

Applying Natural Language Processing to Dear Colleague Letters

(with Prof. Dino Christenson)

Rushi Ganmukhi

Introduction

Goal: Want to find a relationship between Politicians, Interest Groups and Bills

- ❖ Which Politicians work with which Interest Groups on which Bills?
- ❖ Which Politicians and Interest Groups work typically work together?
- ❖ What issues are these Politicians and Interest Groups are interested in?
- ❖ Are some combinations more likely to succeed?

DPC Original Problem

Amicus Curiae Network

- SCOTUS amicus curiae briefs
- Support on an issue before the Court
- Co Signing the brief denotes a tie
- Network corroborated with interviews



Dear Colleague Letters

- ❖ Letters between House Representatives asking to support, oppose or cosponsor a bill.
- ❖ Can be sent by one-to-many, many-to-many, many-to-one or one-to-one
- ❖ Often contain names of interest groups involved
- ❖ These are publicly available letters and are thought to contain little valuable information
- ❖ Can we use the available information in these letters to gain further insights?

The data: Dear Colleague Letters

From: Duncan, Jeff

Sent: Thursday, June 29, 2006 1:42 PM

To: Dear Colleague

Subject: Dear Colleague;Energy;Environment;Budget;White House Agrees that HR 4761 is a Budget Buster

Attachments: SAP on HR4761.pdf

June 29, 2006

Dear Colleague:

I would like to call your attention to what the Bush Administration had to say today in its official Statement of Administration Policy, commenting on the budget-busting implications of H.R. 4761:

"...the Administration strongly opposes revenue- sharing provisions that do not incentivize production and that would reduce Federal receipts relative to current law and have a long-term impact on the Federal deficit. The Administration's preliminary estimate is that the revenue-sharing provisions of H.R. 4761 would reduce Federal receipts by several hundred billion dollars over 60 years..." [Emphasis Added]

A copy of the full Administration Statement of Administration Policy is attached. I urge you to vote against this ill-conceived legislation, that will not only lift the current moratoria on drilling, but which will also result in a substantial loss in revenues to the federal Treasury for decades to come.

Sincerely,

Edward J. Markey

The Problem

Problem: Given the corpus of Dear Colleague letters can we find some interesting relationships

- 1) Parse out fields(To,From,Subject etc.)
- 2) Fuzzy Matching of names (Bills, Reps., Interest Groups)
- 3) Sentiment Analysis
- 4) Topic Modeling
- 5) Look at mined data and formulate a method to reach a conclusions (discussed later)

Basic Email Parsing

- ❖ Emails are in text files categorized by Congress and Session
- ❖ Separate Emails in file by breaking along “From” field
- ❖ Parse out “From”, “To”, “Subject”, “Session”, “Congress” as well as content for each email

Fuzzy Matching of Names

- ❖ Need to match Representative names, Interest Groups and Bill names in the context of each email to the names available in Prof. Christenson's DB
- ❖ Fuzzy matching of Representative names such as "Representative John Smith", "esteemed colleague John Smith" gentleman John" or "John Smith (MA).
- ❖ Matching of Bill names and numbers of the form HR XXX, HR/SB XXXX, Amdt XXX

