SAND2011-0376P

Multilingual Text Analysis of Large Data

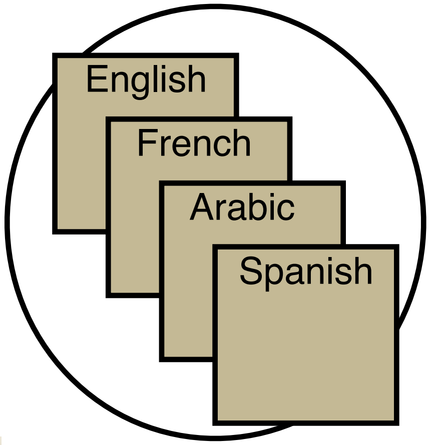
Brett Bader *and Ron Oldfield*
Sandia National Laboratories

January 20, 2011

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

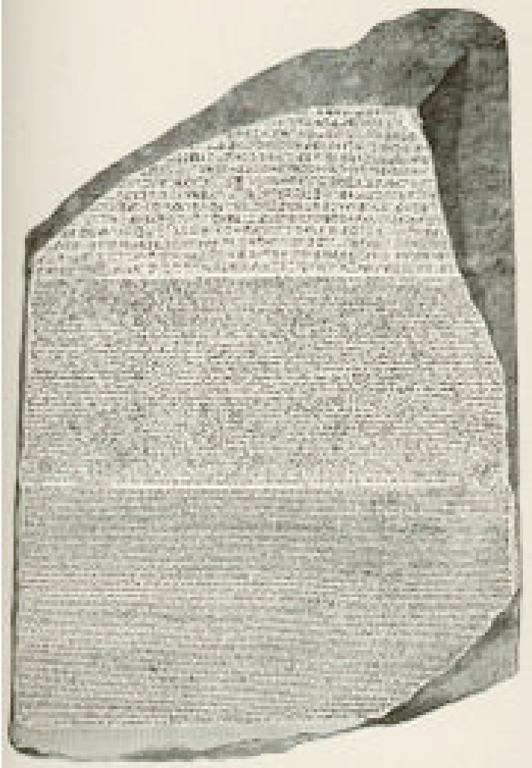
SNL has developed multilingual text analysis to link threats across multiple languages

- “Translate” new documents into a language-independent concept space, which is useful for:
  - Translation triage (i.e., translate documents in clusters of interest)
  - Ideological classification (e.g., hostile to democracy)
  - Multilingual sentiment analysis



Sandia’s database: 54 languages: >99% coverage of web

Afrikaans	Estonian	Norwegian
Albanian	Finnish	Persian (Farsi)
Amharic	French	Polish
Arabic	German	Portuguese
Aramaic	Greek (New Testament)	Romani
Armenian Eastern	Greek (Modern)	Romanian
Armenian Western	Hebrew (Old Testament)	Russian
Basque	Hebrew (Modern)	Scots Gaelic
Breton	Hungarian	Spanish
Chamorro	Indonesian	Swahili
Chinese (Simplified)	Italian	Swedish
Chinese (Traditional)	Japanese	Tagalog
Croatian	Korean	Thai
Czech	Latin	Turkish
Danish	Latvian	Ukrainian
Dutch	Lithuanian	Vietnamese
English	Manx Gaelic	Wolof
Esperanto	Maori	Xhosa



Bag of Words/Vector Space Model

example from (Berry, Drmac, Jessup, 1999)

Documents

D1: How to Bake Bread Without Recipes
D2: The Classic Art of Viennese Pastry
D3: Numerical Recipes: The Art of Scientific Computing
D4: Breads, Pastries, Pies and Cakes: Quantity Baking Recipes
D5: Pastry: A Book of Best French Recipes

