# Machine Learning: Example Remote Text Classification in Rapid-i

Dara (loss...) & Wayne (devzing)

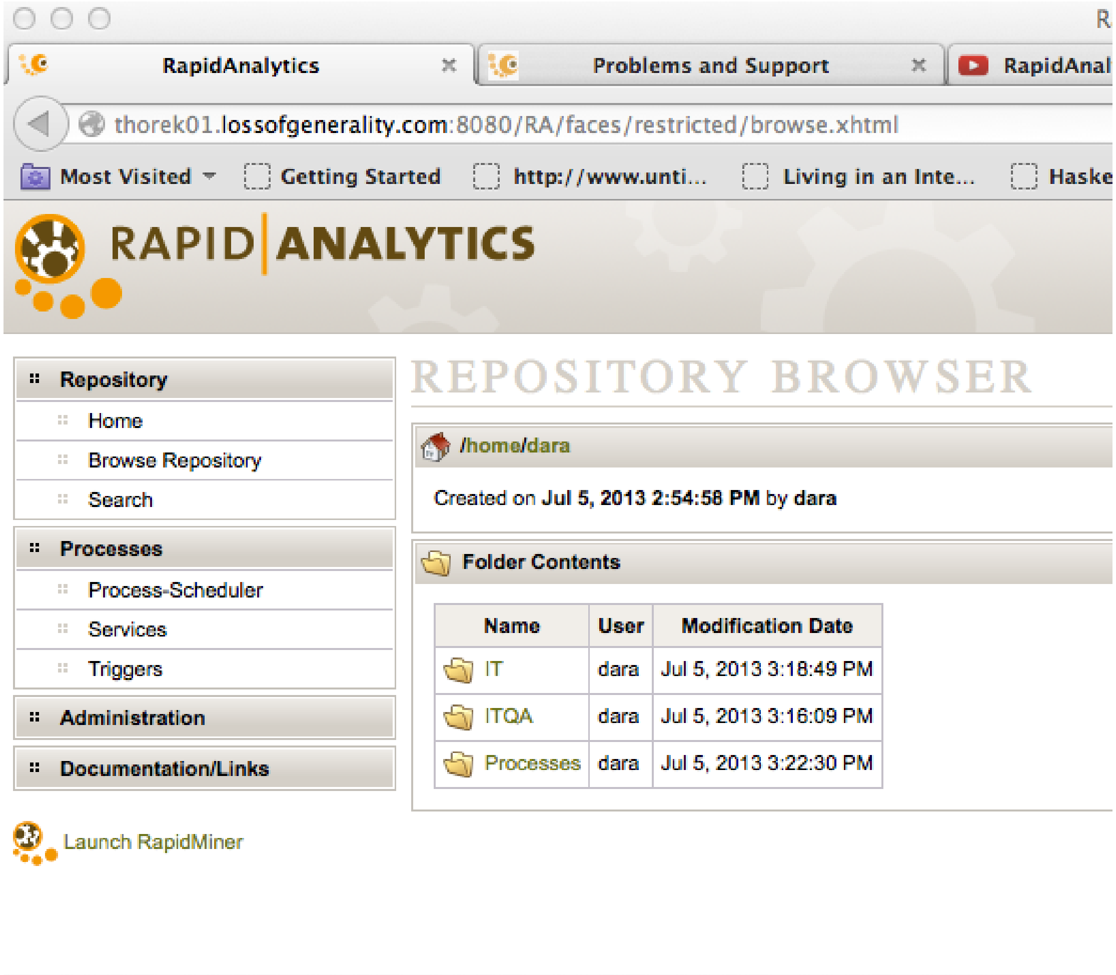
dara@lossofgenerality.com

## Setup

The architecture of the system is a mixed client-server:

1. Client: remote desktop running Rapid Miner, or local client on the remote server running Rapid Miner (RM)

2. The processes could all be executed on Rapid Analytics server (RA) in USA, or they could be executed on the desktop running RM but remote-calling the RA:



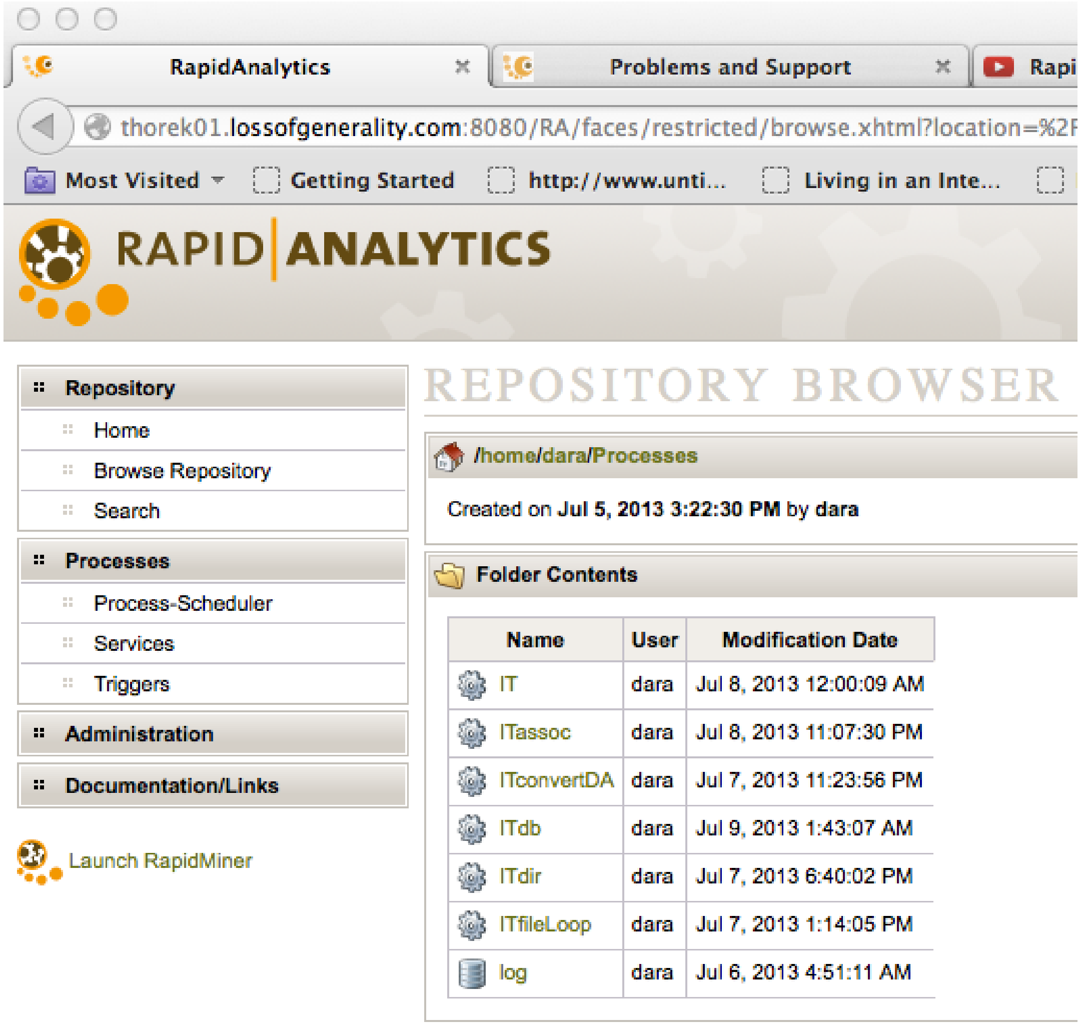
REPOSITORY BROWSER

/home/dara

Created on **Jul 5, 2013 2:54:58 PM** by **dara**

Folder Contents

| Name | User | Modification Date |
|---|---|---|
| IT | dara | Jul 5, 2013 3:18:49 PM |
| ITQA | dara | Jul 5, 2013 3:16:09 PM |
| Processes | dara | Jul 5, 2013 3:22:30 PM |