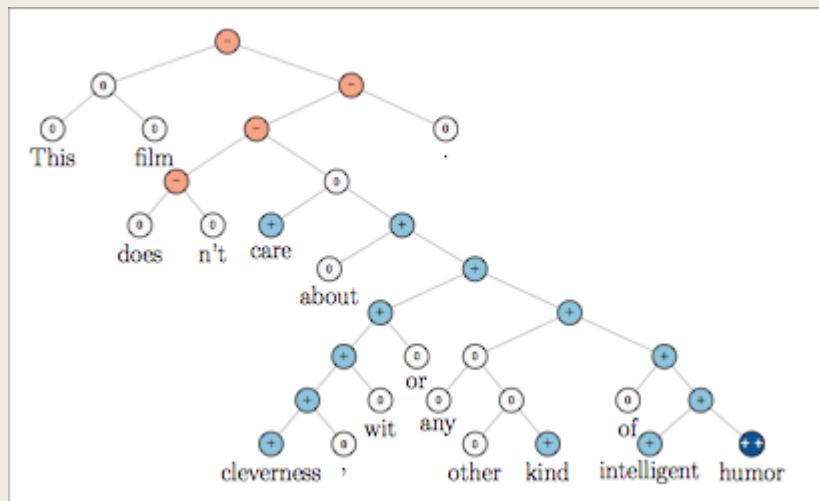
Sentiment Analysis

- ❖ Wish to see whether individual sentences in the emails or the overall emails contain strong positive or negative sentiment.
- ❖ Perhaps sentiment can be related success of the bill, party or representative.
- ❖ Stanford Core NLP Sentiment Analysis
- ❖ Based on: Recursive Deep Models for Semantic Compositionality over and Sentiment Treebank
- ❖ Model uses recursive neural-networks

Sentiment Analysis

- ❖ Categorizes into 5 sentiment classes (-- , - , 0 , + , ++)
- ❖ Looks at multi-grams of words in order to better understand meaning and is the only sentiment analysis tool that can accurately understand negation.

[2]



Topic Modeling

- ❖ Goal is to automatically discover topics set for documents as well as a distribution of topics for each individual document
- ❖ probabilistic topic modeling-> a class of algorithms used to annotate large corpora with information regarding the topic or theme of each document.[1]
- ❖ analyze words in the original text to find topics and categorizes documents by discovered themes
- ❖ can also be used to discover patterns in genetic data, images and social networks

Document Modeling

- ❖ tf-idf: Type of word frequency (frequency of term in document[normalized])
* $\log(\# \text{ of docs} / \# \text{ docs containing the term})$
- ❖ Latent Semantic Indexing: Uses Singular Value Decomposition along w/ tf-idf
- ❖ pLSI: Probabilistic version of Latent Semantic Indexing
- ❖ focus on one particular topic model: Latent Dirichlet Allocation (LDA)
[blei, D., ng, a., Jordan, m. latent Dirichlet allocation. *J. Mach. Learn. Res.* 3 (January 2003), 993–1022.]

LDA (intuition)

- ❖ model each document in a corpus as a finite mixture of underlying topics
- ❖ topics are modeled as a mixture over an underlying set of topic probabilities
- ❖ topic: a distribution over a fixed vocabulary
- ❖ all the documents in the collection share the same set of topics, but each document exhibits those topics in different proportion
- ❖ LDA is a hierarchical Bayesian Model

LDA (example)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Sample topics
generated from a
document.

Titles are created from
the words
corresponding to each
topic

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

LDA (generative process)

- Imaginary and random process by which the LDA model assumes that the documents we created

Generative Process:

Randomly Choose a distribution over topics

For each word in document:

1. Randomly choose a topic
2. Randomly choose a word over the distribution of the vocabulary



LDA (generative process diagram)

Topics

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
...

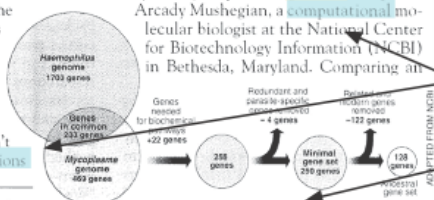
Documents

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 scientific 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 all



* 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

Topic proportions and assignments

Left to Right

- ❖ Topic proportions
- ❖ Topic assignments per document
- ❖ Document & words
- ❖ Topics and word distributions

LDA (prior assumptions)

- Bag of words
order of words does not matter
- Number of topics is fixed
- Order of Documents does not matter

LDA (structure)

- ❖ Observed -> documents, words of each document
- ❖ Latent(hidden) -> topics, per-document topic distributions, the per-document per-word topic assignments, topic structure
- ❖ Need to infer the hidden structure from the observed documents (i.e. what hidden structure generated the observed documents?)

LDA (probabilistic model)

- ❖ Treat data as arising from a generative process that includes hidden-variables
- ❖ defines a joint probability distribution over observed and hidden variables
- ❖ compute posterior distribution as conditional distribution of hidden variables given the observed variables.

