Topic Distance and Coherence for Latent Dirichlet Allocation

Renata Chai Sept. 23, 2016

Topic Model

Definition: A tool to extract thematic structures in a discrete data collection.

"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

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.

Topic Model

Application

- Searching
- Classification
- Similarity and Relevance Judgment
- Summarization

Latent Dirichlet Allocation

Topic Distance

"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

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.

Topic Distance

Similar Topics

- Topic 1
- Program
- Algorithm
- Command
- Computation
- Software
- Graphics

…

- Topic 2
- Software
- Algorithm
- Command
- Engineering
- Program
- UI

…

- Topic 3
- Program
- Computation
- Command
- Computation
- UI
- Software

…

- Topic 4
- Program
- Command
- Algorithm
- Computation
- Graphics
- Software

…

Topic Coherence

Coherent Topic

- Topic – Computer Science
- Program
- Algorithm
- Command
- Computation
- Software
- Graphics

…

Incoherent Topic

- Topic – ?
- Program
- Planet
- Human
- Bank
- Fish
- Olympic

…

Motivation

In the usage of LDA, to improve topic distance and coherence

- What corpus type should we choose?
- What the number of topics should we choose?
- To represent a topic, how many top words should we choose?

Experiment

Varied the number of topics, the number of top words, and corpus types

Distance Measures:

- Distribution-based distance measures
- Ranking - based distance measures
- Set-based distance measures
- Vector-based distance measures

Coherence Measures:

- Co-occurrence based coherence measures
- WordNet Topic coherence measures

Outline

- Background
- Distance Experiments
- Coherence Experiments
- Discussion

Outline

- Background
  - Corpus Representation
  - Latent Dirichlet Allocation
  - Distance Measures
  - Coherence Measures
- Distance Experiments
- Coherence Experiments
- Discussion

Corpus Representation

Corpus: A collection of documents (i.e. 20000 news collections)

Raw Corpus $\longrightarrow$ Input ???

How to use a small amount of information to represent a large corpus?

- What **features** should be extracted to represent a corpus?
- What **relationships** should be kept?

Corpus Representation

Features:

Discrete units: Words, Documents   (A document: a string with >=2 words)

Relationships:

- Words belong to documents
- Semantical relevance between documents (new or current)
  a) Searching: Relevance between querying strings(new documents) and documents
  b) Relevance Recommendation: Relevance between current documents

Vector Space Model

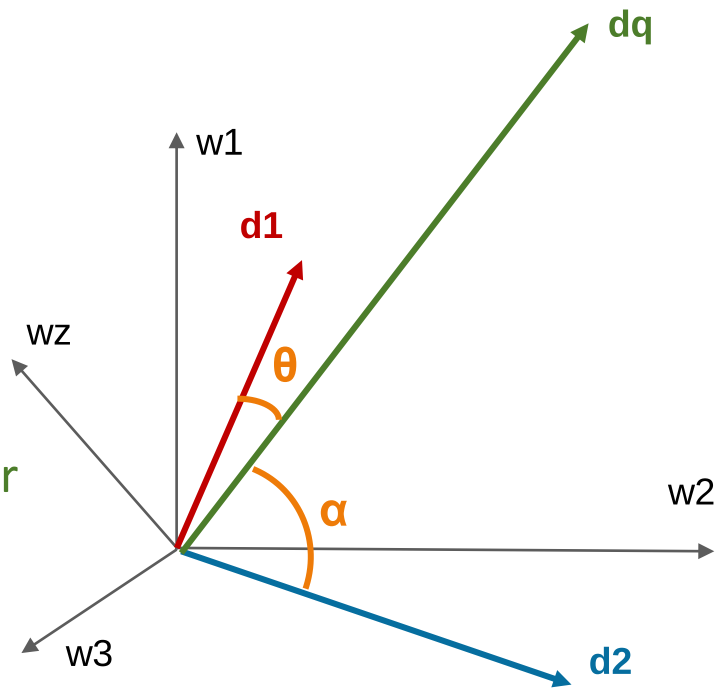
- Vocabulary: All unique words in the corpus.

  Each word is given an arbitrary ID

  i.e. sun: 1, night: 2, day: 3 … space: z

- Vocabulary Space – Each word is a unit vector

- Document: (1: c_1, 2: c_2, 3: c_3 …… z: c_z)



Different corpus types use different methods to determine the weight c for each word in a document vector

Corpus Types

- Binary – Whether a word exists in the doc or not (1:0, 2:1, 3:1, 4:0…z:0)
- Bow – How many times a word appears in a doc (1:0, 2:4, 3:10, 4:0…z:0)
- Tfidf – Specificity of a word to a doc (1:0, 2:0.33, 3:0.52,4:0…z:0)
  - Term frequency inverse document frequency
  - Weight (w_i) = frequency(w,d) $\times$ $\log \frac{Total\ number\ of\ docs}{Number\ of\ docs\ where\ w\ appears}$

