" or "Dr."), generational suffixes (like "Jr." or "III"), professional suffixes (like "M.D.") and descriptive suffixes (like "Board Member" or "Trustee").

Places

Finch for Text extracts and disambiguates geographic names including cities, provinces, states, and countries – even bridges, monuments, oil fields or street addresses – from unstructured text. For each name, it extracts latitude and longitude, class (road, populated area, vegetation, etc.) and enclosing region (where applicable).

Phone Numbers

Globally, telephone numbers are governed by a specification known as a numbering plan. Using this specification, Finch for Text identifies and extracts North American phone numbers whether or not they are abbreviated; include parenthesis around area codes; or include dashes, dots or spaces between the various groupings of numbers.

Currency Values

Finch for Text extracts currency values based on the identifying attributes surrounding a numeric digit. These include symbols (like $, €, £, ¥, etc.) numeric amounts, text-amount modifiers (words hundred, thousand, etc.), currency terms (like dollar, euro, etc.), country names, ISO 4217 alphabetic currency codes, decimal representations of amounts, as well as specific, user-defined patterns.

Cyber Entities

Finch for Text also identifies cyber-related entities like email addresses, host names and Twitter IDs, even when there are deviations from the standard format, e.g. – spaces between characters in an email address or semicolons between multiple addresses. Finch for Text also allows for parsing, so users can find all email addresses from a single domain, for example.

IP Addresses

Finch for Text extracts Internet Protocol version 4 (IPv4) addresses based on the RFC 791 internet protocol specifications in dotted decimal, dotted hexadecimal or dotted octal notation.

Organizations

Finch for Text categorizes "organizations" as either companies and businesses, criminal organizations, educational institutions, governments, performing organizations (like The Metropolitan Opera), and sporting organizations (like the New York Yankees).

Dates, Times and Date-Time Ranges

Finch for Text can extract dates, times, and date and time ranges. This includes multiple date formats: 16 October, 2012; October 16, 2012; 10/16/12; or the sixteenth of October, etc. Time can be expressed as Standard Time (11:11, 11:11 PM, 11:11 p.m., 23:59:59, 23:59:59.00Z); Military Time (1800 hours, 0500 hrs or 0930GMT); or OCLOCK Time (5 o'clock, 10 o'clock, 3 PM). Ranges can also be expressed in myriad ways, such as: "the 1970s" or "1995 through 2011" or in days or months, such as "February - March 2014" or "August through February," etc.

Finch COMPUTING

Regular Expressions

Adding to Finch for Text's customizability, is the solution's ability to capture "regular expressions" in text, or RegEx entities.

For example, suppose a supply chain manager for a textile manufacturer wanted to identify references to all of the company's red dye suppliers in text. He or she could extract the word "red" as well as the word immediately after it. Think: "red fibers" or "red pigment" or "red hue."

The user would simply install a RegEx pattern configuration in Finch for Text. To do it, he or she would tell the system to find the word "red," set confidence thresholds, describe the entity (in this case "red") and set the extraction pattern (in this case, red + one word after it).

```
"red" : {
    "entity_type" : "RED",
    "confidence" : "0.75",
    "description" : "A pattern to identify references to red",
    "pattern" : "(red\\s+)(\\w+)"
}
```

Then, if the document under review included the sentence: *"Our dyes are developed to deliver a rich red pigment and the truest red tones," Finch for Text would return "red pigment" and "red tones,"* indicating to the user that this document contains information of value to their search for content on their red dye suppliers.

Users find this to be an incredibly valuable feature – because this capability is completely under their control and not bound to a specific entity type.

Additionally, users can also use the RegEx feature to identify and extract things like **Social Security numbers, well-head numbers, product numbers, product names, event names, and more.**

Finch COMPUTING

Performance Metrics

The primary and most commonly used metric for evaluating text analysis solutions is the F-Score, which is the average of two other measures of performance: precision and recall.

- **Precision** refers to the percentage of retrieved results that are correct. e.g. How much of what is returned is relevant.
- **Recall** is the percentage of relevant results that are retrieved. e.g. Of all the possible relevant results, how many were returned?

Essentially, precision and recall are measures of quality. (Precision and recall measure extraction quality; disambiguation quality is measured by precision alone.) And the quality of returned results depends on the corpus of data being analyzed.

Every Finch for Text deployment begins with a knowledge base-building and model-training exercise to ensure optimal performance. Below are the extraction and disambiguation results that Finch for Text is able to return on varied data corpora – from news and government data, to industry-specific data, to various large and publicly available datasets.

Extraction

	Precision	Recall	F-Score
People	93.5%	91.8%	92.6
Places	91.8%	93.0%	92.4
Organizations	96.9%	90.6%	93.6

Disambiguation

	Precision
People	89.6%
Places	94.1%
Organizations	91.8%

**Disambiguating other entity types, with the exception of cyber entities, is not necessary given that they're primarily numeric values. We do, as noted on page 6, correct and resolve these entity types so that they can be enriched and yield greater value.*

Anecdotally, customers tell us that F-scores in the 80s are ideal – and hard to replicate with other products or custom-built solutions.

We deliver 90+ F-scores across all entity types.

Finch COMPUTING

Finch for Text Case Studies

Geotagging 85,000 Documents in 45 Minutes

A team of five geologists, working on behalf of a major global energy company spent two years manually identifying the geographical references in more than 3,000 geology reports. 82,000 reports still remained unprocessed. Finch for Text processed the outstanding reports – and reprocessed the completed ones – *in just 45 minutes* – and with greater precision and accuracy than those done manually.

