



Sentiment Analysis

in

Open Source Information

for the

US Government

Ronald P. Reck (SAP) &
Kenneth B. Sall (SAIC)
XML-in-Practice 2008
December 9, 2008
Arlington, VA

Outline



- **Problem Statement**
- **Data Source Collection and Preprocessing**
- **Redland RDF Framework (open source)**
- **Inxight ThingFinder (commercial)**
- **Name Catalogs**
- **Processing Steps**
- **Query Redland using SPARQL**
- **Conclusions**
- **Next Steps**

Problem Statement



- **Detect sentiment expressed toward government officials in news articles.**
- **Large volume of articles to process.**
- **Subtleties of natural language.**

Data Source Collection and Preprocessing



- Wanted to use Kapow for screen scraping.
- Due to time constraints, we used static Reuters news articles, dated 1996-08-20 to 1997-08-19 available from NIST.
- 806,794 news stories: 3,316 MB of XML
- Sentiment analysis results: 457,696,507 RDF assertions, 36,635 MB of disk space.

Sample Reuters news article

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <newsitem itemid="257631" id="root" date="1996-12-15" xml:lang="en">
  <title>USA: Senators raise doubts on CIA nominee Lake.</title>
  <headline>Senators raise doubts on CIA nominee Lake.</headline>
  <byline>Patricia Wilson</byline>
  <dateline>WASHINGTON 1996-12-15</dateline>
- <text>
  <p>Republicans on the Senate Intelligence Committee said on Sunday they were troubled by President Bill Clinton's nomination of Anthony Lake to head the CIA and declined to predict he would be confirmed.</p>
  <p>Sen. Richard Shelby, the panel's incoming chairman, denied a suggestion that Republicans had chosen Lake, currently the White House national security adviser, to be their "whipping boy" among Clinton's second-term Cabinet nominees.</p>
  <p>"Any nominee, including Mr. Lake, should be one with the highest integrity above everything. And some past conduct troubles me, as it troubles a lot of other people in the Senate," Shelby said on the ABC programme "This Week."</p>
  <p>The Alabama Republican was referring to Lake's role in keeping secret a 1994 White House decision that the United States would not object when Croatia allowed Iranian arms to be shipped through its territory to Bosnia.</p>
  <p>Lake also faces questions about possible conflicts of interest because he failed to take timely action on orders to sell off his personal holdings in energy stocks.</p>
  <p>Shelby said the hearings would be "searching and thorough" and the committee would "scrutinize" both issues. He added, "It's too early to predict what will happen."</p>
  <p>Sen. Arlen Specter of Pennsylvania, the outgoing Republican Intelligence Committee chairman, said in a separate interview on CNN's "Late Edition" that Lake's confirmation was "up in the air."</p>
  <p>Clinton has defended Lake, saying nothing disqualified him from serving as CIA director. White House Chief of Staff Leon Panetta said on Sunday that Lake had made "a simple oversight" in the stocks matter and conceded it probably would have been better to work more closely with Congress on the arms issue.</p>
</text>
  <copyright>(c) Reuters Limited 1996</copyright>
- <metadata>
  - <codes class="bip:countries:1.0">
    - <code code="USA">
      <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-12-15" />
    </code>
  </codes>
  - <codes class="bip:topics:1.0">
    - <code code="GCAT">
      <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-12-15" />
    </code>
    - <code code="GCRIM">
      <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-12-15" />
    </code>
    - <code code="GPOL">
      <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-12-15" />
    </code>
  </codes>
  <dc element="dc.date.created" value="1996-12-15" />
  <dc element="dc.publisher" value="Reuters Holdings Plc" />
  <dc element="dc.date.published" value="1996-12-15" />
  <dc element="dc.source" value="Reuters" />
  <dc element="dc.creator.location" value="WASHINGTON" />
  <dc element="dc.creator.location.country.name" value="USA" />
  <dc element="dc.source" value="Reuters" />
</metadata>
</newsitem>
```

Reuters Metadata Not Processed



- Itemid
- Headline
- Byline
- Dateline
- Copyright
- Attribution
- dc.date.created
- dc.publisher
- dc.date.published
- dc.source
- dc.creator.location
- dc.creator.location.country.name
- dc.source

Redland RDF Libraries



- Redland is a set of free software C libraries that provide support for RDF by Dave Beckett; open source.
 - **Raptor** RDF Parser Toolkit for parsing and serializing RDF syntaxes (RDF/XML, N-Triples, Turtle, RSS tag soup, Atom)
 - **Rasqal** RDF Query Library for executing RDF queries using SPARQL or RDQL.
 - **Redland RDF Library** provides the RDF C API and triple stores; includes Raptor and Rasqal
 - **Redland Language Bindings** for C#, Java, Objective-C, Perl, PHP, Python, Ruby and Tcl.

Inxight ThingFinder



- **Commercial; owned and distributed by SAP.**
- **ThingFinder provides advanced text analysis technology that automatically identifies and extracts key entities or user-defined "things" from any text data source.**
- **35 pre-defined key entities (e.g., people, dates, places, companies, etc.).**
- **No setup or manual creation of rules required, other than user-defined entities.**

Name Catalogs – user-defined entities

Name Catalog Concept	List Size	Part of Speech	Example Instances in Reuters Corpus
Administrative offices	11	Nouns	President of the United States, Vice President of the United States
Bush Cabinet	22	Nouns	Attorney General, Secretary of Defense
Clinton Cabinet	15	Nouns	Vice President Al Gore, Secretary of State Madeleine Albright, Secretary of Defense
Intelligence agencies	17	Nouns	National Reconnaissance Office, Federal Bureau of Investigation
Negative adjectives	104	Adjectives	troubled, weak, bad, dangerous
Cabinet positions	15	Noun	[none]
U.S. Department of Defense components	200	Nouns	U.S. Army, U.S. Air Force, U.S. Navy, Northern Command
Positive Adjectives	435	Adjectives	confident, powerful, hopeful, strong
Negative Admire Verbs	16	Verbs	regret, fear, deplore, loathe
Positive Admire Verbs	61	Verbs	support, respect, favor, trust

Example: Name Catalog for Administrative Offices (excerpt)

