Finch for Text

Software that Reads and Reasons

A Powerful Entity Extraction & 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*
Finch for Text reads freeform text as a human would. Lots of it.

It’s the foundation for an unstructured text strategy in the enterprise.
An Introduction

Finch for Text is a patented entity extraction and disambiguation service that employs natural language processing, sophisticated statistical models and other heuristics to extract **15 distinct entity types** from unstructured UTF-8 documents.

**It identifies:** people, places, organizations (businesses, governments, criminal organizations, educational institutions, sports and performing organizations, and news and social media outlets), cyber entities, IP addresses, phone numbers, currency values and, dates and times (including ranges).

It features **custom dictionaries** and **regular expressions** capabilities and is available as an API or on-premise solution.

By taking in all of the relevant context surrounding these entities, the API version of this product **correctly distinguishes** between same-named or similarly named entities.

Finch for Text draws on **a rich IP portfolio** to offer:

- Extreme accuracy and precision.
- Easy customization, installation and use.
- The ability to write new applications utilizing its JSON outputs.
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-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.
A Demonstration

The input text, at right yields the tagged text at the far right. Note the multiple entity types tagged; the implied references (northwest Indiana, Rehnquist); and how those references are tagged using context (ex: Supreme Court as an organization not a place).

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.

At the far right, 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 for Text’s **person-based extraction and disambiguation method** takes into account attributes such as:

- Gender
- First, Middle and Last Names
- Titles (“Judge” or “Dr.”)
- Generational Suffixes (“Jr.” or “III”)
- Professional Suffixes (“M.D.” or “J.D.”)
- Descriptive Suffixes (“Board Member” or “Trustee”)

Precision = 93.5%
Recall = 91.8%
Finch for Text’s sophisticated, algorithmic models also allow it to **correctly extract and disambiguate organizations**.

“Organizations” are defined as:

- Businesses
- Criminal Organizations
- Educational Institutions
- Governments
- News & Social Media Organizations
- Performing Organizations (ex: The Metropolitan Opera)
- Sporting Organizations (ex: The New York Yankees)

**Precision = 96.9%**  
**Recall = 90.6%**
Finch for Text extracts and disambiguates geographic names including cities, provinces, states, and countries.

For each name, it extracts:

- Latitude and Longitude Coordinates
- Class (roads, populated areas, vegetation, etc.)
- Enclosing Region (where applicable)

Precision = 91.8%
Recall = 93.0%
Cyber Entities

Finch for Text can also identify cyber-related entities like email addresses, host names and Twitter IDs, even when there are deviations from the standard format, such as spaces between characters in an email address or semicolons between multiple addresses.

• The extraction of hostnames is based on RFC 1123 standards.

• The extraction of a Twitter ID is based upon Twitter’s username specifications: the @ sign plus 1-15 word-characters.
IP Addresses

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

An IPv4 entity consists of four octets, each separated by a dot (.), and may contain a port.

We examine each of these attributes, as well as normalized text and user patterns.
Phone Numbers

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
- Include the national or international trunk prefixes of 1 and 011, respectively
- 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 (the words hundred, thousand, etc.)
- Currency terms (like dollar, euro, etc.)
- Country names
- ISO 4217 alphabetic currency codes
- Decimal representations of amounts
- Specific, user-defined patterns
Dates and Times, and Date-Time Ranges

Finch for Text can also extract both dates and times, and date-time ranges. This includes:

- **Dates:** 16 October, 2012; October 16, 2012; Oct. 16; 10/16/12; 10/16/2012; 16/10/12; the sixteenth of October, etc.

- **Time:** 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 O’CLOCK Time (5 o’clock, 10 o’clock, 3 PM).

- **Ranges:** “the 1970s” or “1995 through 2011” or simply “1997-08”. Ranges can also appear in days or months, such as “February - March 2014” or “August through February” or “Nov 10 – Dec 10,” etc.
Three Examples

- **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. And Finch for Text more than paid for itself.
Where We Stand Out
Fast & Accurate

We firmly believe that misinformation is worse than no information. And the algorithms we’ve developed to govern Finch for Text’s approach were developed with speed and accuracy in mind.

- 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 233 disambiguation queries per second.

- 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.
Customizable

Finch for Text is at its best when it is customized for end-users’ needs...

Finch for Text models can be trained for a particular industry or lexicon, and it offers user-defined regular expressions, custom dictionaries and 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

Finch for Text’s **minimal hardware requirements** are: 8 Gigabytes of RAM, 2 CPU cores, and approximately 25 gigabytes of disk-space.

It requires either **Redhat 6.X or CentOS 6.x operating system**.

Finch for Text **ships with Python scripts** to assist in installation and removal.

It 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**.
Contact Us

sales@finchcomputing.com