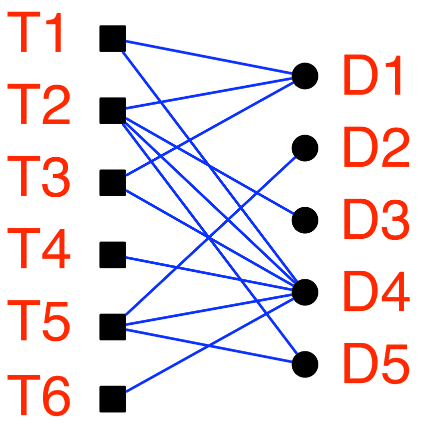
Terms

T1: bak(e,ing)
T2: recipes
T3: bread
T4: cake
T5: pastr(y,ies)
T6: pie

Key concepts

- Bag of words
- Stemming
- Vector space model
- Scaling for information content

Bipartite graph



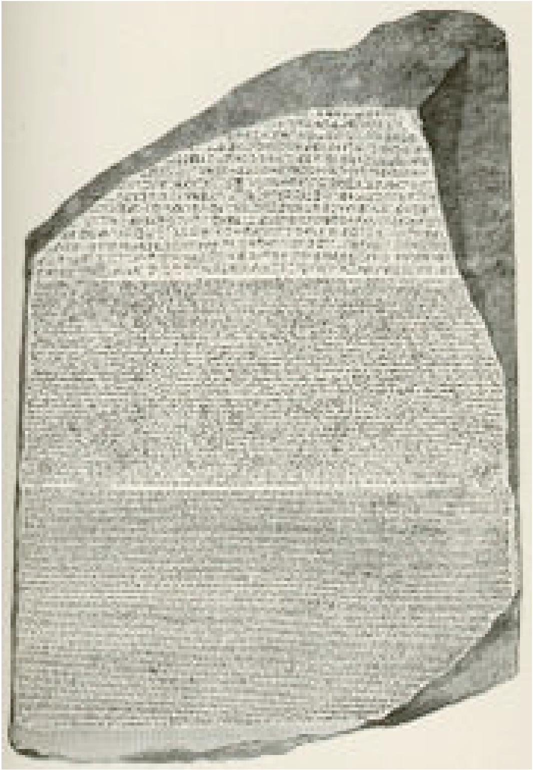
Term-by-doc (adjacency) matrix

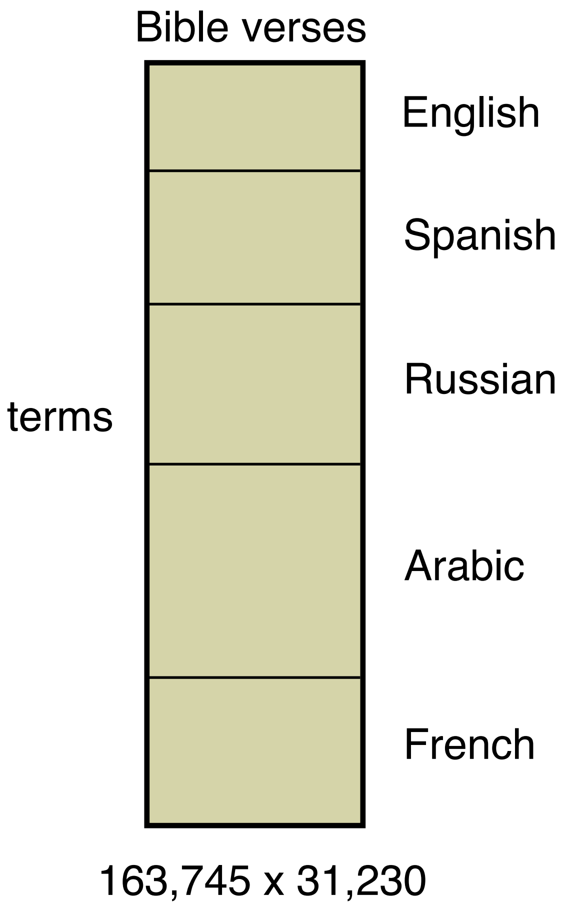
$$\hat{A} = \begin{array}{cccccc} D1 & D2 & D3 & D4 & D5 & \\ 1 & 0 & 0 & 1 & 0 & T1 \\ 1 & 0 & 1 & 1 & 1 & T2 \\ 1 & 0 & 0 & 1 & 0 & T3 \\ 0 & 0 & 0 & 1 & 0 & T4 \\ 0 & 1 & 0 & 1 & 1 & T5 \\ 0 & 0 & 0 & 1 & 0 & T6 \end{array}$$

Term-Document Matrix

Term-by-verse matrix
for all languages

Rosetta Stone





Look for co-occurrence of terms in the same verses and across languages to capture latent concepts

- *Multi-parallel* corpora
  - Bible
  - Quran
  - European Parliament proceedings (Europarl)*

