



Unlock Content™



An SVM Classifier for XML

Kelly Stirman

Director of Business Development

- An XML Content Platform

- A platform for building content applications
 - Content Integration
 - Content Analytics
 - Content Enrichment
 - Content Delivery
 - Search

- A platform with many, many tools
 - Including an XML Classifier

How Did We Get Here?

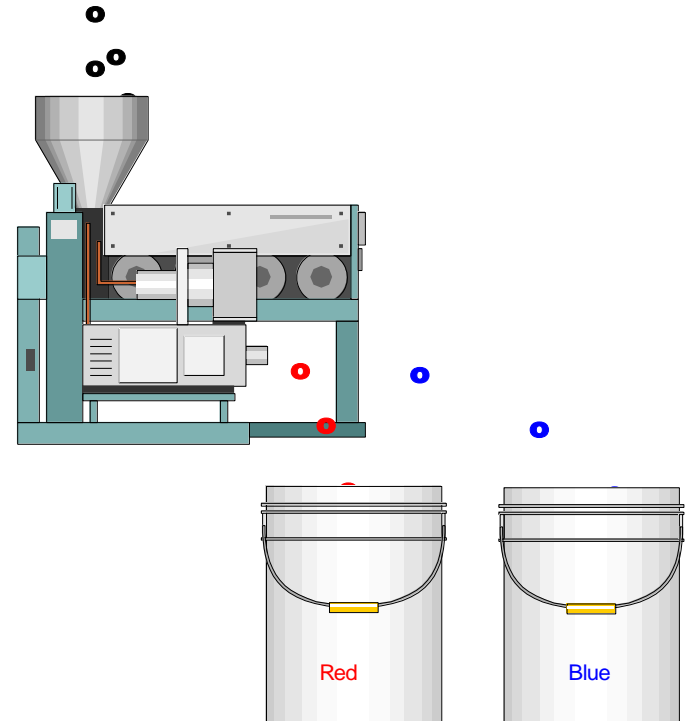
- The answer to a hard problem:
 - Rich query language
 - Guaranteed correctness
 - Transactions

- Turns out to be a database
 - ACID compliant
 - High availability (clustering, journaling, fail over, etc)
 - XQuery interface

- Built with search engine principles
 - Automatic indexing
 - Term lists for text *and* structure

What is a Classifier?

- Associates items to classes
 - Eg, which categories does this document belong to?
- The classes are known
 - Compare to *clustering*, where the classes are inferred
- Some applications
 - Assigning metadata
 - Empowering navigation
 - Routing/alerting



Rules Based

- Explicit rules defined for each class
- Costs primarily incurred in rule creation, testing, maintenance
- Are your people better at writing rules

Statistics Based

- Positive and negative examples provided for each class
- Costs primarily incurred in finding good examples
- Or finding good examples?

What is an XML Classifier?

- Classifies **nodes** not documents
 - Assign granular metadata
 - Exclude or include whatever you wish
 - Assemble a node dynamically from multiple sources

- Classifies based on text **and** structure
 - Titles, figure captions, footnotes, etc
 - Context of the terms is meaningful

- Classification is an **extension** to XQuery
 - In the middle of a query – read or update
 - As part of rendering

Text

- Words
- Stems
- Word Pairs
- Diacritic sensitivity
- Case sensitivity
- 1/2/3 character
- Language

Structure

- Element/attribute name + namespace
- Element/attribute value
- Hierarchical context
- Phrase-around (eg, footnote)
- Phrase-through (eg, style)
- Collections
- Directories
- Security
- Fields

- SVM Classifiers have a reputation of being “black magical”
- Fields provide more “control” to the end user
- A field in MarkLogic
 - Defined as one or more nodes, and their descendents
 - Include/exclude criteria
 - Weights per node
 - Field-specific indexes

Included Elements

Localname(s)	Namespace	Attribute	Attribute Namespace	Value	Weight
title	http://marklogic.com/namespace-examples/foo				7 [delete]
keywords	http://marklogic.com/namespace-examples/foo				15.0 [delete]
abstract	http://marklogic.com/namespace-examples/foo				3.0 [delete]
section	http://marklogic.com/namespace-examples/foo				1.0 [delete]
footnote	http://marklogic.com/namespace-examples/foo				1 [delete]

Excluded Elements

Localname(s)	Namespace	Weight
references	http://marklogic.com/namespace-examples/foo	[delete]

Example: Wikipedia

Log in / Create account

[article](#) [discussion](#) [edit this page](#) [history](#)

April 4

From Wikipedia, the free encyclopedia

April 4 is the 94th day of the year (95th in leap years) in the [Gregorian calendar](#). There are 271 days remaining until the end of the year.