Turning One Man-Month to One Hour

A customer in the federal intelligence community needed to identify the textual, geographic references within its entire content archive. The process was estimated to take one man-month to achieve, and an additional significant effort to verify. Finch for Text did it *in under an hour*, serving up valuable insights to an agency for whom accurate, timely information is critical.

Preventing Costly Mistakes

An energy customer was preparing to drill a test well, deploying significant capital and manpower in the process. A scan of its massive geological content library revealed the company had *already* drilled a test well in the exact location decades earlier. Quick and easy access to this type of information, buried in volumes of text, saved the company millions. Meaning Finch for Text more than paid for itself.

Finch COMPUTING

Why Buy?

Fast and Accurate

We firmly believe that misinformation is worse than no information. Accuracy matters. And the algorithms we've developed to govern Finch for Text's approach are a reflection of that belief. Once models are trained, we produce precision and recall scores in the 90s – far above the industry standard and enabled by the massive knowledge bases we've built. We leverage our expertise in in-memory computing to process disambiguation queries exceptionally quickly. On a streaming feed of news (800,000 documents per day), we're able to process an impressive 233 disambiguation queries per second.

Customizable

Finch for Text is at its best when it is customized for end-users' needs; from training our models for a particular industry or lexicon, to offering user-defined regular expressions, to custom settings, and white and black lists governing its classifications. Its tuning and auto-tuning features are customizable, and prove valuable in environments with a high number of requests per second, where responsiveness and timeout errors can persist. Its logging and reporting features can also be configured to user preferences and desired specifications.

Easy to Use and Install

Finch for Text's minimal hardware requirements are: 16 Gigabytes of RAM, 4 CPU cores, and approximately 50 gigabytes of disk-space. Out of the box, it runs on either Redhat 6.X or the CentOS 6.x operating system, but can be configured to run on Windows, Ubuntu, Solaris, BSD or Mac OS X. Finch for Text ships with Python scripts to assist in installation and removal. It can also be used via a cloud-based RESTful API for testing purposes and is available as a hosted solution. Finch for Text was designed to enable the rapid, accurate processing of large amounts of unstructured text and to create JSON outputs upon which developers can write new applications.

Finch COMPUTING

A Look Ahead: Unstructured Data in the Enterprise

Enterprise data managers, by and large, will freely admit to lacking a coordinated strategy to effectively understand their unstructured text resources. This problem is becoming more apparent – and more serious – as data volumes grow and as more and more enterprise content is unstructured. Industry analysts, press coverage and anecdotal evidence all support this premise.

IDG estimates that by 2022, 93% of all data in the digital universe will be unstructured. Gartner says global data volumes are set to grow 800% between now and 2020, and 80% of it will reside as unstructured data. Enterprises are huge contributors to this data deluge.

Finch for Text serves as the foundation for an unstructured text strategy within the enterprise.

The primary users of a solution like Finch for Text include data managers, scientists and analysts – not limited to one industry or sector. These professionals, long familiar with traditional BI tools, are largely just beginning to explore text analytics solutions that complement their existing data strategies.

Widely adopted enterprise analytics solutions like Tableau and QlickView offer these users traditional BI dashboards, but the tools' underlying technology prevents them from assessing unstructured text in a meaningful way.

Other tools, like NetOwl, are capable of managing unstructured text and claim to do entity disambiguation as well. However, in reality they're often only resolving entity types (determining whether an entity is a person versus a place, or versus an organization, etc.). Unlike Finch for Text, most do not, and cannot, distinguish specific, same-type entities from one another. Our proprietary in-memory computing platform, FinchDB, when coupled with Finch for Text, enables this capability.

Customers in need of text analysis solutions will find immense value in Finch for Text because it is not only capable of producing highly accurate results across a variety of entity types, it can handle extremely large data volumes and integrate with existing solutions easily.

> "Did you know 90% of the world's data was created in the last two years? There's going to be 10 times more mobile data by 2020, 19 times more unstructured data, and 50 times more product data by 2020,' [said Salesforce CEO Marc Benioff]."
>
> *Business Insider, 10/14/14: In 2 Sentences, Salesforce.com CEO Marc Benioff Explains Why Something Called 'Analytics' Is The Hot New Thing*

Finch COMPUTING

About Finch Computing

Finch Computing, formerly Synthos Technologies, is a division of Qbase, LLC. Together, we build and support new ways of interacting with information – via three innovative products that address complex and never-before-addressable big data needs at various points in the software stack.

FinchDB

FinchDB is an in-memory computing platform with embedded analytics that is changing the expectations of database technology. Part database, part search engine it enables radically different data experiences.

Finch for Text

Finch for Text is an entity extraction and disambiguation engine. It reads free-form text as a human would, extracting eight distinct entity types and disambiguating them against massive knowledge bases. It turns documents into data points and is the foundation for an effective unstructured text strategy in the enterprise.

Finch Analyst

Finch Analyst is an end-to-end data discovery solution that enables customers across a variety of industries and use cases to find greater meaning and insight from data. Whether it's streaming or static. Internal or external. Words or numbers.

We believe the search tools of today are insufficient. We understand that the rate at which the world creates information will never be this slow again. And we know that analytics, including predictive analytics, are going to become a larger and larger part of every professional's job.

Finch Computing enables dramatically different data experiences. And meets an intensifying market need for a better, faster, more accurate picture of the environments in which our customers operate.

Finch COMPUTING

Finch for Text

Contact Us

sales@finchcomputing.com

Washington, DC

12018 Sunrise Valley Drive
Suite 300
Reston, VA 20191
+1 888 458 0345 *toll free*

San Francisco, CA

28 Second Street
Floor 3
San Francisco, CA 94105
+1 415 314 7110

Beavercreek, OH

3800 Pentagon Boulevard
Suite 110
Beavercreek, OH 45431
+1 937 521 4200