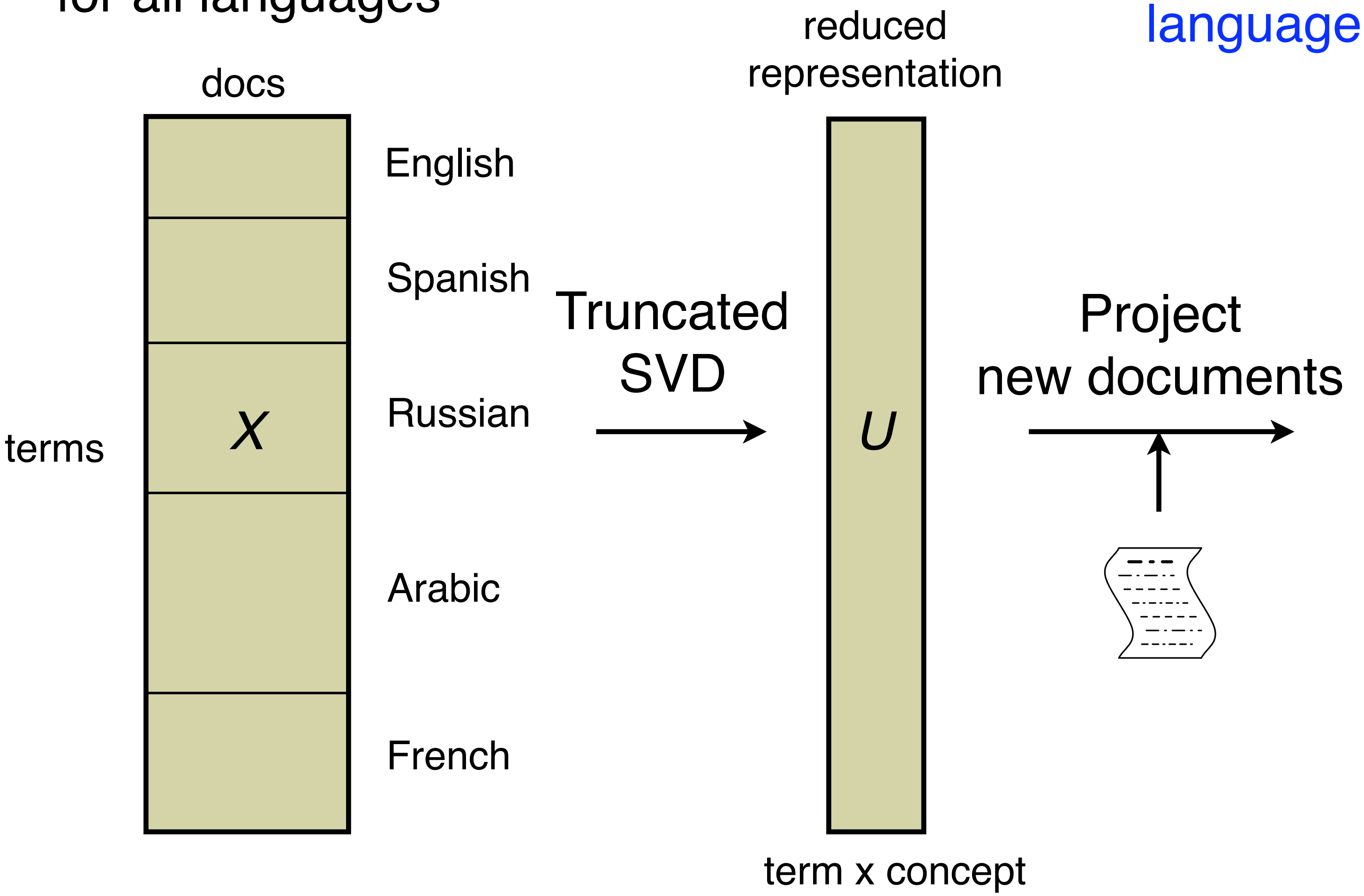
Europarl Corpus

- Extracted from the proceedings of the European Parliament
- Translations in 11 languages
  - French, Italian, Spanish, Portuguese (Romantic)
  - English, Dutch, German, Danish, Swedish (Germanic)
  - Greek
  - Finnish
- Sentence aligned text (16 M sentences across 11 languages)
- 1,247,832 speeches (including translations)
- 1,249,253 terms (from all 11 languages)
  - English terms: 46,074

Multilingual Latent Semantic Analysis

Term-by-doc matrix
for all languages

“Translate” new documents
into a small number of
language-independent features



reduced
representation

docs

terms

X

English

Spanish

Russian

Arabic

French

Truncated SVD

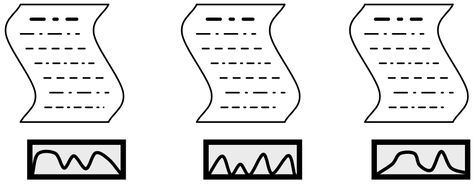
U

term x concept

Project
new documents

dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

Document feature
vector



Applications

- cross-language retrieval
- pairwise similarities for clustering
- machine learning applications

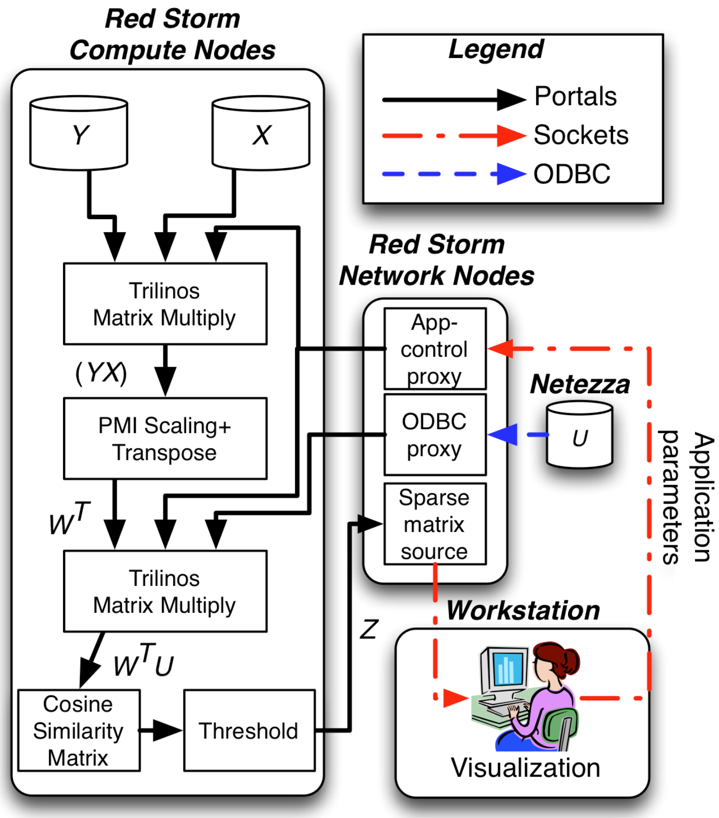
Multilingual Clustering is a Great Candidate for HPC

