NUBBI: An introduction and a non-examination of Sudanese Newspapers

Peter Landwehr

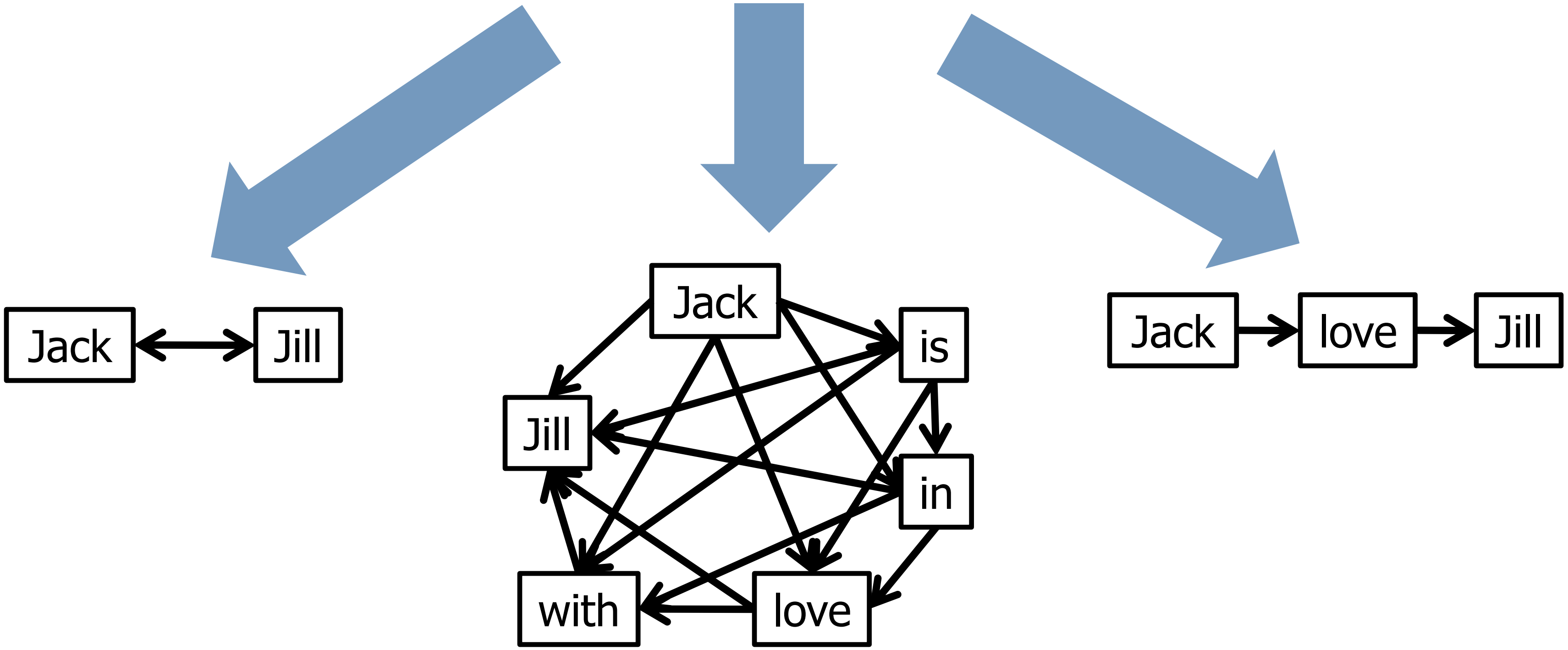
plandweh@cs.cmu.edu

What I’m covering

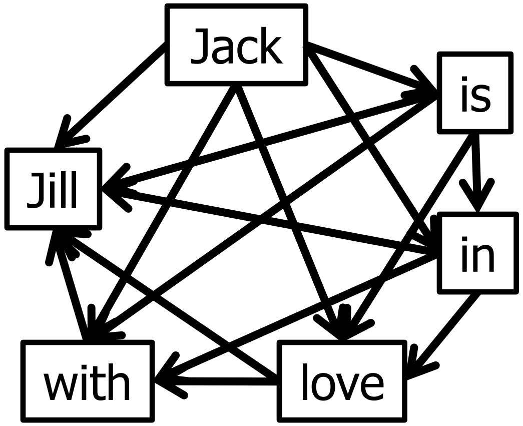
1. A difficulty with semantic networks
2. One way to try and resolve it.
3. My preliminary work with this method.