REPOSITORY BROWSER

/home/dara/Processes

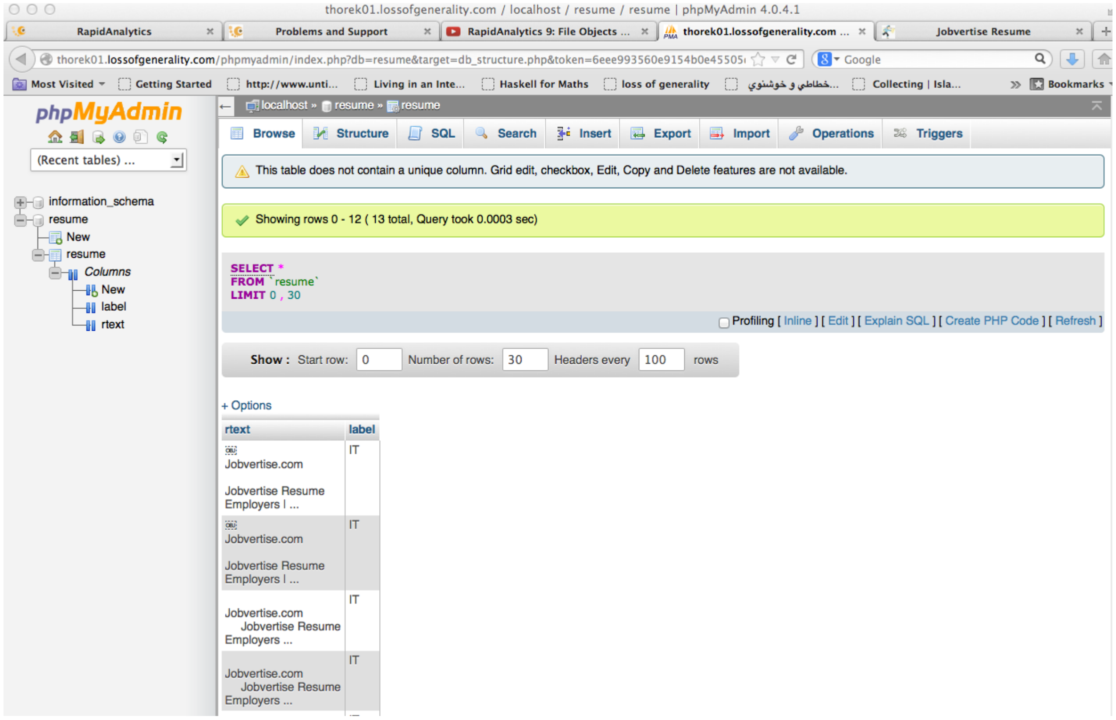
Created on **Jul 5, 2013 3:22:30 PM** by **dara**

**Folder Contents**

| Name | User | Modification Date |
|---|---|---|
| IT | dara | Jul 8, 2013 12:00:09 AM |
| ITassoc | dara | Jul 8, 2013 11:07:30 PM |
| ITconvertDA | dara | Jul 7, 2013 11:23:56 PM |
| ITdb | dara | Jul 9, 2013 1:43:07 AM |
| ITdir | dara | Jul 7, 2013 6:40:02 PM |
| ITfileLoop | dara | Jul 7, 2013 1:14:05 PM |
| log | dara | Jul 6, 2013 4:51:11 AM |

3. Files could be read from web URLs, local directories, server directories, or mySQL database on RA, or mix of all of that.

mySQL



thorek01.lossofgenerality.com/phpmyadmin/index.php?db=resume&target=db_structure.php&token=6eee993560e9154b0e45505

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

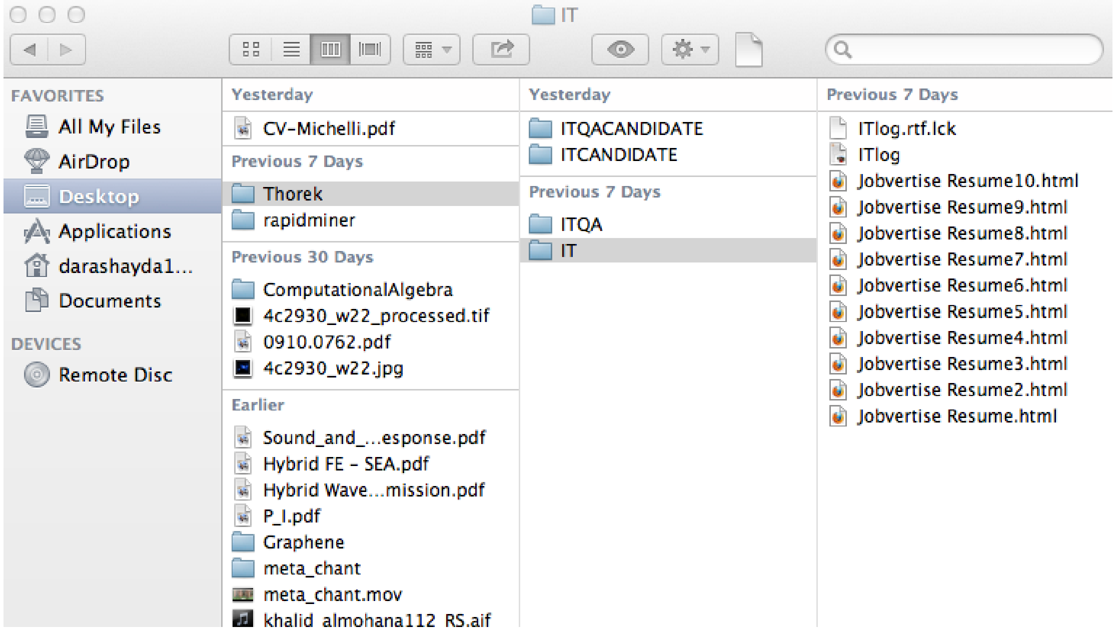
Showing rows 0 - 12 ( 13 total, Query took 0.0003 sec)

```
SELECT *
FROM `resume`
LIMIT 0 , 30
```

Show : Start row: 0 Number of rows: 30 Headers every 100 rows

| rtext | label |
|---|---|
| Jobvertise.com<br>Jobvertise Resume Employers I ... | IT |
| Jobvertise.com<br>Jobvertise Resume Employers I ... | IT |
| Jobvertise.com<br>Jobvertise Resume Employers ... | IT |
| Jobvertise.com<br>Jobvertise Resume Employers ... | IT |

Desktop directories:



Yesterday
CV-Michelli.pdf
Previous 7 Days
Thorek
rapidminer
Previous 30 Days
ComputationalAlgebra
4c2930_w22_processed.tif
0910.0762.pdf
4c2930_w22.jpg
Earlier
Sound_and_...esponse.pdf
Hybrid FE - SEA.pdf
Hybrid Wave...mission.pdf
P_I.pdf
Graphene
meta_chant
meta_chant.mov
khalid_almohana112_RS.aif