```
<?xml version="1.0" encoding="UTF-8" ?>
- <catalog name="US_administrative_offices" source="Catalogs\administrative_offices.input.xml">
  <entity_category name="US_administrative_offices">
    - <entity_name canonical="President of the United States">
      <variant name="The President" />
    </entity_name>
    - <entity_name canonical="Vice President of the United States">
      <variant name="Vice President" />
    </entity_name>
    - <entity_name canonical="White House Chief of Staff">
      <variant name="Chief of Staff" />
    </entity_name>
    <entity_name canonical="Administrator of the Environmental Protection Agency" />
    <entity_name canonical="Director of the Office of Management and Budget" />
    <entity_name canonical="Director of the National Drug Control Policy" />
    <entity_name canonical="United States Trade Representative" />
    - <entity_name canonical="Chairman, Board of Governors of the Federal Reserve System">
      <variant name="Chairman of the Federal Reserve" />
      <variant name="Federal Reserve Chairman" />
    </entity_name>
    <entity_name canonical="Director of National Intelligence" />
  </entity_category>
</catalog>
```

Processing Steps - 1



- Extract information from Reuters Corpus CDs supplied by NIST: 806,794 news stories from 1996 and 1997.
- Configure Inxight ThingFinder; Task used load balancing software called Pen.
- Enhance ThingFinder CSV to produce CSV+ (see paper for details).
- Extraction of the 33,780,421 entities from the 806 thousand files took less than a day to complete.

Processing Steps - 3

■ Convert CSV+ files to RDF.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:sap="http://iama.rrecktek.com/ont/sap#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/" >

<!-- stuff deleted -->

<sap:entity rdf:about="http://inright.rrecktek.com/entity./tf-257631newsML_xml.csv+/16">
  <sap:Type>PERSON</sap:Type>
  <sap:Name>President Bill Clinton</sap:Name>
  <sap:Occurrences rdf:datatype="http://www.w3.org/2001/XMLSchema#int">4</sap:Occurrences>
  <sap:Method>Unique</sap:Method>
  <sap:Confidence rdf:datatype="http://www.w3.org/2001/XMLSchema#int">10</sap:Confidence>
  <sap:Relevance rdf:datatype="http://www.w3.org/2001/XMLSchema#int">68</sap:Relevance>
  <sap:Offset rdf:datatype="http://www.w3.org/2001/XMLSchema#int">224</sap:Offset>
  <sap:Length rdf:datatype="http://www.w3.org/2001/XMLSchema#int">22</sap:Length>
  <sap:ParagraphNum rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</sap:ParagraphNum>
  <sap:SentenceInPara rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3</sap:SentenceInPara>
  <sap:SentenceInDoc rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3</sap:SentenceInDoc>
  <sap:SubEntityCount rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3</sap:SubEntityCount>
  <sap:SubEntityType>PERSON_POS</sap:SubEntityType>
  <sap:SubEntityName>President Bill Clinton</sap:SubEntityName>
  <sap:SubEntityMethod>Unique</sap:SubEntityMethod>
  <sap:SubEntityOffset rdf:datatype="http://www.w3.org/2001/XMLSchema#int">224</sap:SubEntityOffset>
  <sap:SubEntityOffset rdf:datatype="http://www.w3.org/2001/XMLSchema#int">224</sap:SubEntityOffset>
</sap:entity>
```