- Scale of Data
  - Millions of elements (Bible, Quran, Wikipedia, Europarl)
  - Computationally expensive (matrix multiplies for large matrices)
- Time to Solution
  - Interactive control/vis is a motivating factor
  - Focus on “strong scaling” capabilities of HPC platform
- Leveraging Existing Sandia Libraries
  - LMSA for dataset generation
  - Trilinos for computation
  - Titan for visualization
  - Nessie for data services (provides “glue” to integrate systems)

Architectural Challenges

Exploiting specialized architectures

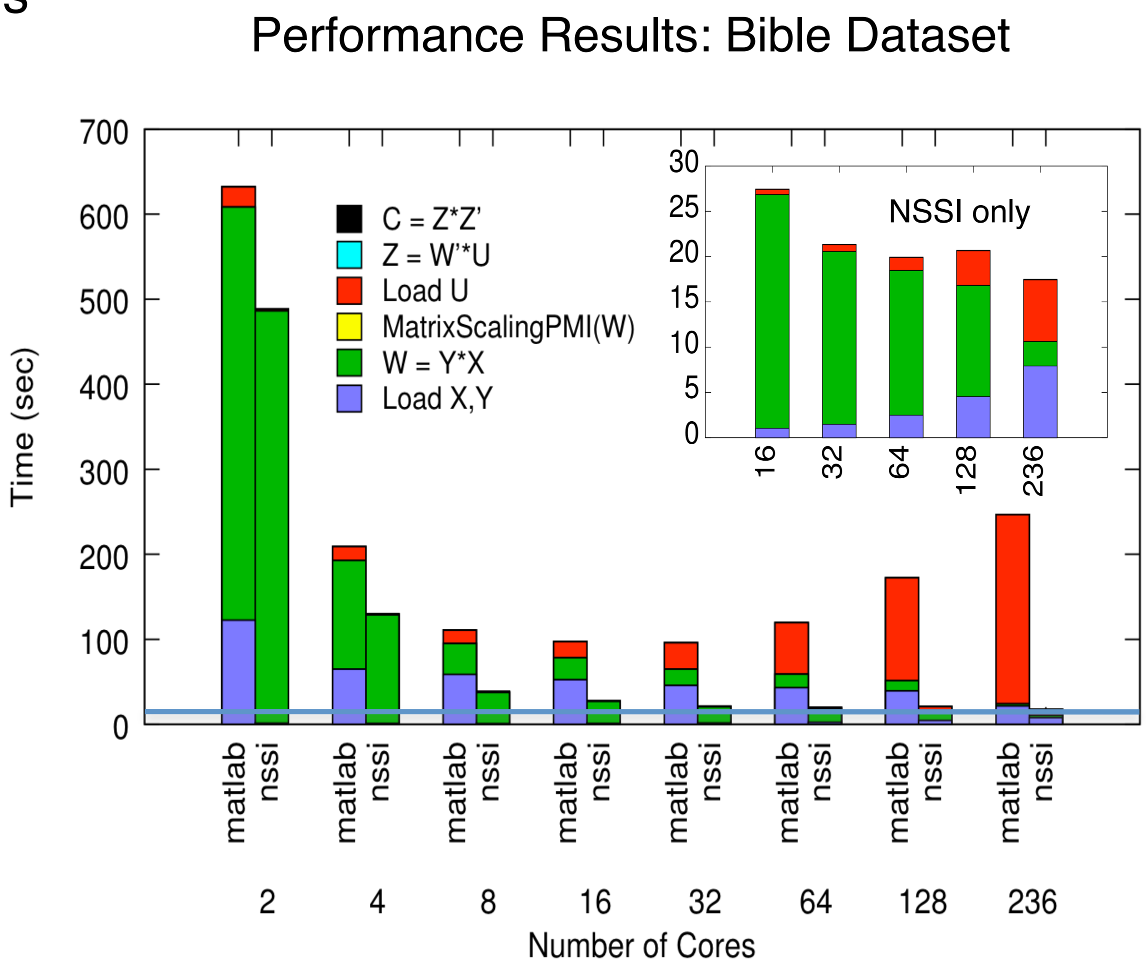
- Red Storm for numerics
- Clusters/Workstations for vis and interactive control
- Data Warehouse Appliances for database functionality