LDA (formal definition, variables)

- ❖ $\beta_{1:k}$; Each β represents a topic and is the distribution over the words
- ❖ θ_d ; Topic proportions for d th document, $\theta_{k,d}$ is the topic proportion for topic k in document d
- ❖ z_d , Topic assignment for document d , $z_{d,n}$ topic assignment for n th word in document d
- ❖ w_d , observed words for document d , $w_{d,n}$ is n th word in document d

LDA (formal definition, variables [image])

Beta_k

Topics

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
...	

Documents

w_d

Topic proportions and assignments

Θ_d

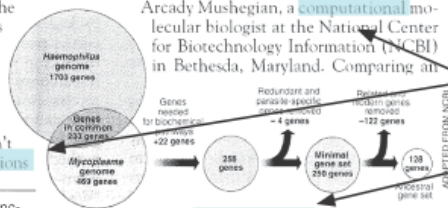
z_d

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

LDA (formal definition, generative process)

- ❖ Probabilities for generative process relate to the following joint probabilities

$$\begin{aligned} p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\ = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \\ \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right). \end{aligned}$$

- ❖ Topic Proportions ($z_{d,n}$) depend on per-document topic proportions (θ_d)
- ❖ Observed word ($w_{d,n}$) depend on topic assignment ($z_{d,n}$) and topics ($\beta_{1:k}$)
- ❖ $p(\theta_d)$ is modeled as a Dirichlet distribution as it is conjugate to the multinomial distribution


LDA (formal definition, computation of posterior)

$$\begin{aligned} p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) \\ = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \end{aligned} \quad (2)$$

- ❖ Numerator is joint distribution of the variables
- ❖ Denominator is marginal probability of observations, probability of seeing corpus under any topic model

LDA (formal definition, computation of posterior)

- ❖ Intractable -> number of possible topic structures is exponentially large, theoretically could be calculated by summing the joint distribution over every possible hidden structure
- ❖ Denominator -> need to marginalize over all hidden variables [3]

$$p(\mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta,$$


LDA (formal definition, computation of posterior)