Processing Steps - 4

■ Name catalog RDF case:

```
<sap:entity rdf:about="http://inright.rrecktek.com/entitytf-25763lnewsML_xml.csv+/66">
<sap:Name>excellent</sap:Name>
<sap:Occurrences rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</sap:Occurrences>
<sap:Method>Name Catalog</sap:Method>
<sap:Confidence rdf:datatype="http://www.w3.org/2001/XMLSchema#int">20</sap:Confidence>
<sap:Relevance rdf:datatype="http://www.w3.org/2001/XMLSchema#int">10</sap:Relevance>
<sap:Offset rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1863</sap:Offset>
<sap:Length rdf:datatype="http://www.w3.org/2001/XMLSchema#int">9</sap:Length>
<sap:ParagraphNum rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</sap:ParagraphNum>
<sap:SentenceInPara rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</sap:SentenceInPara>
<sap:SentenceInDoc rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</sap:SentenceInDoc>
<sap:SubEntityCount rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</sap:SubEntityCount>
</sap:entity>
```

Processing Steps - 5



- Enter data into Redland RDF data store.
 - Took more than a week due to size of source.
 - Entering one result file at a time might be streamlined by entering multiple RDF result files in a single pass.
- Use SPARQL to query the RDF data.

Query Redland - 1

Find all NEGATIVE statements (adjectives or verbs).
UNION version; more general solution but NOT IMPLEMENTED in Redland.

```
PREFIX sap: <http://iama.rrecketk.com/ont/sap#>
SELECT ?ent ?name
WHERE {
    {
        { ?ent sap:Type "negative_adjectives" } UNION { ?ent
sap:Type "negative_admire_verbs" }
    }
    ?ent sap:Name ?name
}
```

Query Redland - 2

Find all NEGATIVE statements (adjectives or verbs).
Regex version.

```
PREFIX sap: <http://iama.rrecktek.com/ont/sap#>
SELECT ?ent ?type ?name
WHERE {
    ?ent sap:Type ?type . FILTER regex (?type, "negative", "i")
    ?ent sap:Name ?name
}
```

Query Redland - 3

Find all articles from named catalogs.