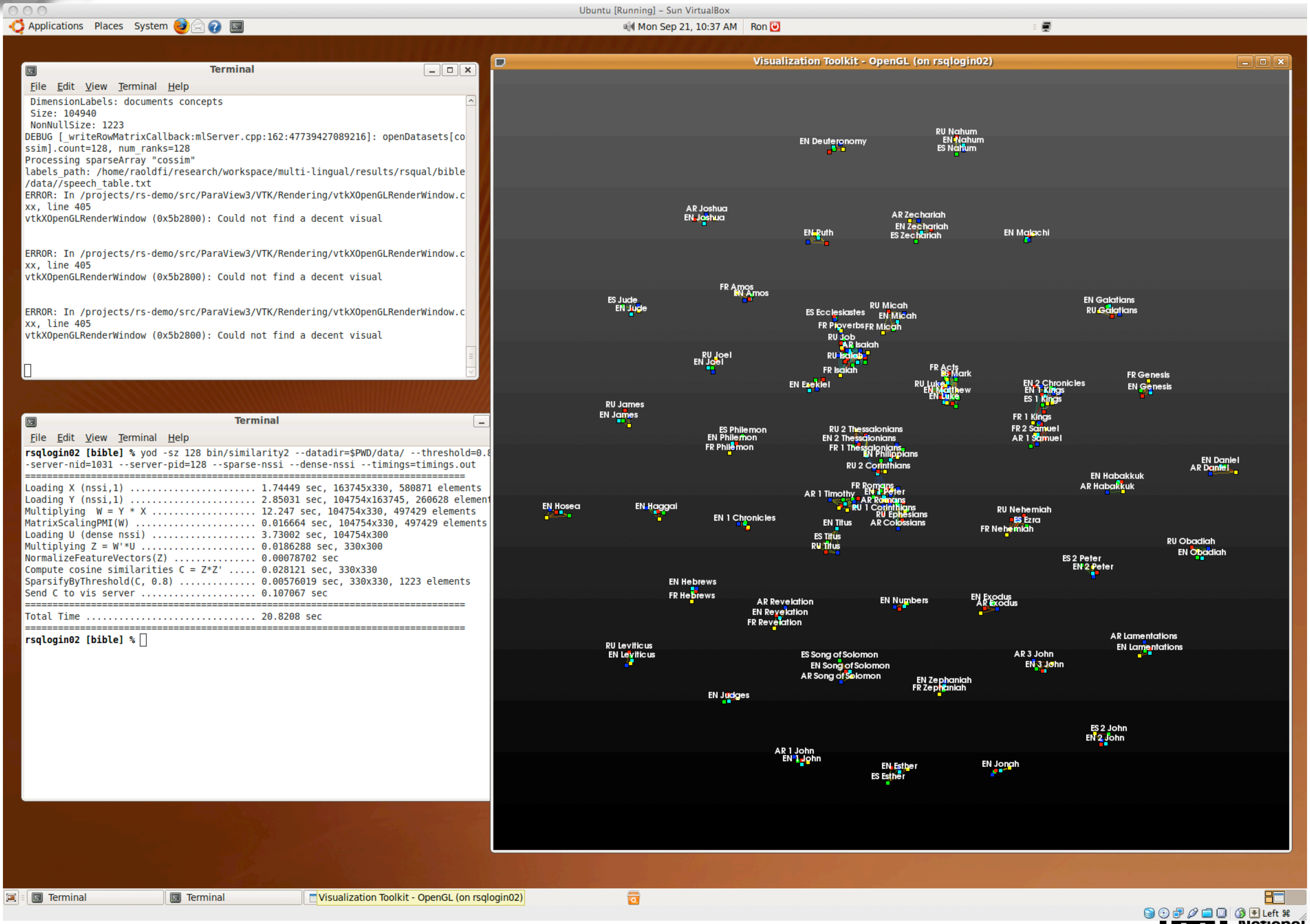
Integrating these systems for interactive jobs has never been done

Scaling Challenges for Multilingual Clustering

- Strong scaling exposes weaknesses in loading
  - Original methods for loading were not designed for production use.
- Improvements
  - Sparse Reads
    - Keep track of processor mapping information
    - Parallel I/O
  - Dense Reads
    - Convert to binary format
    - Parallel I/O
    - Data ordering
- Status on Red Storm (Cray XT4)
  - 250K docs of Europarl dataset requires 2048 nodes to execute (memory constrained)
  - At 4096 cores, we overwhelm network communication layer when reading input
  - Our target data set has over 1M docs