Semantic Networks!

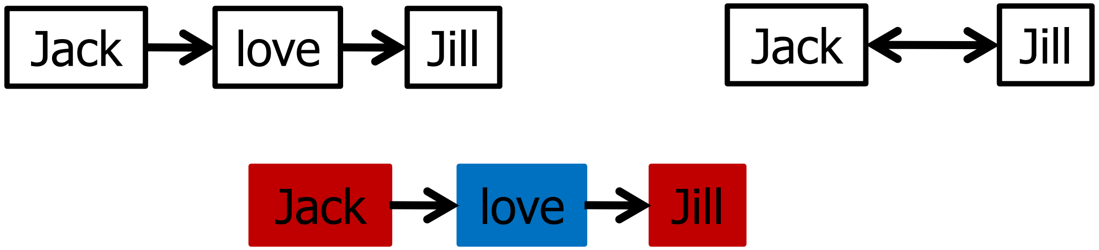
Jack is in love with Jill.



Map Theory



Jack
is
Jill
in
with
love



Jack → love → Jill

Jack ↔ Jill

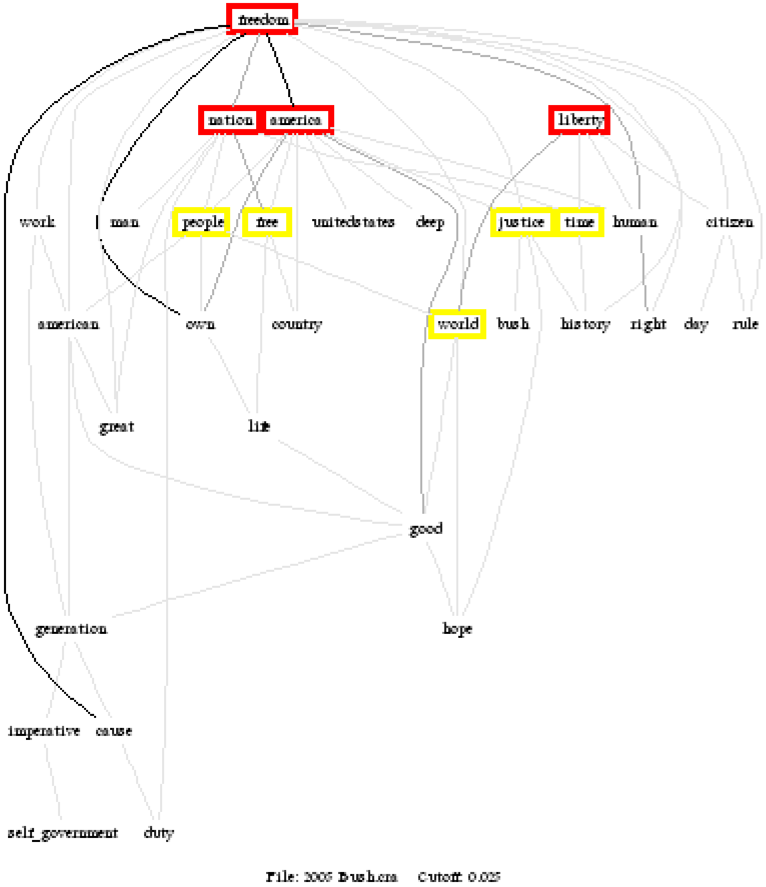
Jack → love → Jill

- Words in a document gain significance based on their juxtaposition with other words.
- Situate words in a network based on windowed proximity.
- Words can be classed into particular roles based on these network statistics.