Contents [\[hide\]](#)

- 1 [Events](#)
- 2 [Births](#)
- 3 [Deaths](#)
- 4 [Holidays and observances](#)
- 5 [External links](#)

<< **April 2008** >>

Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			
MMVIII						

Events [\[edit\]](#)

- **1581** - [Francis Drake](#) completes a circumnavigation of the world and is knighted by [Elizabeth I](#).
- **1655** - The miraculous statue entitled the [Infant of Prague](#) is solemnly crowned by command of Cardinal [Harrach](#).
- **1660** - [Declaration of Breda](#) by [King Charles II of England](#).
- **1721** - [Sir Robert Walpole](#) enters office as the first [Prime Minister of the United Kingdom](#) under [King George I](#).
- **1812** - U.S. President [James Madison](#) enacted a ninety-day embargo on trade with the [United Kingdom](#).
- **1814** - [Napoleon](#) abdicates for the first time.
- **1818** - The [United States Congress](#) adopts the [flag of the United States](#) with 13 red and white stripes and one star for each state (then 20).
- **1841** - [William Henry Harrison](#) dies of [pneumonia](#) becoming the first [President of the United States](#) to die in office and the one with the shortest term served.

April 4 in recent years

- 2008 (Friday)
- 2007 (Wednesday)
- 2006 (Tuesday)
- 2005 (Monday)
- 2004 (Sunday)
- 2003 (Friday)
- 2002 (Thursday)
- 2001 (Wednesday)

Example: Configure

Classify Train **Configure**

Classifier URI

Classifier Type and Kernel

Type:

weights

supports

Kernel

- sqrt
- sqrt-normalized
- linear-normalized
- gaussian
- geodesic

Database search options

word case-sensitive range-element-attribute-indexes

stemmed diacritic-sensitive range-element-indexes

one-character phrase phrase-arounds

two-character element-word phrase-throughs

three-character element-phrase element-word-query-throughs

trailing-wildcard element-character field

element-trailing-wildcard

Numeric options

max-terms

max-support

min-weight

tolerance

epsilon

...

Sessions

Example: Wikipedia

Content

[Refresh](#)

Allah
Alexios III Angelos
Azerbaijan/People
AppliedStatistics
Argot
Abydos, Hellespont
Tsarevich Alexei Petrovich of Russia
April 4
April 30
Afrika Islam
Andriscus
Aldous Huxley
Alexander of Battenberg
AlbaniaCommunications
AxiomOfChoice
Andronicus of Cyrrhus
Aberdeen (disambiguation)
Albert Alcibiades, Margrave of Brandenburg-Kulmbach
AlbaniaHistory
August 2

Results



Science

This is a very poor "Science" example.
It is a distance of 59.7 out of 100 outside the class.



Biography

This is a very poor "Biography" example.
It is a distance of 57.9 out of 100 outside the class.



History

This is an excellent "History" example.
It is a distance of 91.6 out of 100 inside the class.



Geography

This is a terrible "Geography" example.
It is a distance of 74.6 out of 100 outside the class.



Culture

This is a very poor "Culture" example.
It is a distance of 56.2 out of 100 outside the class.

Example: Genre Classification of Books



- **Application:** accurate automatic genre classification

- **Problem:**
 - Whole book too big
 - First paragraph misleading

- **Solution:**
 - Synthesized document includes TOC, Reviews, Summary, Metadata
 - Parts can be dynamically defined
 - Granular assignment of classes – not just the whole book

- **Application:** user-specific sentiment analysis of financial news

- **Problem:**
 - Analysts and editors cannot agree on “positive”, ”neutral”, ”negative”
 - Sentiment more granular than articles

- **Solution:**
 - User-defined classifiers, iterative process
 - **Or**, use the classifier of your favorite analyst
 - Classify individual sentences or paragraphs, order by sentiment strength
 - Classify at render time, color-code for sentiment



Unlock Content™



Thank You

Kelly Stirman
kelly.stirman@marklogic.com
t: 267.496.2759