HPC clustering


Ubuntu [Running] - Sun VirtualBox
Applications
Places
System
Mon Sep 21, 10:37 AM
Ron
Terminal
File
Edit
View
Terminal
Help
Visualization Toolkit - OpenGL (on rsqlogin02)

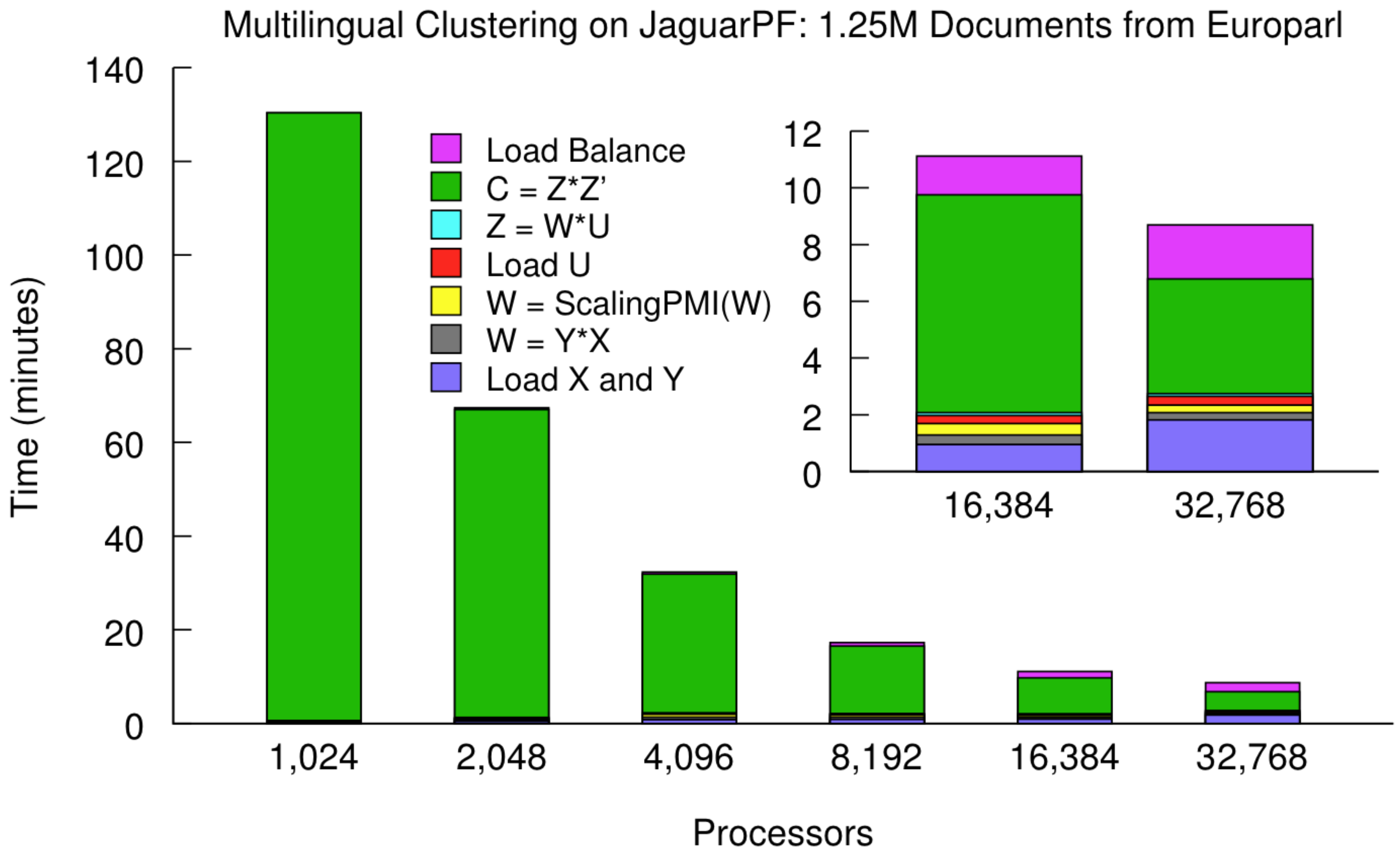
```
 DimensionLabels: documents concepts
 Size: 104940
 NonNullSize: 1223
DEBUG [_writeRowMatrixCallback:mlServer.cpp:162:47739427089216]: openDatasets[co
ssim].count=128, num_ranks=128
Processing sparseArray "cossim"
labels_path: /home/raoldfi/research/workspace/multi-lingual/results/rsqual/bible
/data//speech_table.txt
ERROR: In /projects/rs-demo/src/ParaView3/VTK/Rendering/vtkXOpenGLRenderWindow.c
xx, line 405
vtkXOpenGLRenderWindow (0x5b2800): Could not find a decent visual


ERROR: In /projects/rs-demo/src/ParaView3/VTK/Rendering/vtkXOpenGLRenderWindow.c
xx, line 405
vtkXOpenGLRenderWindow (0x5b2800): Could not find a decent visual


ERROR: In /projects/rs-demo/src/ParaView3/VTK/Rendering/vtkXOpenGLRenderWindow.c
xx, line 405
vtkXOpenGLRenderWindow (0x5b2800): Could not find a decent visual
```

```
rsqlogin02 [bible] % yod -sz 128 bin/similarity2 --datadir=$PWD/data/ --threshold=0.8
-server-nid=1031 --server-pid=128 --sparse-nssi --dense-nssi --timings=timings.out
================================================================================
Loading X (nssi,1) ...................... 1.74449 sec, 163745x330, 580871 elements
Loading Y (nssi,1) ...................... 2.85031 sec, 104754x163745, 260628 element
Multiplying  W = Y * X .................. 12.247 sec, 104754x330, 497429 elements
MatrixScalingPMI(W) ..................... 0.016664 sec, 104754x330, 497429 elements
Loading U (dense nssi) .................. 3.73002 sec, 104754x330
Multiplying Z = W'*U .................... 0.0186288 sec, 330x300
NormalizeFeatureVectors(Z) .............. 0.00078702 sec
Compute cosine similarities C = Z*Z' ..... 0.028121 sec, 330x330
SparsifyByThreshold(C, 0.8) ............. 0.00576019 sec, 330x330, 1223 elements
Send C to vis server .................... 0.107067 sec
================================================================================
Total Time .............................. 20.8208 sec
================================================================================
rsqlogin02 [bible] %
```