Yesterday
ITQACANDIDATE
ITCANDIDATE
Previous 7 Days
ITQA
IT

Previous 7 Days
ITlog.rtf.lck
ITlog
Jobvertise Resume10.html
Jobvertise Resume9.html
Jobvertise Resume8.html
Jobvertise Resume7.html
Jobvertise Resume6.html
Jobvertise Resume5.html
Jobvertise Resume4.html
Jobvertise Resume3.html
Jobvertise Resume2.html
Jobvertise Resume.html

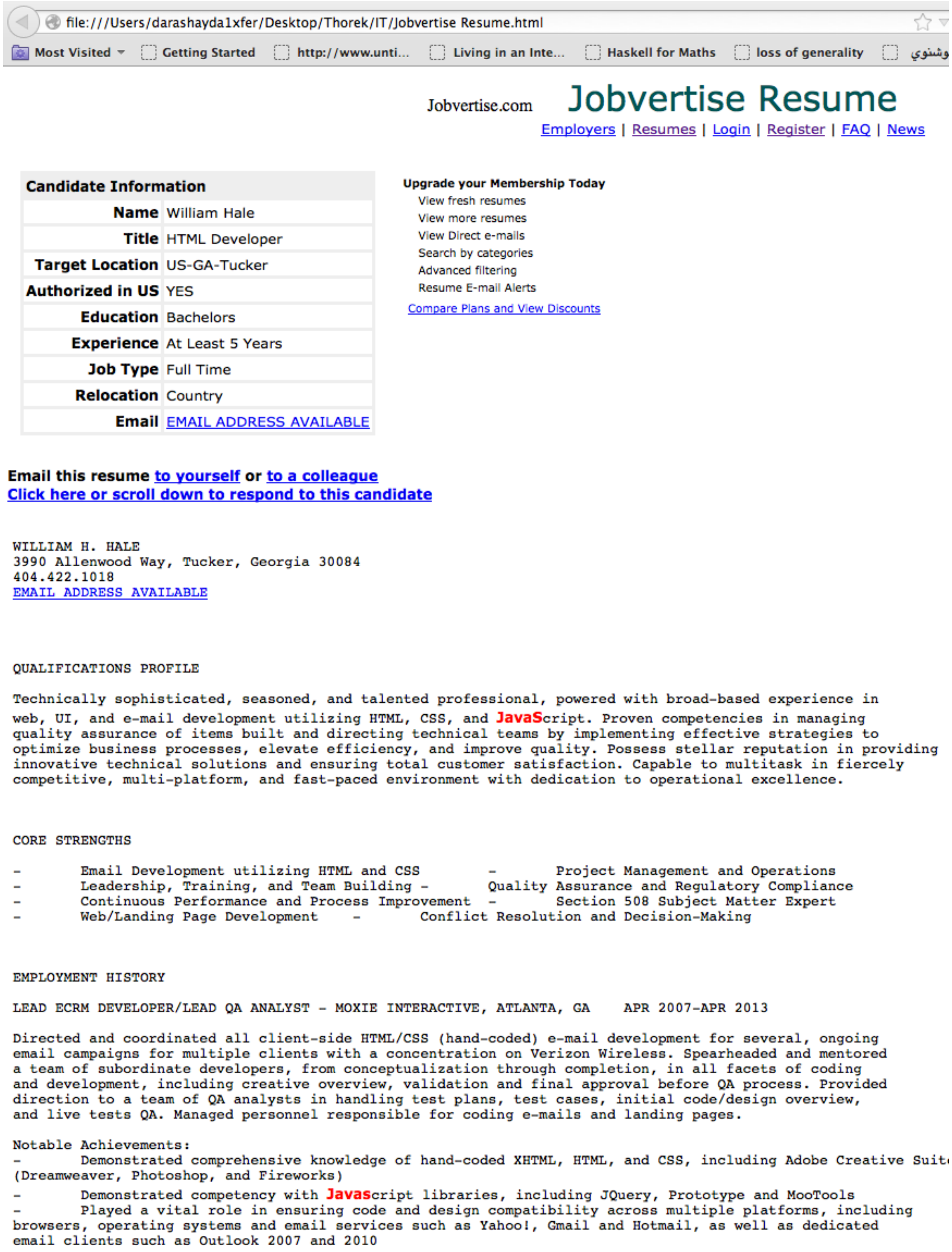
# Accent

Please read this document with a immix of suave Dutch Farsi accent.

# Example

Less than 20 resumes were randomly selected from a job site and stored in HTML format in a mySQL database on RA.

Sample IT resume, corpus for learning algorithm:



Jobvertise.com **Jobvertise Resume**

Employers | Resumes | Login | Register | FAQ | News

| **Candidate Information** | |
|---|---|
| **Name** | William Hale |
| **Title** | HTML Developer |
| **Target Location** | US-GA-Tucker |
| **Authorized in US** | YES |
| **Education** | Bachelors |
| **Experience** | At Least 5 Years |
| **Job Type** | Full Time |
| **Relocation** | Country |
| **Email** | EMAIL ADDRESS AVAILABLE |

**Upgrade your Membership Today**

View fresh resumes
View more resumes
View Direct e-mails
Search by categories
Advanced filtering
Resume E-mail Alerts

Compare Plans and View Discounts

**Email this resume to yourself or to a colleague**
**Click here or scroll down to respond to this candidate**

WILLIAM H. HALE
3990 Allenwood Way, Tucker, Georgia 30084
404.422.1018
EMAIL ADDRESS AVAILABLE

QUALIFICATIONS PROFILE

Technically sophisticated, seasoned, and talented professional, powered with broad-based experience in web, UI, and e-mail development utilizing HTML, CSS, and JavaScript. Proven competencies in managing quality assurance of items built and directing technical teams by implementing effective strategies to optimize business processes, elevate efficiency, and improve quality. Possess stellar reputation in providing innovative technical solutions and ensuring total customer satisfaction. Capable to multitask in fiercely competitive, multi-platform, and fast-paced environment with dedication to operational excellence.

CORE STRENGTHS

```
-       Email Development utilizing HTML and CSS        -       Project Management and Operations
-       Leadership, Training, and Team Building -       Quality Assurance and Regulatory Compliance
-       Continuous Performance and Process Improvement  -       Section 508 Subject Matter Expert
-       Web/Landing Page Development    -       Conflict Resolution and Decision-Making
```