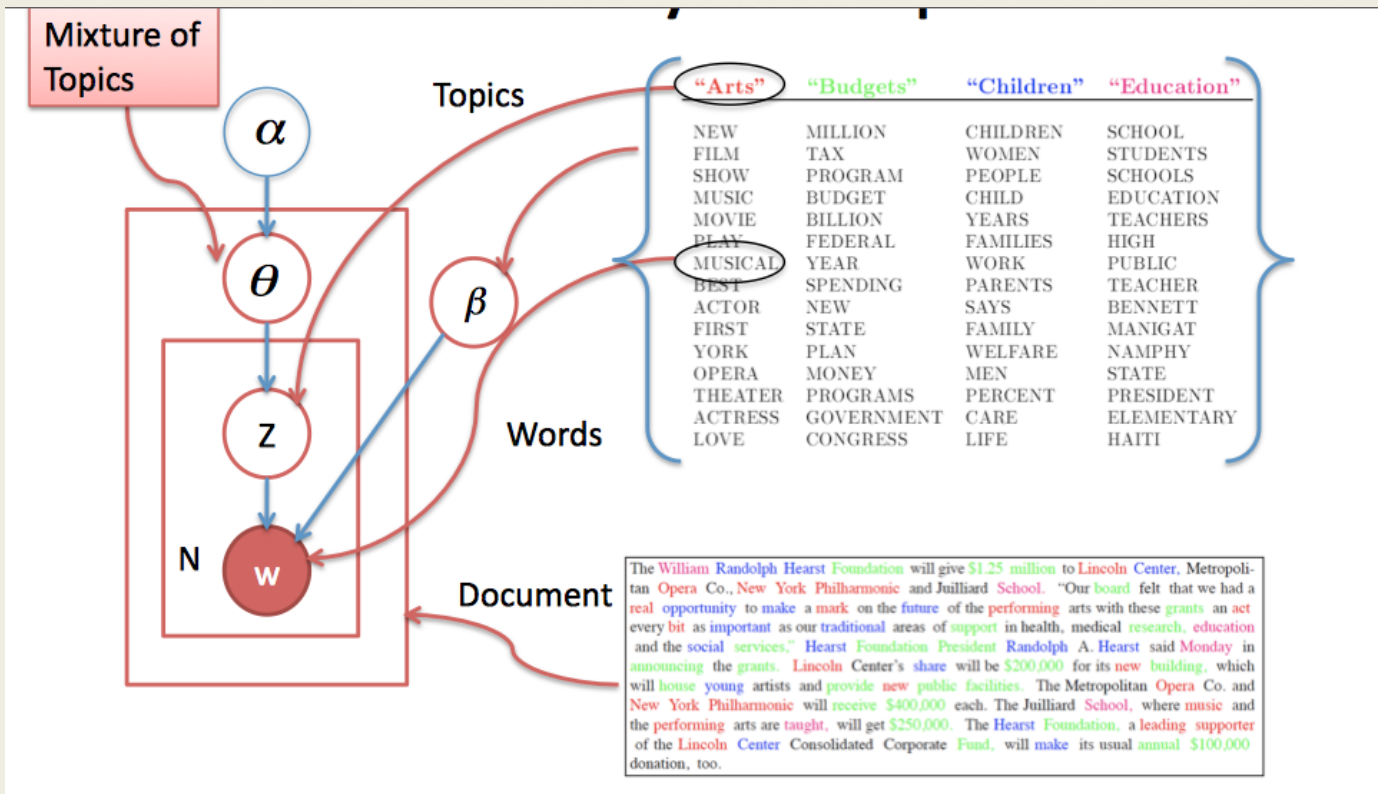
Two methods to compute denominator

- ❖ Both methods perform a structured search over the possible topic structure space guided by the observed documents
- ❖ Sampling -> collect samples from posterior in order to approximate its distribution, Gibbs sampling is commonly used
- ❖ Variational -> assume parameterized distributions over the hidden structure and find the member of the family that is closest to the posterior [1]

LDA (graphical model)

alpha is a parameter for the dirichlet distribution

w is marked in red as it is the only observed variable



[3]

Inner plate represents words in a document

Outer plate represents documents in the corpus

Roads Ahead

- ❖ Problem is more exploratory/open ended
- ❖ Depending on how the data looks, I can take different paths:
 - 1) Improvements to the LDA model
 - 2) Apply other language processing algorithms to mine data
 - 3) Find an interesting relationship and further explore it
 - 4) Use gained knowledge to improve the Amicus Curia Network

LDA (improvements)

❖ Relax Assumption that number of topics is fixed

[teh, y., Jordan, m., beal, m., blei, D. hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* 101, 476 (2006), 1566–1581.]

❖ Relax Bag of words assumption. Condition generated word on previous word

[Wallach, h. Topic Modeling: beyond bag of words. in *Proceedings of the 23rd International Conference on Machine Learning* (2006).]