National Laboratories

Scaling Challenges for Multilingual Clustering

- Performance on JaguarPF (Cray XT5)
  - 1.25M docs of Europarl data set
  - With 32K cores, it takes 470 seconds

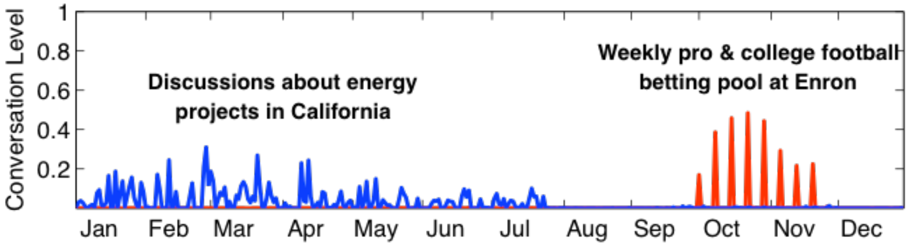


Sandia National Laboratories

Related Text Analysis Projects

- Discussion tracking in Enron emails
- Uncovering plots in text (scenario discovery)
- Network data exfiltration analysis
- Higher-order web link analysis
- Unsupervised part-of-speech tagging
- Multilingual sentiment analysis
- Identifying emerging keywords of interest

Identifying unusual activity in Enron emails



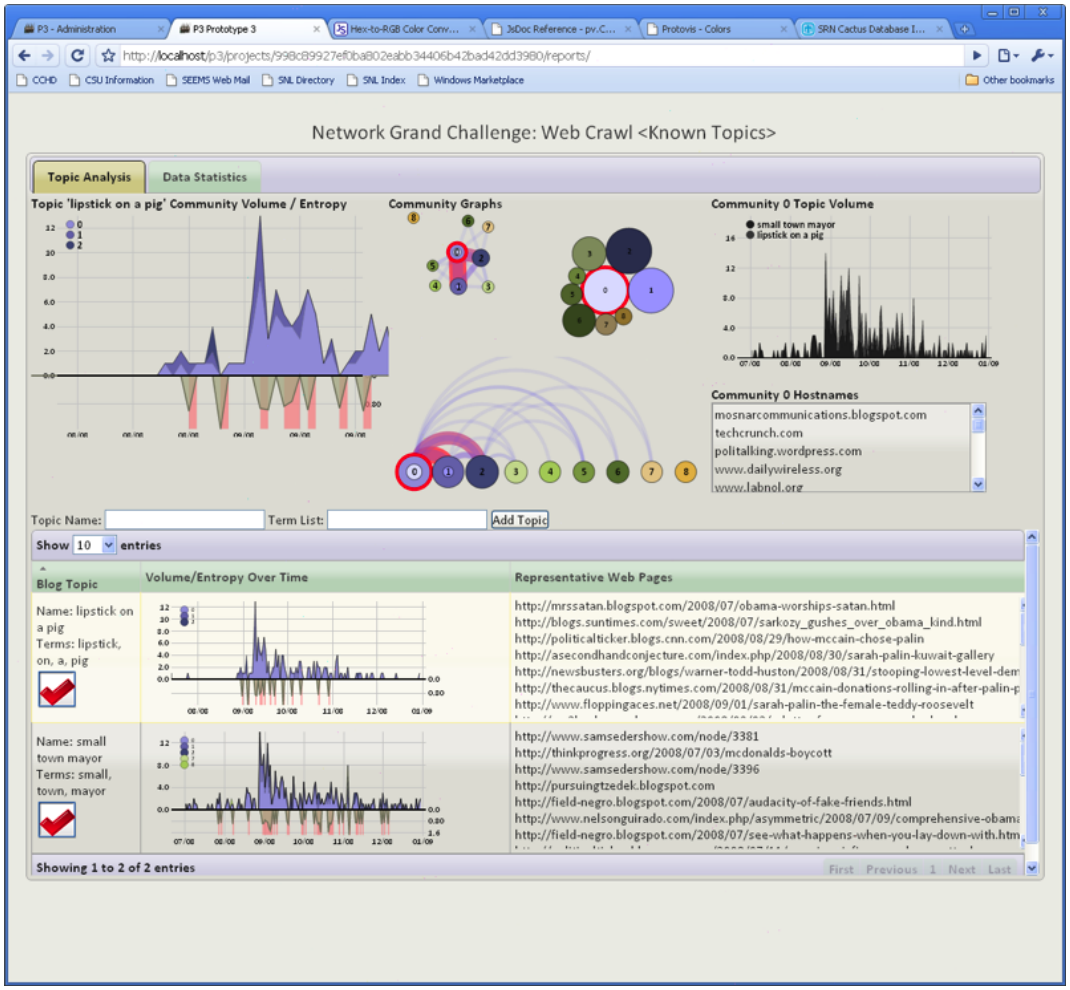
California, power, utilities, energy, utility, governor, market