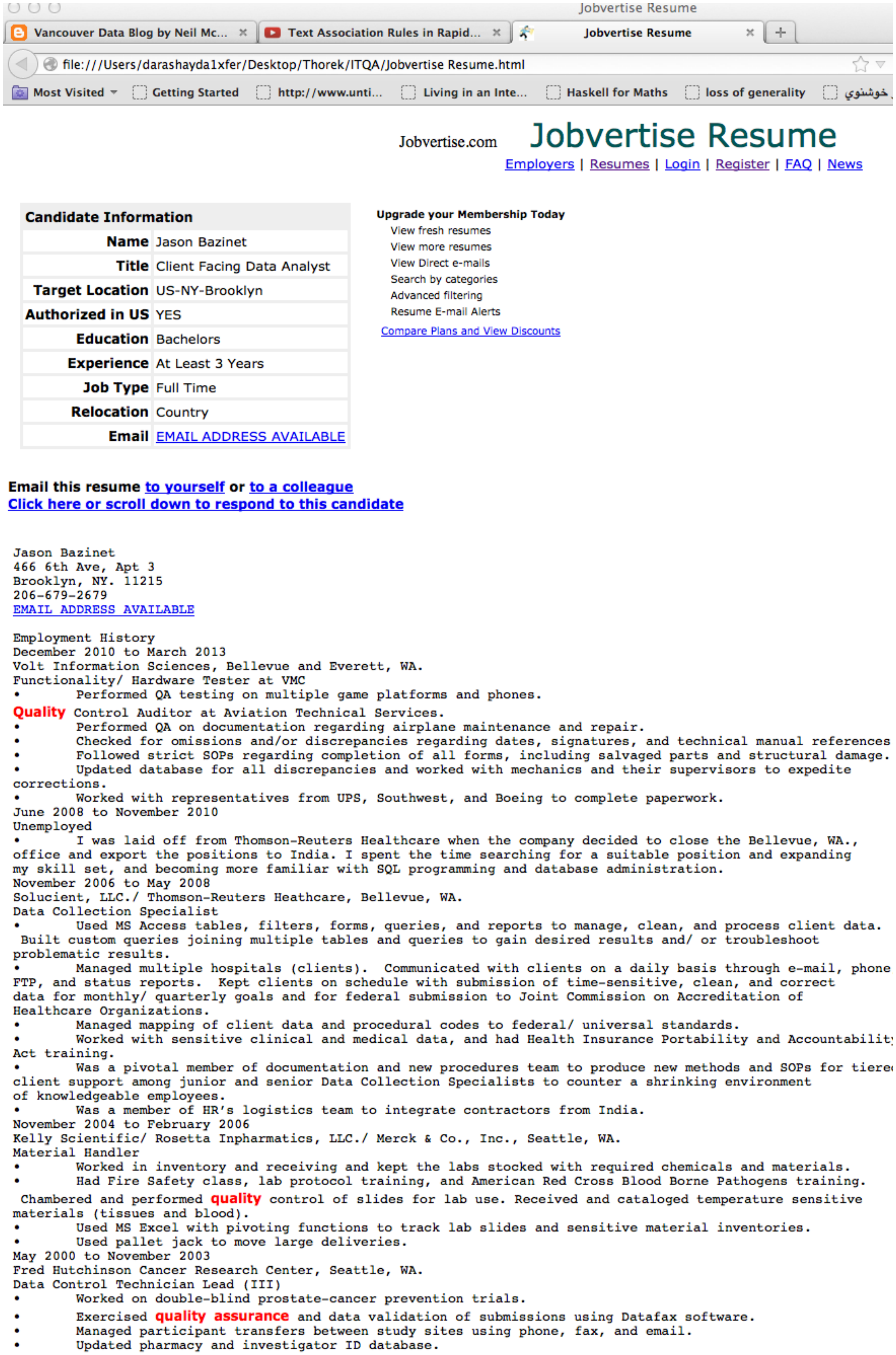
EMPLOYMENT HISTORY

LEAD ECRM DEVELOPER/LEAD QA ANALYST - MOXIE INTERACTIVE, ATLANTA, GA    APR 2007-APR 2013

Directed and coordinated all client-side HTML/CSS (hand-coded) e-mail development for several, ongoing email campaigns for multiple clients with a concentration on Verizon Wireless. Spearheaded and mentored a team of subordinate developers, from conceptualization through completion, in all facets of coding and development, including creative overview, validation and final approval before QA process. Provided direction to a team of QA analysts in handling test plans, test cases, initial code/design overview, and live tests QA. Managed personnel responsible for coding e-mails and landing pages.

Notable Achievements:
- Demonstrated comprehensive knowledge of hand-coded XHTML, HTML, and CSS, including Adobe Creative Suit (Dreamweaver, Photoshop, and Fireworks)
- Demonstrated competency with Javascript libraries, including JQuery, Prototype and MooTools
- Played a vital role in ensuring code and design compatibility across multiple platforms, including browsers, operating systems and email services such as Yahoo!, Gmail and Hotmail, as well as dedicated email clients such as Outlook 2007 and 2010

Sample IT related Quality Assurance :



Jobvertise.com **Jobvertise Resume**

Employers | Resumes | Login | Register | FAQ | News

| **Candidate Information** | |
|---|---|
| **Name** | Jason Bazinet |
| **Title** | Client Facing Data Analyst |
| **Target Location** | US-NY-Brooklyn |
| **Authorized in US** | YES |
| **Education** | Bachelors |
| **Experience** | At Least 3 Years |
| **Job Type** | Full Time |
| **Relocation** | Country |
| **Email** | EMAIL ADDRESS AVAILABLE |

**Upgrade your Membership Today**

View fresh resumes
View more resumes
View Direct e-mails
Search by categories
Advanced filtering
Resume E-mail Alerts

Compare Plans and View Discounts

**Email this resume to yourself or to a colleague**
**Click here or scroll down to respond to this candidate**

Jason Bazinet
466 6th Ave, Apt 3
Brooklyn, NY. 11215
206-679-2679
EMAIL ADDRESS AVAILABLE

Employment History
December 2010 to March 2013
Volt Information Sciences, Bellevue and Everett, WA.
Functionality/ Hardware Tester at VMC
• Performed QA testing on multiple game platforms and phones.
**Quality** Control Auditor at Aviation Technical Services.
• Performed QA on documentation regarding airplane maintenance and repair.
• Checked for omissions and/or discrepancies regarding dates, signatures, and technical manual references
• Followed strict SOPs regarding completion of all forms, including salvaged parts and structural damage.
• Updated database for all discrepancies and worked with mechanics and their supervisors to expedite corrections.
• Worked with representatives from UPS, Southwest, and Boeing to complete paperwork.
June 2008 to November 2010
Unemployed
• I was laid off from Thomson-Reuters Healthcare when the company decided to close the Bellevue, WA., office and export the positions to India. I spent the time searching for a suitable position and expanding my skill set, and becoming more familiar with SQL programming and database administration.
November 2006 to May 2008
Solucient, LLC./ Thomson-Reuters Heathcare, Bellevue, WA.
Data Collection Specialist
• Used MS Access tables, filters, forms, queries, and reports to manage, clean, and process client data. Built custom queries joining multiple tables and queries to gain desired results and/ or troubleshoot problematic results.
• Managed multiple hospitals (clients). Communicated with clients on a daily basis through e-mail, phone FTP, and status reports. Kept clients on schedule with submission of time-sensitive, clean, and correct data for monthly/ quarterly goals and for federal submission to Joint Commission on Accreditation of Healthcare Organizations.
• Managed mapping of client data and procedural codes to federal/ universal standards.
• Worked with sensitive clinical and medical data, and had Health Insurance Portability and Accountabilit Act training.
• Was a pivotal member of documentation and new procedures team to produce new methods and SOPs for tiere client support among junior and senior Data Collection Specialists to counter a shrinking environment of knowledgeable employees.
• Was a member of HR's logistics team to integrate contractors from India.
November 2004 to February 2006
Kelly Scientific/ Rosetta Inpharmatics, LLC./ Merck & Co., Inc., Seattle, WA.
Material Handler
• Worked in inventory and receiving and kept the labs stocked with required chemicals and materials.
• Had Fire Safety class, lab protocol training, and American Red Cross Blood Borne Pathogens training. Chambered and performed **quality** control of slides for lab use. Received and cataloged temperature sensitive materials (tissues and blood).
• Used MS Excel with pivoting functions to track lab slides and sensitive material inventories.
• Used pallet jack to move large deliveries.
May 2000 to November 2003
Fred Hutchinson Cancer Research Center, Seattle, WA.
Data Control Technician Lead (III)
• Worked on double-blind prostate-cancer prevention trials.
• Exercised **quality assurance** and data validation of submissions using Datafax software.
• Managed participant transfers between study sites using phone, fax, and email.
• Updated pharmacy and investigator ID database.