[Carley 1997] [Carley & Kaufer 1993]

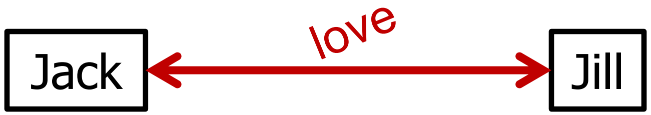
Centering Resonance Analysis

- Drop all terms that aren’t nouns
- Link nouns, noun phrases
- Compare betweenness centralities of words in networks to determine “resonance”.



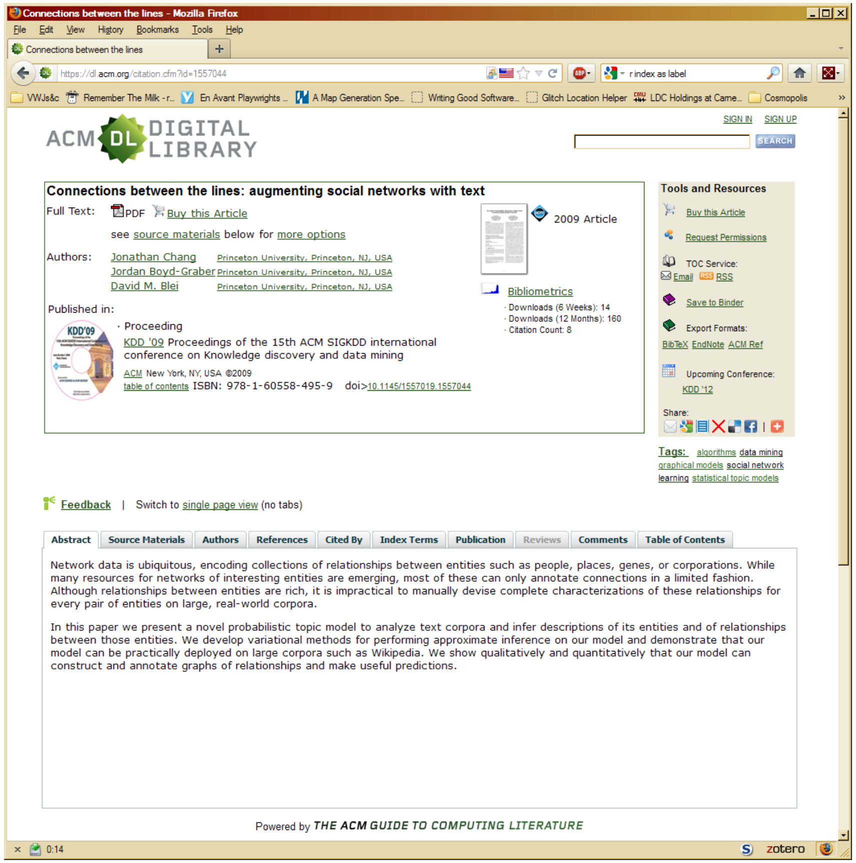
[Corman et al. 2002]

An appealing improvement



One way towards this: NUBBI

- Networks Uncovered By Bayesian Inference
- “Connections between the lines: augmenting social networks with text” by Chang, Boyd-Graber, & Blei. *KDD 2009*



Connections between the lines: augmenting social networks with text

Full Text: PDF Buy this Article

see source materials below for more options

2009 Article

Authors:
Jonathan Chang Princeton University, Princeton, NJ, USA
Jordan Boyd-Graber Princeton University, Princeton, NJ, USA
David M. Blei Princeton University, Princeton, NJ, USA

Bibliometrics
· Downloads (6 Weeks): 14
· Downloads (12 Months): 160
· Citation Count: 8

Published in:

· Proceeding
KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining
ACM New York, NY, USA ©2009
table of contents ISBN: 978-1-60558-495-9 doi>10.1145/1557019.1557044

Tools and Resources

Buy this Article
Request Permissions
TOC Service:
Email RSS RSS
Save to Binder
Export Formats:
BibTeX EndNote ACM Ref
Upcoming Conference:
KDD '12
Share:

Tags: algorithms data mining graphical models social network learning statistical topic models