Corpus Implementation

- Vocabulary: a dictionary {id: word, id: word…}

- Document: a list of (word_id, weight value)

words with weight value 0 are ignored in the list

[(1,0.22), (3,0.5), (4,0.3), (27,0.78)…]

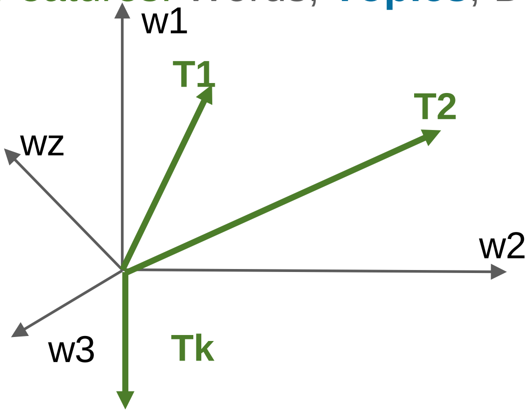
- Corpus: a list of document

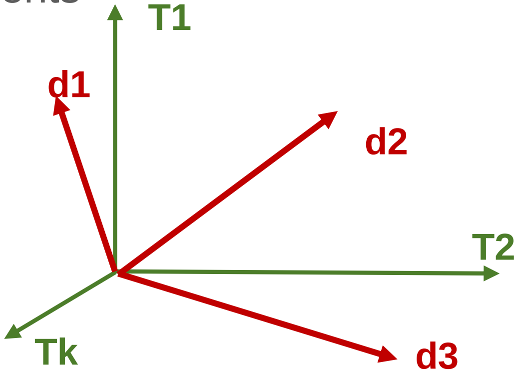
Dimension Reduction

Vocabulary Size: 10,000 ~ 100,000

How to reduce dimension?

Features: Words, **Topics**, Documents





Vector Space: Vocabulary Space -> **Topic Space** -> Documents

Dimension Reduction

Vocabulary Size: 10,000 ~ 100,000

How to reduce dimension?

Features: Words, **Topics**, Documents

How to extract topics from a corpus? How to define weight values?

- Topic ($word_1$:c_1, w_2:c_2, w_3:c_3….w_z:c_z)
- Document ($topic_1$:c_1, t_2:c_2, t_3:c_3…w_z:c_z)

Vector Space: Vocabulary Space -> **Topic Space** -> Documents

Topic Model

- Vector Space Approach:

Singular Value Decomposition (D: document T: topic W: word)

$$\text{D-W Matrix} = \text{D-T Matrix} \times \text{T Strength Matrix} \times (\text{W-T Matrix})^{\text{T}}$$

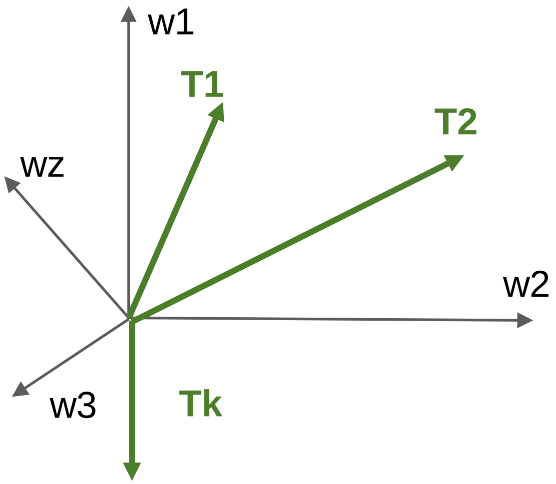
- Probability Distribution Approach:

Generative Probabilistic Model

Latent Dirichlet Allocation

Probability Distribution Representation

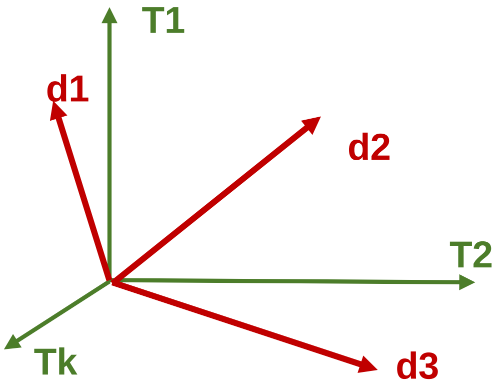
A topic is a multinomial distribution over the vocabulary



	Topic 1	Topic 2	Topic 3
computer	0.035	8.3e-5	2.3e-5
earth	3.5e-5	0.033	3.4e-7
file	0.003	3.2e-8	1.0e-8
organic	2.7e-4	0.010	9.2e-5
planet	1.9e-9	1.3e-7	0.012
…	…	…	…
Sum	**1**	**1**	**1**