They were unevenly divided into IT development professional and IT Quality Assurance.

Then a Naive Bayes statistical classification algorithm ran over these resumes, to machine-learn how to classify a future resume with similar wordings.

Certain keywords from the resumes were selected for the learn algorithm:

management, design, work, assurance

Finally another set of new resumes selected without any classification and applied to the model built from the above using the same keywords.

## Processes

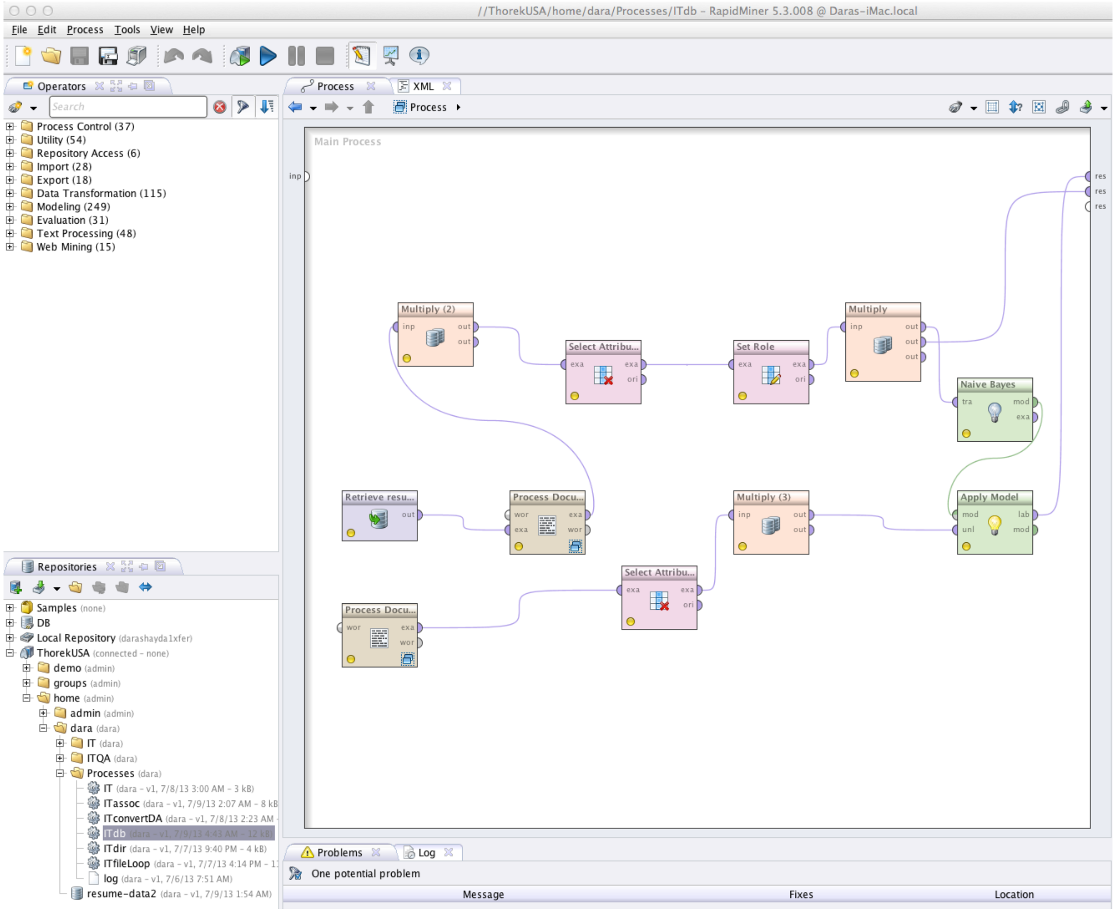
Following is the screen view of the Rapid-Miner running on the remote desktop:

1. Lower Left two icons RETRIEVE resu... and PROCSS DOCU..., are correspondingly for reading the database on the RA server in USA called resme-data2 (see the left most lowest part of the image), and PROCESS DOCU... reading the local directories in the remote desktop in Toronto.

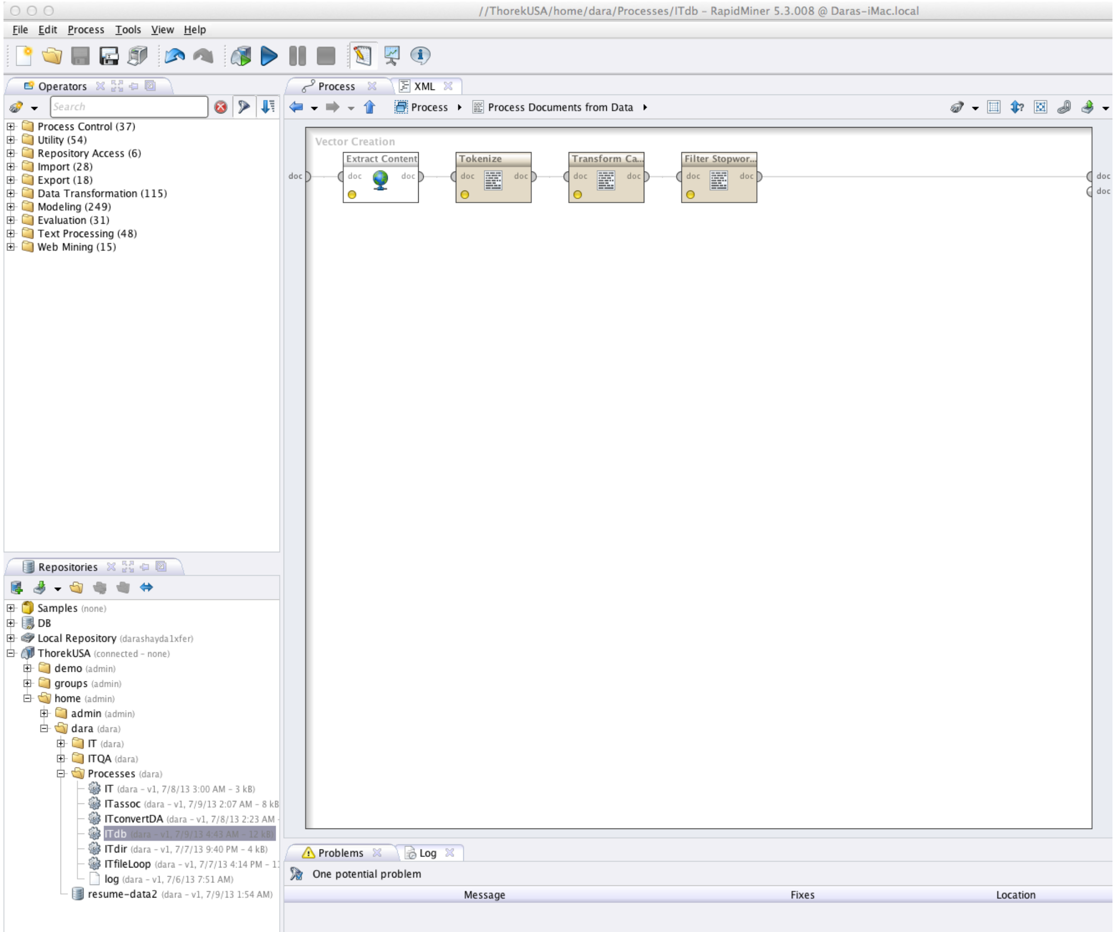
2. ITdb is the name of the process that does all the work, partially running on the desktop (RM) and for the most part running the RA in USA.