games, week, missed, picked, prize, wins, scored, upsets

J. Steffes, S. Kean, J. Dasovich, R. Shapiro, P. Allen, ...

A. Pace, L. Campbell, C. Dean

Analysis tools for web forecasting



http://localhost/p3/projects/998c89927ef0ba802eabb34406b42bad42dd3980/reports/

Network Grand Challenge: Web Crawl <Known Topics>

Community 0 Hostnames

mosnarcommunications.blogspot.com
techcrunch.com
politalking.wordpress.com
www.dailywireless.org
www.labnol.org

Topic Name: Term List:

Show 10 entries

Blog Topic	Volume/Entropy Over Time	Representative Web Pages
Name: lipstick on a pig Terms: lipstick, on, a, pig		http://mrssatan.blogspot.com/2008/07/obama-worships-satan.html http://blogs.suntimes.com/sweet/2008/07/sarkozy_gushes_over_obama_kind.html http://politicalticker.blogs.cnn.com/2008/08/29/how-mccain-chose-palin http://asecondhandconjecture.com/index.php/2008/08/30/sarah-palin-kuwait-gallery http://newsbusters.org/blogs/warner-todd-huston/2008/08/31/stooping-lowest-level-dem http://thecaucus.blogs.nytimes.com/2008/08/31/mccain-donations-rolling-in-after-palin-p http://www.floppingaces.net/2008/09/01/sarah-palin-the-female-teddy-roosevelt
Name: small town mayor Terms: small, town, mayor		http://www.samsedershow.com/node/3381 http://thinkprogress.org/2008/07/03/mcdonalds-boycott http://www.samsedershow.com/node/3396 http://pursuingtzedek.blogspot.com http://field-negro.blogspot.com/2008/07/audacity-of-fake-friends.html http://www.nelsonguirado.com/index.php/asymmetric/2008/07/09/comprehensive-obama http://field-negro.blogspot.com/2008/07/see-what-happens-when-you-lay-down-with.htm

Showing 1 to 2 of 2 entries

Selected References

- US Patent Application SD11033, “Technique for Information Retrieval Using Enhanced Latent Semantic Analysis,” Peter Chew and Brett Bader, filed 2009.
- Bader, Kegelmeyer, and Chew. 2011. Multilingual sentiment analysis using latent semantic indexing and machine learning. Submitted to ACL.
- Bader and Chew. 2008. Enhancing multilingual latent semantic analysis with term alignment information. *COLING 2008.*
- Chew, Bader, and Abdelali. 2008. Latent Morpho-Semantic Analysis: Multilingual Retrieval with character N-grams and mutual information. *COLING 2008*.
- Chew, Kegelmeyer, Bader and Abdelali. 2008. The Knowledge of Good and Evil: Multilingual Ideology Classification with PARAFAC2 and Machine Learning. *Language Forum* (34), 37-52.
- Chew, Bader, Kolda and Abdelali. 2007. Cross-language information retrieval using PARAFAC2. *Proceedings of KDD 2007.*
- Chew and Abdelali. 2007. Benefits of the ‘massively parallel Rosetta Stone’: cross-language information retrieval with over 30 languages, *Proceedings of the Association for Computational Linguistics conference,* 2007.
- Chew, Verzi, Bauer and McClain. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability,* 2006, 68–74.

Brett Bader (bwbader@sandia.gov)
http://www.sandia.gov/~bwbader