It produces 21,732 hits (all the catalog ones).

```
PREFIX sap: <http://iama.rrecktek.com/ont/sap#>
```

```
SELECT ?ent ?type ?name
```

```
WHERE {
```

```
    ?ent sap:Method "Name Catalog" .
```

```
    ?ent sap:Type ?type .
```

```
    # Next line needed only if we want to display the negative  
    sentiment term.
```

```
    ?ent sap:Name ?name
```

```
}
```

Query Redland - 4

All 10 name catalogs with full, explicit catalog names; 21,732 hits.

```
PREFIX sap: <http://iama.rrecktek.com/ont/sap#>
SELECT ?ent ?type ?name
WHERE {
    ?ent sap:Type ?type . FILTER regex (?type,
    "Bush Cabinet|Clinton Cabinet|cabinet_positions|Intelligence
    Agencies|DOD_components|US_administrative_offices|negative_adjectiv
    es|negative_admire_verbs|positive_adjectives|positive_admire_verbs",
    "i")
    ?ent sap:Method "Name Catalog" .
    ?ent sap:Name ?name
}
```

Query Redland - 5

More compact query taking advantage of regex and catalog name patterns.

Same 21,732 hits.

```
PREFIX sap: <http://iama.rrecktek.com/ont/sap#>
```

```
SELECT ?ent ?type ?name
```

```
WHERE {
```

```
  ?ent sap:Type ?type . FILTER regex (?type,  
    "cabinet|Intelligence
```

```
Agencies|DOD_components|US_administrative_offices|negative|positive",  
  "i")
```

```
    ?ent sap:Method "Name Catalog" .
```

```
    ?ent sap:Name ?name
```

```
}
```

Query Redland - 6 - Sample Output



**ent=[http://inxight.rrecktek.com/entity/339336newsML_xml/25]
type=negative_adjectives
name=hurt**

**ent=[http://inxight.rrecktek.com/entity/694847newsML_xml/20]
type=DOD_components
name=U.S. Army**

**ent=[http://inxight.rrecktek.com/entity/599085newsML_xml/30]
type=Clinton Cabinet
name=Vice President Al Gore**

Query Redland - 7



- **Incomplete query; difficult learning curve.**
- **Need to apply substring to entity URIs to determine the news article's URIs.**
- **Upon retrieval of article headline and date our desired output would include:**
 - The article date & headline**
 - The sentence that contained the entity and sentiment term as a link to full text of article**
 - The sentiment term(s)**
 - The entity & named catalog**

Conclusions



- RDF, SPARQL, and Inxight's ThingFinder enable government intelligence and defense agencies to detect positive or negative sentiment when open source information refers to their agencies and officials.
- It is difficult to deal with hundreds of thousands of files as each step for iterating over the files requires the creation of a program or script.
- The default memory space that is allocated for Linux command line argument is not sufficient.
- The learning curve for SPARQL was more steep than anticipated; few good tutorials.

Next Steps



- Use Kapow to screen scrape current news stories from variety of sources.
- Leverage metadata from original articles.
- RDF should be applied at the sentence level.
- Improve name catalogs (acronyms, name variations, etc.)
- Use rule-based pattern matching engine.
- Perfect SPARQL queries.

Resources



- **Inxight ThingFinder:**
<http://www.inxightfedsys.com/products/sdks/tf/default.asp>
- **Redland RDF Libraries:** <http://librdf.org/>
- **SPARQL Query Language for RDF:**
<http://www.w3.org/TR/rdf-sparql-query/>
- **Reuters Corpus available from NIST:**
<http://trec.nist.gov/data/reuters/reuters.html>
- **Kapow:** <http://www.kapowtech.com/>
- **The slides, demo, and a related paper, see RReckTek:**
<http://iama.rrecktek.com/conferences/XML-in-Practice2008/>