

GETTING FROM RAW DATA TO CONTENT ANALYSIS WITH CUSTOM PROGRAMS

JASON T. KILEY
UNIVERSITY OF GEORGIA

Overview

- * Goal: high-level familiarization with data issues where programming (Python 3.x) can help.
- * Retrieve the structure from semi-structured text.
- * Identify texts of interest within a larger set.
- * Output data in appropriate forms for content analysis.

Semi-structured text

- * **Defined:** Text that has a structure that cannot be used directly.
- * Convert that semi-structure to a form we can use.
- * Examples: LexisNexis (press releases, news), web pages, recent EDGAR documents.
- * Isolate the text we want to analyze, and use the metadata.

PR Newswire US

December 20, 2006 Wednesday 3:34 PM GMT

Seven Summits Research Releases Comments on HPQ, HES, A, MXIM, and FDO

LENGTH: 496 words

DATELINE: CHICAGO Dec. 20

CHICAGO, Dec. 20 /PRNewswire/ -- Seven Summits Research releases NewsBites on key stocks.

Seven Summits Strategic Investments NewsBites are available to all investors at <http://www.go7now.com/nb.aspx?p=1220Z> (Note: You may have to copy this link into your browser then press the [ENTER] key.)

Today's Seven Summits Strategic Investments NewsBites cover the following stocks: Hewlett-Packard Co. (NYSE:HPQ), Hess Corporation (NYSE:HES), Agilent Technologies Inc. (NYSE:A), Mentor Integrated Products Inc. (NASDAQ:MXIM), and

PRESS RELEASE 1

Web site: <http://www.sevensummitsstrategicinvestments.com/>

SOURCE Seven Summits Investment Research

URL: <http://www.prnewswire.com>

LOAD-DATE: December 21, 2006

LANGUAGE: ENGLISH

PUBLICATION-TYPE: Newswire

Copyright 2006 PR Newswire Association LLC.
All Rights Reserved.

4 of 118 DOCUMENTS

PR Newswire US

PRESS RELEASE 1 & 2

Regular expressions

- * **Defined:** a way of specifying search patterns.
- * Any given regex has a binary outcome depending on whether it matches: true or false.
- * In some cases, this true or false result is enough.
- * We can also specify that we want to capture a part of the matching text to use.

Identifying interesting texts

- * Archival raw data is often full of texts that we are not interested in.
- * Some data is naturally noisy.
- * Other times, over-gathering is easier if we can automate the elimination of uninteresting texts.
- * Human labor is smart but expensive and slow; computers are fast but only as smart as you can make it.


```






























j_dict = {'jbl':-
..... {'contact': re.compile('@jabil\.com')},-
..... 'jci':-
..... {'source': re.compile('Johnson Controls')},-
..... 'jcp':-
..... {'contact': re.compile('J.{,2}C.{,2}Penney')},-
..... 'jnj':-
..... {'source': re.compile('Johnson & Johnson$'),-
..... 'contact': re.compile('Johnson(?: |\\n)&(?: |\\n)Johnson')},-
..... 'jny':-
..... {'source': re.compile('Jones Apparel')},-
..... 'jpm':-
..... {'contact': re.compile('J.{,2}P.{,2}Morgan|'-
..... '@jpmorgan\.com|@jpmchase\.com|'-
..... 'jpmorganchase\.com')}-
..... }-
-

```

REGULAR EXPRESSIONS

Output

- * Forms
 - * Individual files: LIWC, comparisons.
 - * Spreadsheets: coding, statistical analysis.
- * Testing and project management: gather descriptive statistics, diagnostics.

Name ▲	Date Modified	Size	Kind
 acls_20020114_366.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20020122_365.txt	Mar 19, 2014, 2:43 PM	11 KB	text
 acls_20020123_364.txt	Mar 19, 2014, 2:43 PM	2 KB	text
 acls_20020129_360.txt	Mar 19, 2014, 2:43 PM	1 KB	text
 acls_20020318_347.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20020522_335.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20020604_330.txt	Mar 19, 2014, 2:43 PM	2 KB	text
 acls_20020903_319.txt	Mar 19, 2014, 2:43 PM	1 KB	text
 acls_20020904_317.txt	Mar 19, 2014, 2:43 PM	4 KB	text
 acls_20020925_315.txt	Mar 19, 2014, 2:43 PM	4 KB	text
 acls_20021016_311.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20021017_310.txt	Mar 19, 2014, 2:43 PM	4 KB	text
 acls_20021023_308.txt	Mar 19, 2014, 2:43 PM	16 KB	text
 acls_20021115_306.txt	Mar 19, 2014, 2:43 PM	1 KB	text
 acls_20021220_297.txt	Mar 19, 2014, 2:43 PM	2 KB	text
 acls_20030107_296.txt	Mar 19, 2014, 2:43 PM	4 KB	text
 acls_20030108_295.txt	Mar 19, 2014, 2:43 PM	1 KB	text
 acls_20030113_294.txt	Mar 19, 2014, 2:43 PM	2 KB	text
 acls_20030207_289.txt	Mar 19, 2014, 2:43 PM	1 KB	text
 acls_20030307_284.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20030312_281.txt	Mar 19, 2014, 2:43 PM	4 KB	text
 acls_20030501_275.txt	Mar 19, 2014, 2:43 PM	15 KB	text
 acls_20030527_273.txt	Mar 19, 2014, 2:43 PM	5 KB	text
 acls_20030609_269.txt	Mar 19, 2014, 2:43 PM	5 KB	text
 acls_20030627_267.txt	Mar 19, 2014, 2:43 PM	2 KB	text
 acls_20030728_256.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20030820_252.txt	Mar 19, 2014, 2:43 PM	3 KB	text
 acls_20031007_247.txt	Mar 19, 2014, 2:43 PM	2 KB	text
 acls_20031027_246.txt	Mar 19, 2014, 2:43 PM	20 KB	text