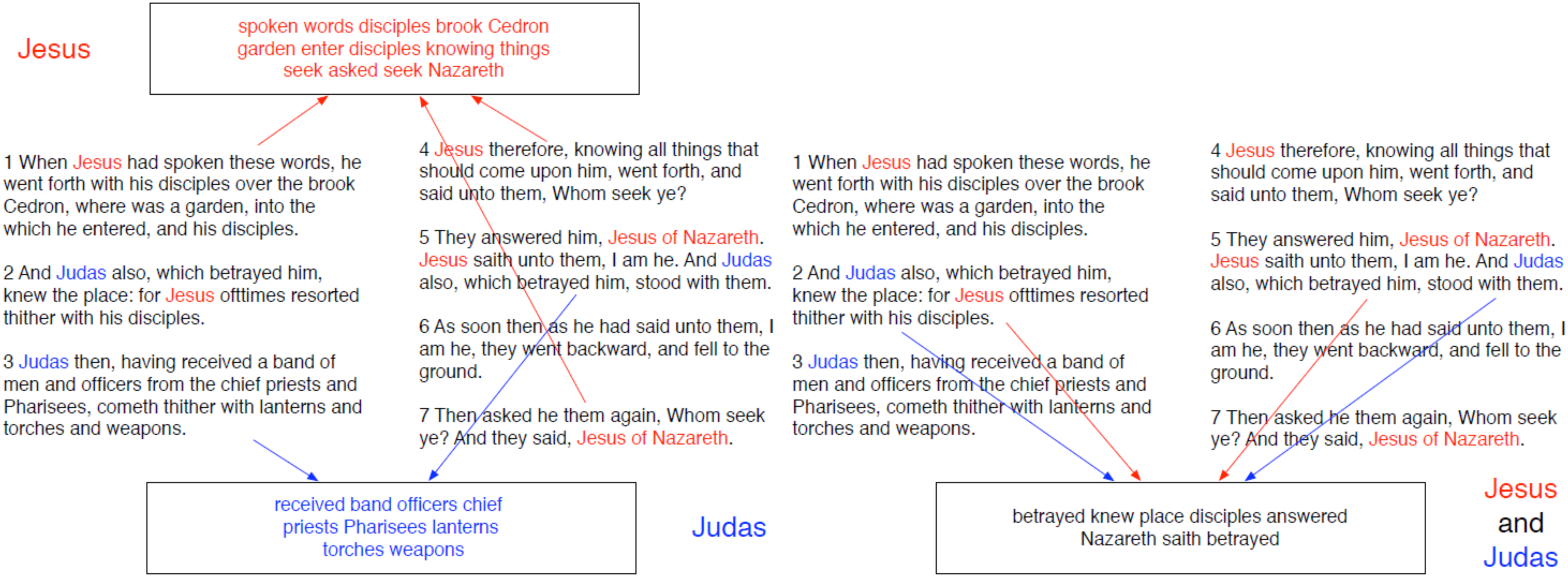
Feedback | Switch to single page view (no tabs)

Network data is ubiquitous, encoding collections of relationships between entities such as people, places, genes, or corporations. While many resources for networks of interesting entities are emerging, most of these can only annotate connections in a limited fashion. Although relationships between entities are rich, it is impractical to manually devise complete characterizations of these relationships for every pair of entities on large, real-world corpora.

In this paper we present a novel probabilistic topic model to analyze text corpora and infer descriptions of its entities and of relationships between those entities. We develop variational methods for performing approximate inference on our model and demonstrate that our model can be practically deployed on large corpora such as Wikipedia. We show qualitatively and quantitatively that our model can construct and annotate graphs of relationships and make useful predictions.

Powered by *THE ACM GUIDE TO COMPUTING LITERATURE*

Stealing a diagram…



Jesus

spoken words disciples brook Cedron garden enter disciples knowing things seek asked seek Nazareth

1 When Jesus had spoken these words, he went forth with his disciples over the brook Cedron, where was a garden, into the which he entered, and his disciples.

2 And Judas also, which betrayed him, knew the place: for Jesus ofttimes resorted thither with his disciples.

3 Judas then, having received a band of men and officers from the chief priests and Pharisees, cometh thither with lanterns and torches and weapons.