Probability Distribution Representation

A document is a multinomial distribution over topics



	Topic 1	Topic 2	Topic 3	Sum
Doc 1	0.87	0.09	0.04	**1**
Doc 2	0.05	0.92	0.03	**1**
…	…	…	…	

Bayes' Theorem

- **Joint Probability**

For two random variables A & B, joint probability of A & B is the probability of the co-occurrence of A & B

i.e. $\mathrm{P}(\text{Red} \,\&\, \text{Bag1}) = 0.3$

If A and B are independent,

$\mathrm{p}(A,B) = \mathrm{p}(A) \times \mathrm{p}(B)$

Balls->	Red	Blue	Total
Bag 1	0.3	0.1	0.4
Bag 2	0.4	0.2	0.6
Total	0.7	0.3	1

Bayes' Theorem

- **Marginal Probability**

Assume A and B (B_1, B_2…B_n)

$$P(A) = \sum_{i=1}^{n} P(A, B_i)$$

Balls->	Red	Blue	Total
Bag 1	0.3	0.1	0.4
Bag 2	0.4	0.2	0.6
Total	**0.7**	0.3	1

i.e.

$P(Red) = \sum_{i=1}^{2} P(Red, Bag_i) = 0.3+0.4 = 0.7$

Bayes' Theorem

- **Conditional Probability**

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$\text{i.e.} P(\text{Red}|\text{Bag1}) = \frac{P(Red,Bag1)}{P(Bag1)} = \frac{0.3}{0.4} = \frac{3}{4}$$

Balls->	Red	Blue	Total
Bag 1	0.3	0.1	0.4
Bag 2	0.4	0.2	0.6
Total	0.7	0.3	1

Bayesian Inference

$$\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(X|\theta)p(\theta)}{p(X)} \propto \boldsymbol{p}(\boldsymbol{X}|\boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta})$$

posterior $\propto$ **Likelihood Function** $\times$ **Prior**

Sequential: Posterior – New Prior

Generative Probabilistic Modeling

Observed data – Observed variables (X)

a) Propose a generative process of observed variables
b) The generative process involves hidden parameters (θ)
c) The generative process defines the joint probability of observed variables and hidden parameters - $p(\theta \cup X)$

Goal – Compute and maximize the conditional distribution of the hidden parameters given observed variables

$$\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{\boldsymbol{p}(\boldsymbol{\theta} \cup \boldsymbol{X})}{\boldsymbol{p}(\boldsymbol{X})}$$

Maximize

Latent Dirichlet Allocation

	Topic 1	Topic 2	Topic 3
computer	0.035	8.3e-5	2.3e-5
earth	3.5e-5	0.033	3.4e-7
file	0.003	3.2e-8	1.0e-8
organic	2.7e-4	0.010	9.2e-5
planet	1.9e-9	1.3e-7	0.012
program	0.025	9.2e-6	7.2e-6
space	1.3e-8	5.4e-6	0.008
soil	5.6e-5	0.009	4.3e-7
universe	7.8e-6	1.9e-7	0.065
...	...	...	...

- **A topic is a distribution over the vocabulary**
- **A document is a distribution over topics**
- **Both distributions are generated by dirichlet processes**

→

Topic 1	Topic 2	Topic 3
computer program File ...	earth organic soil ...	universe planet space ...

	Topic 1	Topic 2	Topic 3
Doc 1	0.87	0.09	0.04
Doc 2	0.05	0.92	0.03

Latent Dirichlet Allocation

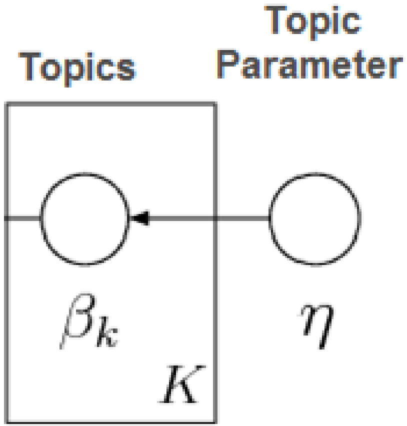
Doc 1: computer program…
Doc 2: earth soil…
Doc 3: sun universe…
Doc 4: …
…
Doc m: ...