3. Box icons are pieces of code running small portions of the entire system. For example SELECT ATTRIBUTE selects the keywords “management, design, work, assurance “ and their stats.

4. Right-most box icon is Naive Bayes and its ouput is fed to the Apply Model which is where the machine-learnt classification is applied to the new resumes coming from PROCESS DOCU... lower-most left side of the screen.

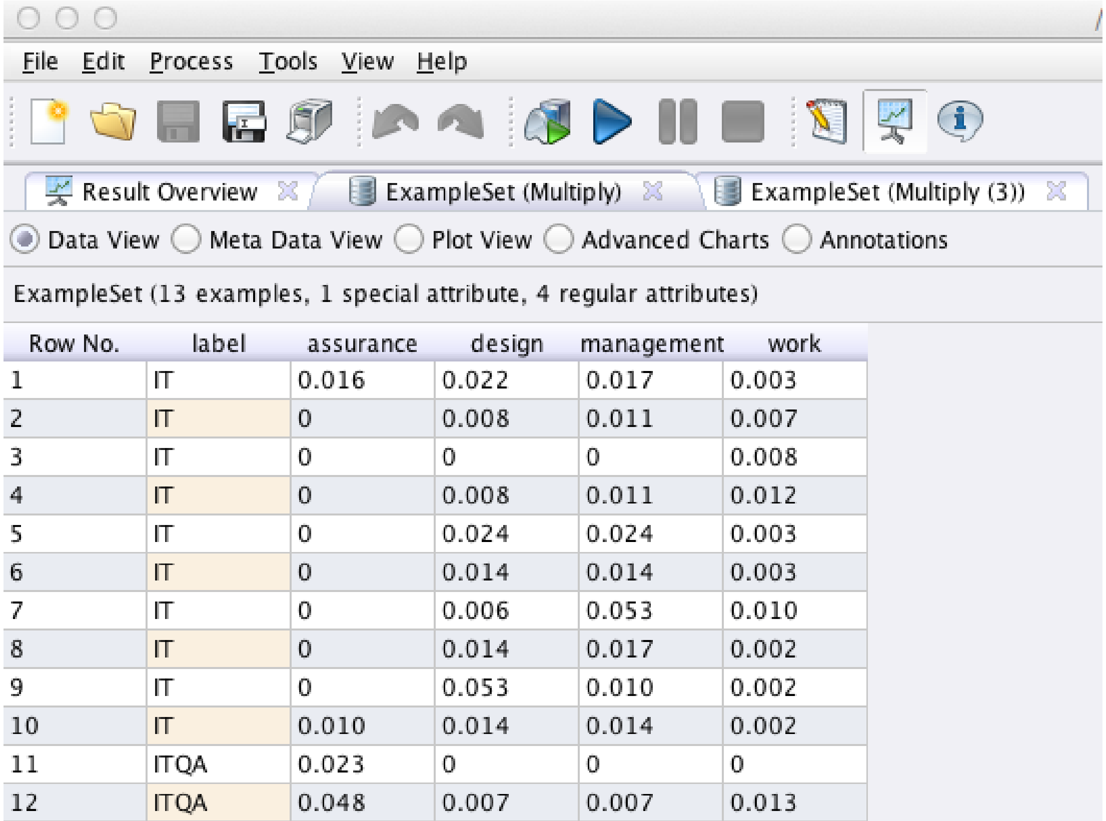


5. DOCU PROC... is where the HTML files i.e. the resumes are read and processed i.e. the words inside the resume are scanned and transformed to lower case and all the useless words like ‘is’ or ‘the’ are removed from the original resume to simplify.


//ThorekUSA/home/dara/Processes/ITdb – RapidMiner 5.3.008 @ Daras-iMac.local
File Edit Process Tools View Help
Operators
Search
Process Control (37)
Utility (54)
Repository Access (6)
Import (28)
Export (18)
Data Transformation (115)
Modeling (249)
Evaluation (31)
Text Processing (48)
Web Mining (15)
Process
XML
Process ▸ Process Documents from Data ▸
Vector Creation
doc
Extract Content
Tokenize
Transform Ca...
Filter Stopwor...
doc
doc
Repositories
Samples (none)
DB
Local Repository (darashayda1xfer)
ThorekUSA (connected – none)
demo (admin)
groups (admin)
home (admin)
admin (admin)
dara (dara)
IT (dara)
ITQA (dara)
Processes (dara)
IT (dara – v1, 7/8/13 3:00 AM – 3 kB)
ITassoc (dara – v1, 7/9/13 2:07 AM – 8 kB
ITconvertDA (dara – v1, 7/8/13 2:23 AM
ITdb (dara – v1, 7/9/13 4:43 AM – 12 kB)
ITdir (dara – v1, 7/7/13 9:40 PM – 4 kB)
ITfileLoop (dara – v1, 7/7/13 4:14 PM – 1
log (dara – v1, 7/6/13 7:51 AM)
resume-data2 (dara – v1, 7/9/13 1:54 AM)
Problems
Log
Message
Fixes
Location

One potential problem

6. This is the output of #5 processing the resumes and their keywords statistics which are then fed to the Naive Bayes machine learning classification algorithm:


File Edit Process Tools View Help
Result Overview
ExampleSet (Multiply)
ExampleSet (Multiply (3))
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

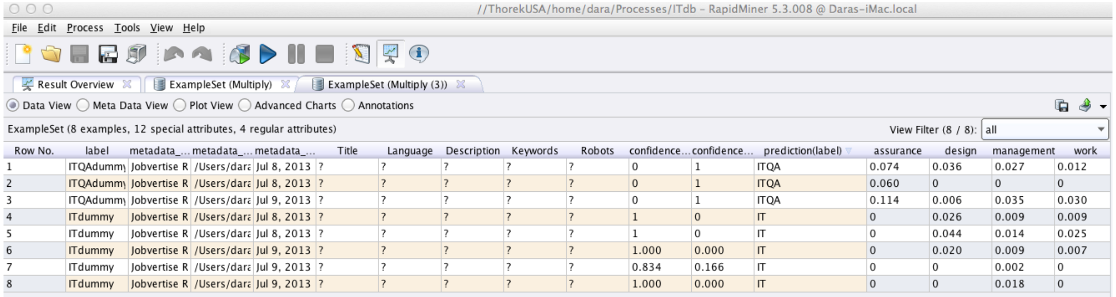
ExampleSet (13 examples, 1 special attribute, 4 regular attributes)

| Row No. | label | assurance | design | management | work |
|---|---|---|---|---|---|
| 1 | IT | 0.016 | 0.022 | 0.017 | 0.003 |
| 2 | IT | 0 | 0.008 | 0.011 | 0.007 |
| 3 | IT | 0 | 0 | 0 | 0.008 |
| 4 | IT | 0 | 0.008 | 0.011 | 0.012 |
| 5 | IT | 0 | 0.024 | 0.024 | 0.003 |
| 6 | IT | 0 | 0.014 | 0.014 | 0.003 |
| 7 | IT | 0 | 0.006 | 0.053 | 0.010 |
| 8 | IT | 0 | 0.014 | 0.017 | 0.002 |
| 9 | IT | 0 | 0.053 | 0.010 | 0.002 |
| 10 | IT | 0.010 | 0.014 | 0.014 | 0.002 |
| 11 | ITQA | 0.023 | 0 | 0 | 0 |
| 12 | ITQA | 0.048 | 0.007 | 0.007 | 0.013 |

7. This is the output of the Naive Bayes model applied to the new set of resumes. The predictions are labeled accordingly with some of their confidence factors.