4 Jesus therefore, knowing all things that should come upon him, went forth, and said unto them, Whom seek ye?

5 They answered him, Jesus of Nazareth. Jesus saith unto them, I am he. And Judas also, which betrayed him, stood with them.

6 As soon then as he had said unto them, I am he, they went backward, and fell to the ground.

7 Then asked he them again, Whom seek ye? And they said, Jesus of Nazareth.

received band officers chief priests Pharisees lanterns torches weapons

Judas

1 When Jesus had spoken these words, he went forth with his disciples over the brook Cedron, where was a garden, into the which he entered, and his disciples.

2 And Judas also, which betrayed him, knew the place: for Jesus ofttimes resorted thither with his disciples.

3 Judas then, having received a band of men and officers from the chief priests and Pharisees, cometh thither with lanterns and torches and weapons.

4 Jesus therefore, knowing all things that should come upon him, went forth, and said unto them, Whom seek ye?

5 They answered him, Jesus of Nazareth. Jesus saith unto them, I am he. And Judas also, which betrayed him, stood with them.

6 As soon then as he had said unto them, I am he, they went backward, and fell to the ground.

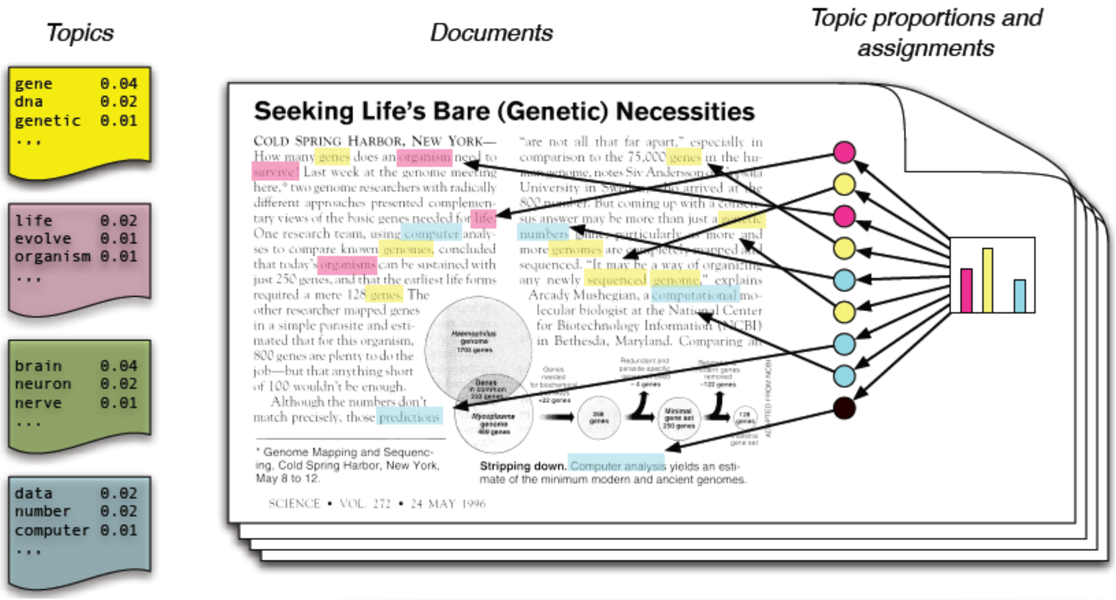
7 Then asked he them again, Whom seek ye? And they said, Jesus of Nazareth.

betrayed knew place disciples answered Nazareth saith betrayed

Jesus
and
Judas

Latent Dirichlet Allocation

- Decompose documents as mixtures of *N* topics
- Randomly assign topics to documents
- Based on random topic mixtures, randomly assign words to topics.
- Expectation Maximization or Gibbs sampling to converge.


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
Redundant and parasite-specific genes removed −4 genes
Related modern genes removed −122 genes
256 genes
Minimal gene set 250 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

[Blei et al. 2003][Blei 2011]

Idealized description of the algorithm

- Assumes text is grouped into entity and pair contexts
- For each entity context, NUBBI assigns a mixture of entity topics.
- For each pair context, NUBBI splits the words into those describing entities and those describing the pair.
- For each pair-specific word in a pair-specific context, NUBBI assigns a pair topic.