OUTPUT: INDIVIDUAL FILES

Id	In_date	title
a	20070406	Agilent Technologies Signs Agreement to Acquire Stratagene Corp., a Developer of Life Science Research and Diagnostic Products
a	20090727	Agilent Technologies to Acquire Varian, Inc. for \$1.5 Billion; Transformational Transaction Establishes Agilent's Position as a Leading Provider of Analytical Instrumentation to the Applied and Life Sciences Markets
aa	20020318	Alcoa Agrees to Acquire Ivex Packaging Corporation; Move Broadens Alcoa's Position in the Food Service and Consumer Packaging Industry
aa	20000314	Alcoa to Acquire Cordant Technologies
aa	19990812	Alcoa Comments on Its Bid to Buy Reynolds Metals Company
aa	19990811	Alcoa Offers to Acquire Reynolds Metals Company in a Cash and Stock Transaction Valued at \$ 5.6 Billion
aa	19990811	Alcoa and Egyptalum Sign Memorandum of Understanding
aa	19990810	Alcoa and Kibar Holding Company Sign Letter of Intent
aa	19980306	Alcoa Division Announces Price Hike
abfs	19950710	ARKANSAS BEST CORPORATION AND WORLDWAY CORPORATION AGREE TO MERGER
abt	20031216	Abbott Announces Purchasing Agreement with Premier for MediSense Point-of-Care Products
abt	20031215	Abbott Laboratories to Acquire I-STAT Corp., a Leading Manufacturer of Point-of-Care Diagnostics; -Acquisition Strengthens Abbott's Position in the Rapidly Growing Point-of- Care Diagnostics Market with Proprietary Technology for Patient Testin
abt	20031215	Abbott Laboratories Names Executive to Lead Ross Products Division; -Gary L. Flynn to Retire After 32 Years of Service -
abt	20040113	Abbott Laboratories to Acquire TheraSense; -Acquisition Strengthens Abbott's Presence in the Large and Growing Blood Glucose Monitoring Market with Advanced Technology -
abt	20040112	Abbott Laboratories Submits Application to U.S. Department of Agriculture For Mad Cow Disease Test
abt	20040112	Abbott Laboratories Announces Agreement With Atria Genetics for HLA Tissue Typing Tests; -HLA Tests Enable Labs to Better Identify Appropriate Bone Marrow Donors -
abt	20061106	Abbott to Expand Presence in Lipid Management Market With Acquisition of Kos Pharmaceuticals; - Acquisition Strengthens Abbott's Late-Stage Pipeline -
abt	20091214	Abbott to Acquire STARLIMS Technologies Ltd., a Leader in Laboratory Information Management Systems; Enhances Abbott's Diagnostics Portfolio and Expertise in Information Management Systems; Expands STARLIMS' Presence Across Labora
abt	20090112	Abbott Announces Earnings Guidance for 2009; Expects to Deliver Another Year of Double-Digit EPS Growth
abt	20090112	Abbott Expands Its Growing Medical Device Business With Acquisition of Advanced Medical Optics (AMO); Adding an established global leader in ophthalmic care will provide long-term sustainable growth platform with more than \$1 billion in an
abt	20090112	Abbott Expands Its Growing Medical Device Business With Acquisition of Advanced Medical Optics (AMO); Adding an established global leader in ophthalmic care will provide long-term sustainable growth platform with more than \$1 billion in an
abt	19990709	Abbott Reports Increase in Sales, Earnings in Second Quarter, First Half
abt	19990708	Abbott Announces Executive Promotions
abt	19990708	Abbott Laboratories to Acquire Perclose
adbe	20050418	Adobe Systems Provides Q2 FY2005 Intra-Quarter Business Update; Company Expects Results Toward the High End of Previously Provided Financial Target Ranges
adbe	20050418	Adobe to Acquire Macromedia; Combined Company to Deliver Industry-Defining Technology Platform for Rich, Interactive Content
adbe	20050417	Adobe Licenses Rotoscoping Technology from Curious Software; Technology Will Enhance Digital Workflow for Adobe Video Collection
adbe	20050417	Adobe Enables High-Definition Workflow for Discovery HD Theater Network Promotions
adbe	20050417	Adobe Aligns with Industry Leaders to Deliver OpenHD Turnkey Desktop Solutions
adbe	19950622	Adobe Systems to Acquire Frame Technology; Electronic Publishing Leader to Add Technical Publishing Products and Strengthen Its Presence in the UNIX Market
adbe	19950622	Adobe Systems Reports Second Quarter 1995 Results; Net Income Increases 96 Percent
adp	20030106	ADP Signs Agreement To Acquire ProBusiness, Inc.; ADP and ProBusiness Produce A Strong Strategic Combination
adp	20030106	Automatic Data Processing to Acquire ProBusiness Services, Inc. for \$17.00 Per Share in Cash
adp	20030106	Automatic Data Processing to Acquire ProBusiness Services, Inc. for \$17.00 Per Share in Cash
ads	20050803	adidas-salomon to Combine with Reebok and Create EUR 9 Blllion Footprint in Global Athletic Footwear, Apparel and Hardware Markets
adsk	20080501	Autodesk Announces Intent to Acquire Moldflow, Leading Provider of Injection Molding Simulation Software; Deal Would Expand Autodesk Digital Prototyping in Plastic Parts Markets

OUTPUT: SPREADSHEET

Scratching the surface

- * If you can precisely describe how to do something, even if that explanation is complex, a computer can do the hard work.
- * Automate raw data gathering.
- * Use open source libraries for analysis.
- * Write your own algorithms to analyze content.

COMMENTS AND QUESTIONS