



SAIC's nLing

A tailored search and data acquisition solution

nLing – Easily Deployed, Highly Interoperable, Text Mining Solution

nLing, developed by Science Applications International Corporation (SAIC), is a software product for text mining. In less than 30 minutes, nLing can be installed onto a network server and users can begin collecting data. Data sources can include Web sites (HTML pages, RSS feeds, Microsoft SharePoint sites), relational database systems (Oracle, MySQL, PostgreSQL, others), and file systems (captured media, "hot folders", etc.). Any of these data acquisition activities can be scheduled to occur at once or periodically (e.g., daily, hourly, monthly). Incoming data can be processed by the bundled named entity extractor, so that instances of person, location, facility, equipment, ip_address, url, and other entities will be tagged in the index.



nLing

Carolyn Copenhaver

434.989.7989

The Process

Once the initial collection process has been completed, users are able to use the search capability in their analytical processes. Experienced *Pathfinder* users will appreciate that nLing supports the rich query syntax they have come to expect in a data mining tool,



Future Enhancements

- Faceted browsing
- Recognizing and highlighting of “statistically improbable terms”
- Tagging/annotation of documents (tags/ annotations can be queried)
- Notifications/alerts based on standing queries (delivery via IM, RSS, email)
- Document level security model
- HTML form-based authentication for Web crawls
- Crawl the client's file system
- User provided relevance feedback incorporated into query expansion
- “Perl Crawl” facilitating flexible mapping of semi-structured data formats

while novice users can begin by using the Google-like queries they are already familiar with. The search results are returned in order of relevance to the user's query with query terms highlighted in the text. Data security is provided at the index level, so users are granted access based on their group membership. Interoperability with other systems was a key design consideration; hence, nLing is able to authenticate users against Lightweight Directory Access Protocol (LDAP) or existing systems in addition to its own data store (a relational database management system).

Features

- Client/server architecture – server currently deployable on Linux, Windows, and Solaris®, client runs on any operating system (Java® 6 JVM required)
- “Query by Example” – find similar documents
- Foreign language capability – Unicode used throughout, pluggable language models enhance foreign language capability (BasisTech modules)
- Vector space model (VSM) of relevance – top documents returned first in list
- Named entity recognition/extraction
- Pluggable “crawler” framework – currently supporting two Open Source products – Nutch and Aperture®
- Query syntax – Boolean operators, wildcards, proximity, phrases

Interoperability

- Export to XML and KML (for Google Earth)
- Index/database format is shared by *Pathfinder* – can immediately augment capabilities of existing *Pathfinder* systems
- Authenticate via LDAP, *Pathfinder* Web services, or internal data store
- All capabilities exposed as Web services for easy integration with other COTS and GOTS tools; scripting via languages such as Perl and Python; others as needs arise

Tailored Solutions

- ESRI ArcGIS plug-in – ArcGIS users can install a plugin, log in to their nLing server, submit queries and access the results without ever leaving their application. Seamless integration with ArcGIS was achieved by developing a plug-in client in C# .NET against nLing's Web services API.
- Web proxy/cache – intercept network traffic, indexing all Web-based objects in nLing for analysis. This solution also permits non-attribution through tunneling to chained proxy/cache systems.

Carolyn Copenhaver, *Pathfinder* Program Manager

mobile: 434.989.7989 | email: copenhaverca@saic.com

Michael Goddard, *Pathfinder* Technical Director

tel: 434.964.3034 | email: goddardm@saic.com

Visit us online at www.saic.com/products/software/pathfinder

Energy | Environment | National Security | Health | Critical Infrastructure