Idealized description of the analyst's role

- Define entity and pair counts
- Define three hyperparameters for topic proportions required by the model
  - Topic mixture
  - Pair mixture
  - Entity/pair selector mixture
- Select appropriate windowing parameters and clean text.

Issues with NUBBI

- Text in general
  - Deciding on context sizes. (e.g. Bible = verses)
  - Text cleaning.
  - Evaluating goodness of results.
- NUBBI-specific
  - # of entity topics = ?
  - # of pair topics = ?
  - Optimal mixing values
  - The best way of assigning multi-context text.
  - Relationships described in text can go beyond pairs to triples,

A personal test

- Eight years of news articles from the Sudan Tribune, documenting key activities in the region
- Detailed thesauri containing ~17000 different actors.
- These texts have been used for a variety of CASOS-affiliated work
  - T. Van-Holt and J. C. Johnson, “A Text and Network Analysis of Natural Resource Conflict in Sudan,” in *Proceedings Sunbelt XXXI*, St. Petersburg Beach, Florida, USA, 2011.
  - J. Diesner and K. M. Carley, “Mapping Socio-Cultural Networks of Sudan From Open-Source, Large-Scale Text Data,” in *Proceedings of the 29th Annual Conference of the Sudan Studies Association*, West Lafayette, Indiana, 2010.

Basic Notes

- Cleaned the text using AutoMap
  - Used standard AutoMap cleaning methods (clean white space, fix a predefined set of typos, convert British to American spellings, expand contractions & abbreviations, convert to lowercase, best-effort resolve pronouns and delete unresolved, remove punctuation and numbers.
  - "A stop list for general text" by Fox. *ACM SIGIR*, volume 24, issue 1-2. Fall 1989.
  - Specialized AutoMap thesaurus on Sudan. (Possible issue in order of application)
- Used the R implementation of NUBBI ("lda" package)
- 12 entity topics, 6 pair topics, windows of size 5.

Corpora Notes

Year	Articles in Corpora	Entities in Corpora	Meaningful Pairs	Vocabulary (Total)
2003	616	379	352	2007 (6539)
2009	1056	536	460	3571 (13283)
2010	1063	481	472	3641 (13644)

Topic Overlaps

- Pair topics
  - 2003 has no overlaps with 2008 or 2009
  - In an optimal matching, the 2008 and 2009 topics overlap by 24.6% (Top 25 words/topic.)
  - Negligible matches between entity pairs in each pair topic per year. (Top 25 entities/topic
- Entity topics
  - 20-24% optimal matching overlap between all entity topics. This says more about LDA than it does about our quality. (Top25 words/topic)
  - 25% and 22% overlaps of entities between 2003 and 2008 and 2009, but 35% overlap between 2008 and 2009

Future Work

- Make this preliminary work less preliminary by…
  - improving data cleaning
  - Improve model parameter choices
  - Full exploration of all years of the corpus
- Improve the "NUBBI experience" by
  - Integrating it with a text cleaning tool (AutoMap) for easier use.
  - Merge with HDPs to avoid having to set topic counts.
  - Experiment with other corpora to develop deeper recommendations for tuning.

References

1. D. M. Blei, “Introduction to Probabilistic Topic Models.”
2. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, Jan. 2003.
3. K. M. Carley, “Extracting team mental models through textual analysis,” *J. Organiz. Behav.*, vol. 18, pp. 533–558, 1997.
4. K. M. Carley and D. Kaufer, “Semantic Connectivity: An Approach for Analyzing Symbols in Semantic Networks,” *Communication Theory*, vol. 3, no. 3, pp. 183–213, 1993.
5. J. Chang, J. Boyd-Graber, and D. M. Blei, “Connections between the lines: augmenting social networks with text,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, 2009, pp. 169-178.
6. S. R. Corman, T. Kuhn, R. D. Mcphee, and K. J. Dooley, “Studying Complex Discursive Systems.,” *Human Communication Research*, vol. 28, no. 2, pp. 157-206, 2002.

The end