Although the predictions are correct the algorithm is not confident on some of the choices.

Right-most side columns are the stats from the new resumes.


//ThorekUSA/home/dara/Processes/ITdb – RapidMiner 5.3.008 @ Daras-iMac.local
File Edit Process Tools View Help
Result Overview
ExampleSet (Multiply)
ExampleSet (Multiply (3))
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (8 examples, 12 special attributes, 4 regular attributes)

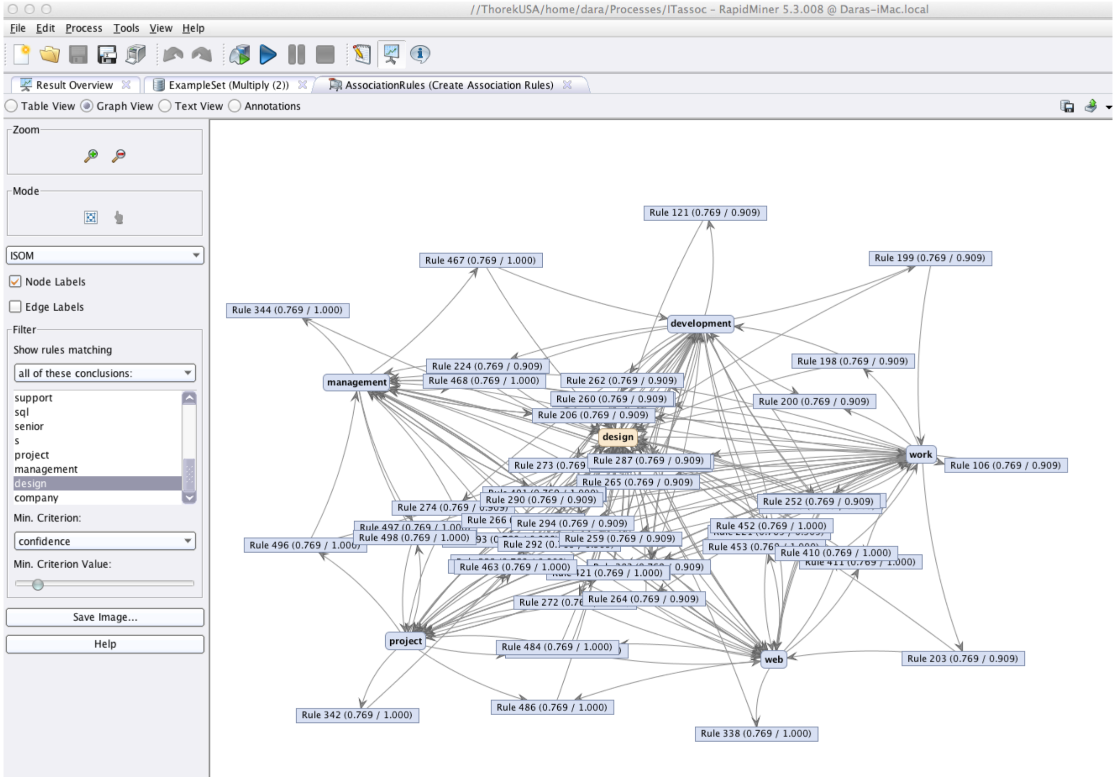
View Filter (8 / 8): all

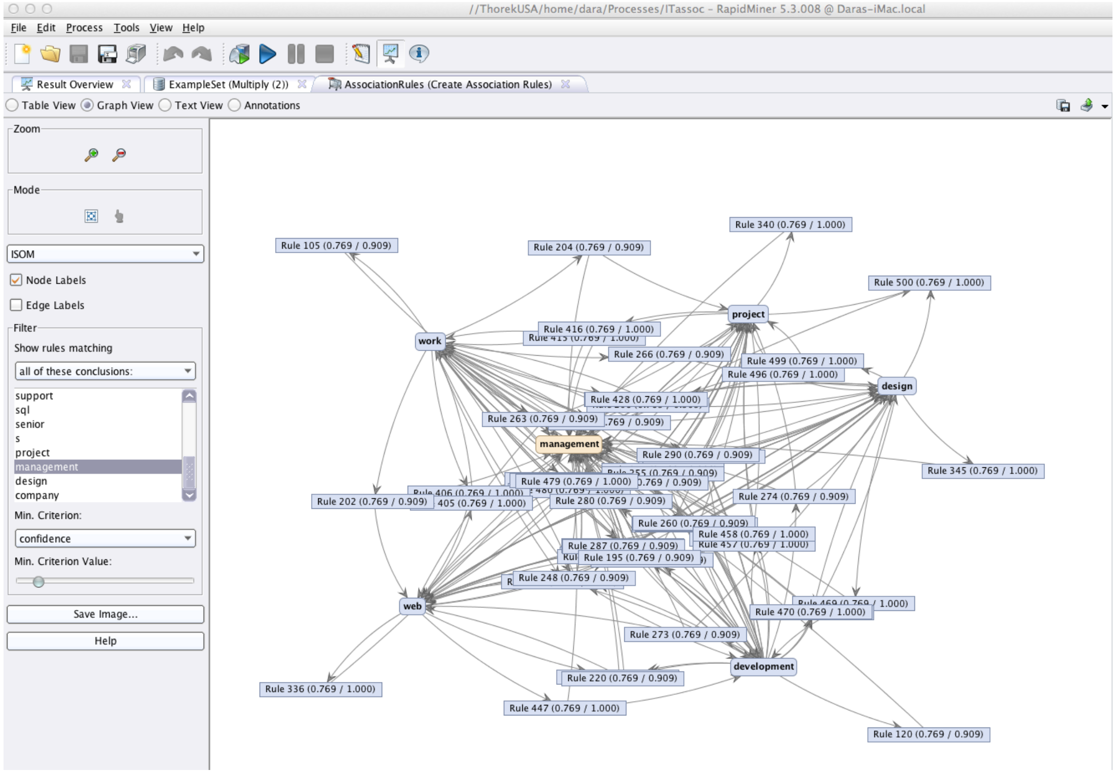
| Row No. | label | metadata_... | metadata_... | metadata_... | Title | Language | Description | Keywords | Robots | confidence... | confidence... | prediction(label) | assurance | design | management | work |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ITQAdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 0 | 1 | ITQA | 0.074 | 0.036 | 0.027 | 0.012 |
| 2 | ITQAdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 0 | 1 | ITQA | 0.060 | 0 | 0 | 0 |
| 3 | ITQAdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 0 | 1 | ITQA | 0.114 | 0.006 | 0.035 | 0.030 |
| 4 | ITdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 1 | 0 | IT | 0 | 0.026 | 0.009 | 0.009 |
| 5 | ITdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 1 | 0 | IT | 0 | 0.044 | 0.014 | 0.025 |
| 6 | ITdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 1.000 | 0.000 | IT | 0 | 0.020 | 0.009 | 0.007 |
| 7 | ITdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 0.834 | 0.166 | IT | 0 | 0 | 0.002 | 0 |
| 8 | ITdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 1.000 | 0.000 | IT | 0 | 0 | 0.018 | 0 |