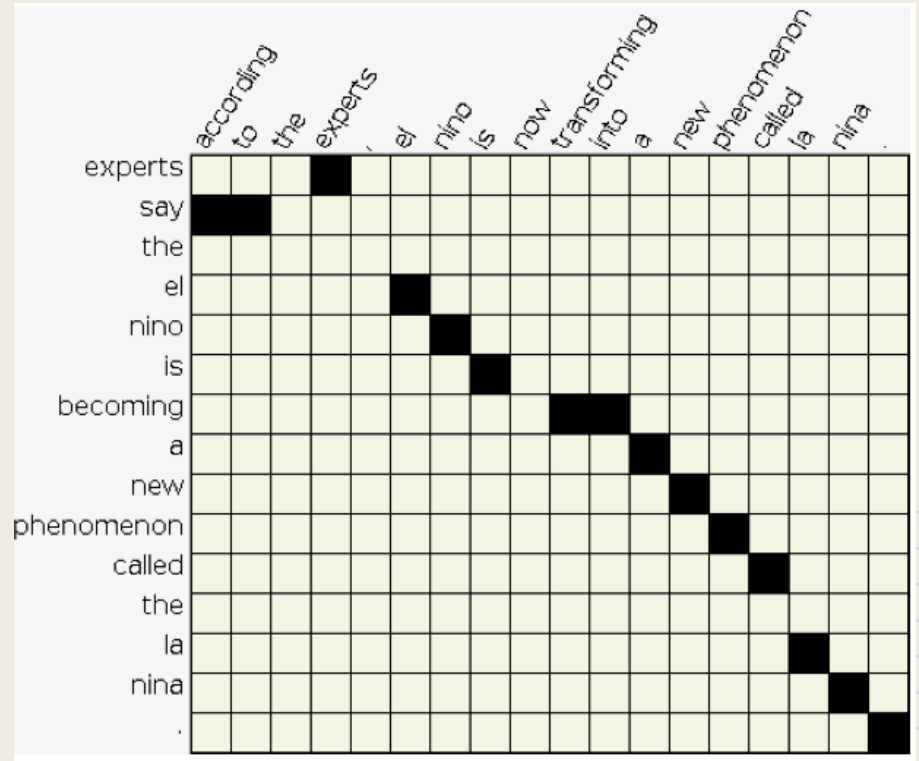
❖ Incorporate Meta-data

[rosen-Zvi, m., griffiths, t., steiyvers, m., smith, P., the author-topic model for authors and documents. in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (2004), auai Press, 487–494.]

[Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression]

Algorithms

- ❖ A word alignment is a pairing of a word in one word string with its counterpart in another word string. [4]
- ❖ For two word strings S_1 and S_2 , a sentence alignment of S_1 to S_2 is a set of word alignments where the first word is in S_1 and the second is in S_2



Sentence Alignment

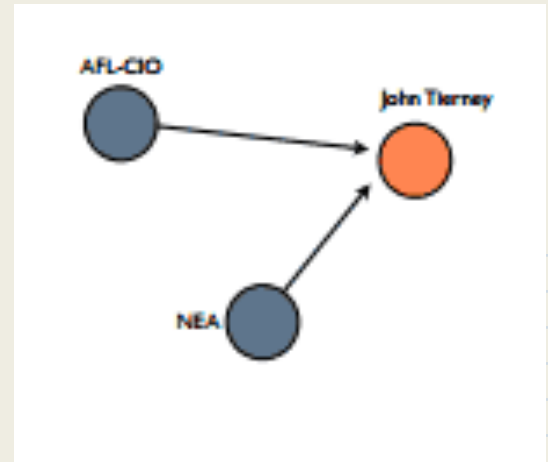
- ❖ Used in machine translation, but paraphrasing can be thought of a special case of translation
- ❖ Look at POS, dependencies, stemming and many other things to accomplish this, both supervised and unsupervised methods
- ❖ Sentence Alignment -> Can we pull out paraphrased sentences that correspond to the same bill/interest group/representative?

Further Explore a data-Relationship

- ❖ A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases <http://www.stanford.edu/~jgrimmer/ExpAgendaFinal.pdf>
- ❖ Based on frequency of topics and collaborations can we determine an agenda for each Representative/Interest Groups?
- ❖ By looking at which bills passed/failed can we determine which Representatives/Interest Groups hold more influence?

Amicus Curia Network

- Can we use the new data to help form or reinforce Links in the Amicus Curia Network?





Questions? Suggestions?

References

- [0] Blei, D., Ng, A., Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (January 2003), 993–1022.
- [1] <http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf>
- [2] Recursive Deep Models for Semantic Compositionality over and Sentiment Treebank; http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf
- [3] <http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf>
- [4] <http://www.ling.ohio-state.edu/~scott/talks/paralign/paralign.pdf>

Timeline

- 1) Parse out fields(To,From,Subject etc.) -> Done
- 2) Fuzzy Matching of names (Bills, Reps., Interest Groups) -> 3/17-3/23
- 3) Sentiment Analysis -> 3/17-3/23
- 4) Topic Modeling -> 3/24-3/30
- 5) Look at mined data and formulate a method to reach a conclusions
 - > Discuss with Profs. first week of April
 - > April 1 onward