**Note**: *The new resume’s correct classification not included in the system is tagged by suffix dummy i.e. ITQAdummy or ITdummy so we can check to see if the classification is correct.*

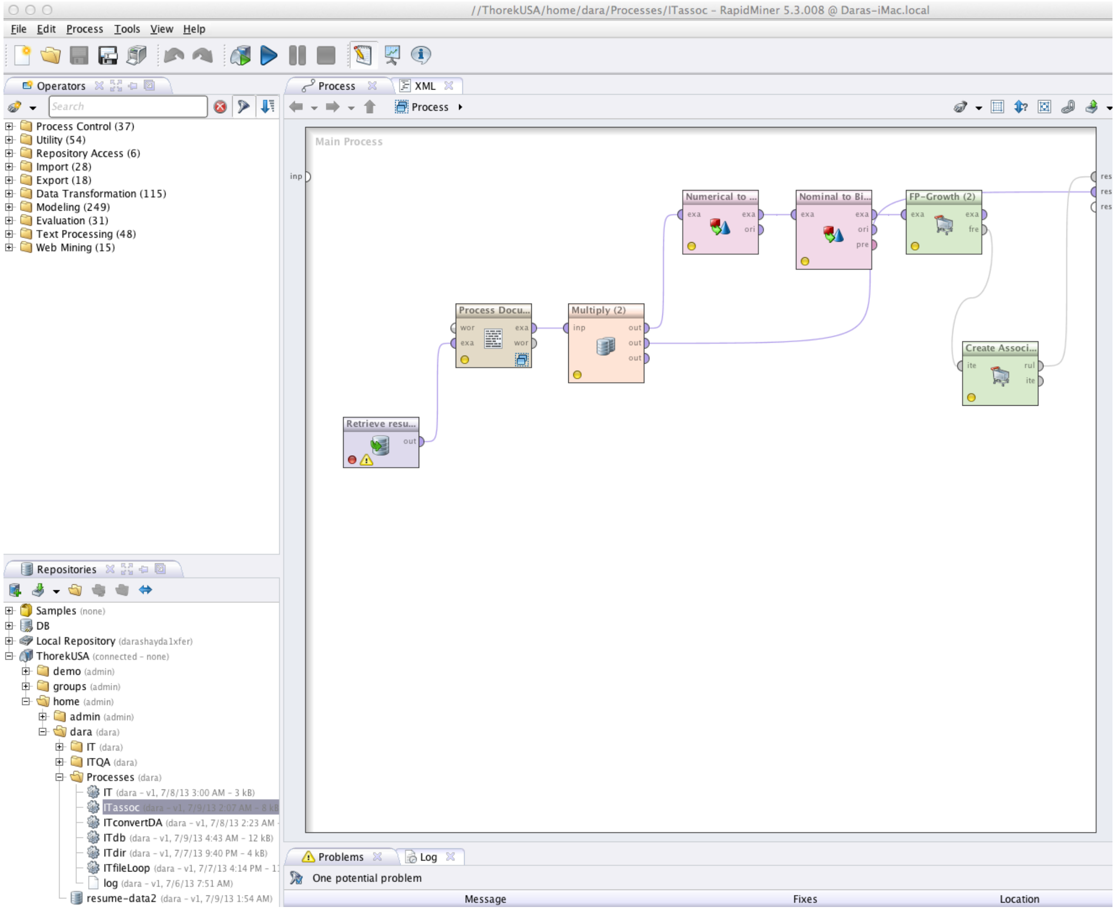
8. The textual connectivity and relationship of the words could be computed and graphed in the following sophisticated manner for the OPERATOR to decide which keywords are suitable for the machine learning application and possibly which type of algorithm:

Clearly you can see the design and management are linked in complex fashion in people’s resumes.





The above graphs were again programmed the same way in Rapid-i using the graphical programmer:

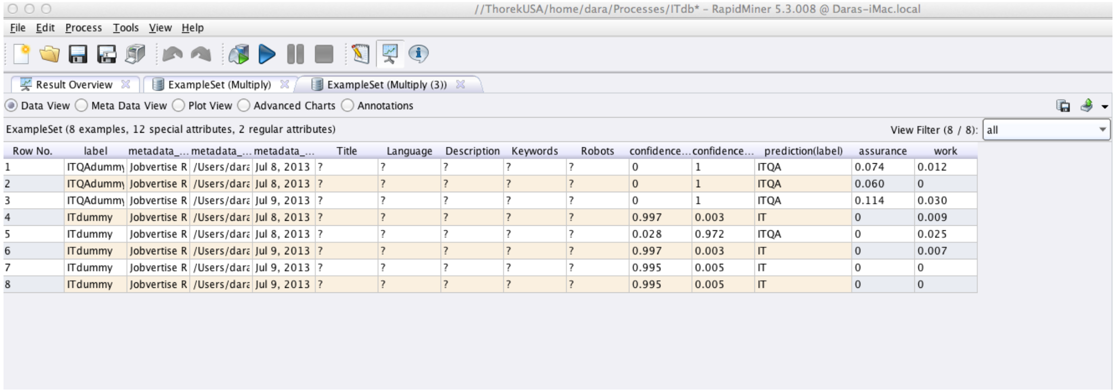


//ThorekUSA/home/dara/Processes/ITassoc – RapidMiner 5.3.008 @ Daras-iMac.local

One potential problem

9. Now by altering the keywords we see some down turn in the classifications and these are not subtle nor easily understood by human cognition, hence the machine-learning:

Keywords are only work and assurance. As you can see the classification predicts another ITQA which should have been
IT!


File Edit Process Tools View Help
Result Overview
ExampleSet (Multiply)
ExampleSet (Multiply (3))
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

//ThorekUSA/home/dara/Processes/ITdb* – RapidMiner 5.3.008 @ Daras-iMac.local

ExampleSet (8 examples, 12 special attributes, 2 regular attributes)

View Filter (8 / 8): all

| Row No. | label | metadata_... | metadata_... | metadata_... | Title | Language | Description | Keywords | Robots | confidence... | confidence... | prediction(label) | assurance | work |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ITQAdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 0 | 1 | ITQA | 0.074 | 0.012 |
| 2 | ITQAdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 0 | 1 | ITQA | 0.060 | 0 |
| 3 | ITQAdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 0 | 1 | ITQA | 0.114 | 0.030 |
| 4 | ITdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 0.997 | 0.003 | IT | 0 | 0.009 |
| 5 | ITdummy | Jobvertise R | /Users/dara | Jul 8, 2013 | ? | ? | ? | ? | ? | 0.028 | 0.972 | ITQA | 0 | 0.025 |
| 6 | ITdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 0.997 | 0.003 | IT | 0 | 0.007 |
| 7 | ITdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 0.995 | 0.005 | IT | 0 | 0 |
| 8 | ITdummy | Jobvertise R | /Users/dara | Jul 9, 2013 | ? | ? | ? | ? | ? | 0.995 | 0.005 | IT | 0 | 0 |