	Topic 1	Topic 2	Topic 3
computer	0.035	8.3e-5	2.3e-5
earth	3.5e-5	0.033	3.4e-7
file	0.003	3.2e-8	1.0e-8
organic	2.7e-4	0.010	9.2e-5
planet	1.9e-9	1.3e-7	0.012
program	0.025	9.2e-6	7.2e-6
space	1.3e-8	5.4e-6	0.008
soil	5.6e-5	0.009	4.3e-7
universe	7.8e-6	1.9e-7	0.065
…	…	…	…

	Topic 1	Topic 2	Topic 3
Doc 1	0.87	0.09	0.04
Doc 2	0.05	0.92	0.03
…	…	…	…

LDA - Corpus Generation Process

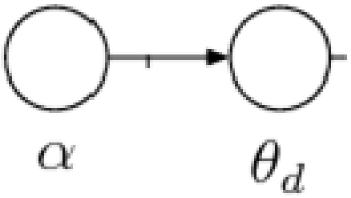
- $\boldsymbol{\beta}_k$: A topic/A probability distribution over the vocabulary
- There are K $\boldsymbol{\beta}$



	Topic 1	Topic 2	Topic 3
computer	0.035	8.3e-5	2.3e-5
earth	3.5e-5	0.033	3.4e-7
file	0.003	3.2e-8	1.0e-8
organic	2.7e-4	0.010	9.2e-5
planet	1.9e-9	1.3e-7	0.012
program	0.025	9.2e-6	7.2e-6
space	1.3e-8	5.4e-6	0.008
soil	5.6e-5	0.009	4.3e-7
universe	7.8e-6	1.9e-7	0.065
…	…	…	…

LDA - Corpus Generation Process

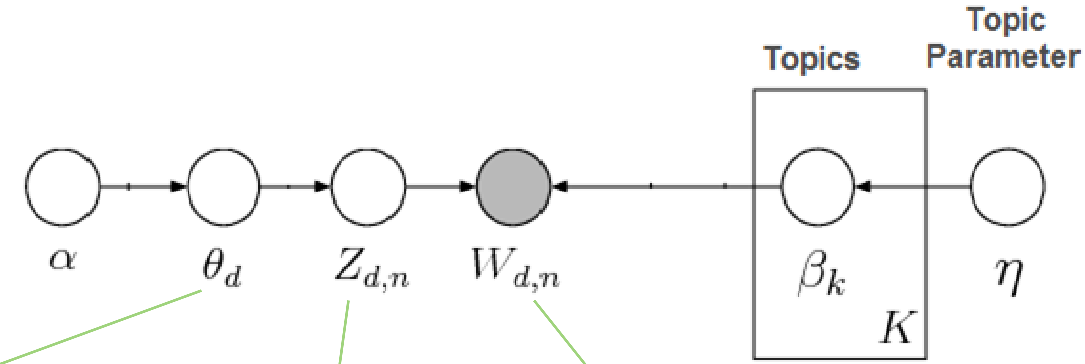
- To generate a document, a probability distribution over topics $\boldsymbol{\theta}_{\mathrm{d}}$ is generated



	Topic 1	Topic 2	Topic 3
Doc 1	0.87	0.09	0.04

LDA - Corpus Generation Process

- To generate a word in the document:
- Based on $\boldsymbol{\theta}_d$, a topic Z is selected
- Based on $\boldsymbol{\beta}_z$, a word W is selected



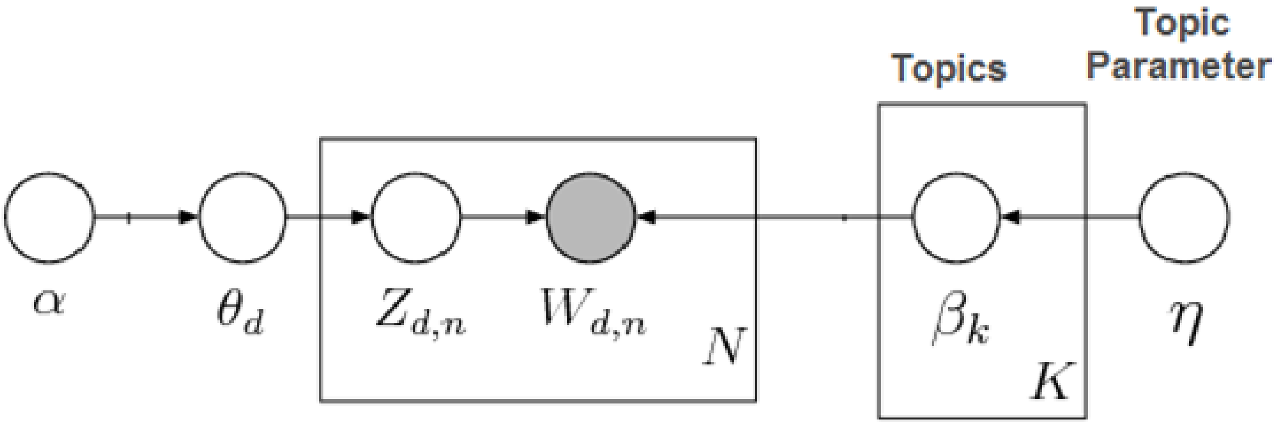
	Topic 1	Topic 2	Topic 3
Doc 1	0.87	0.09	0.04

Topic 1

Computer: 0.035
Program: 0.025
File: 0.003
...

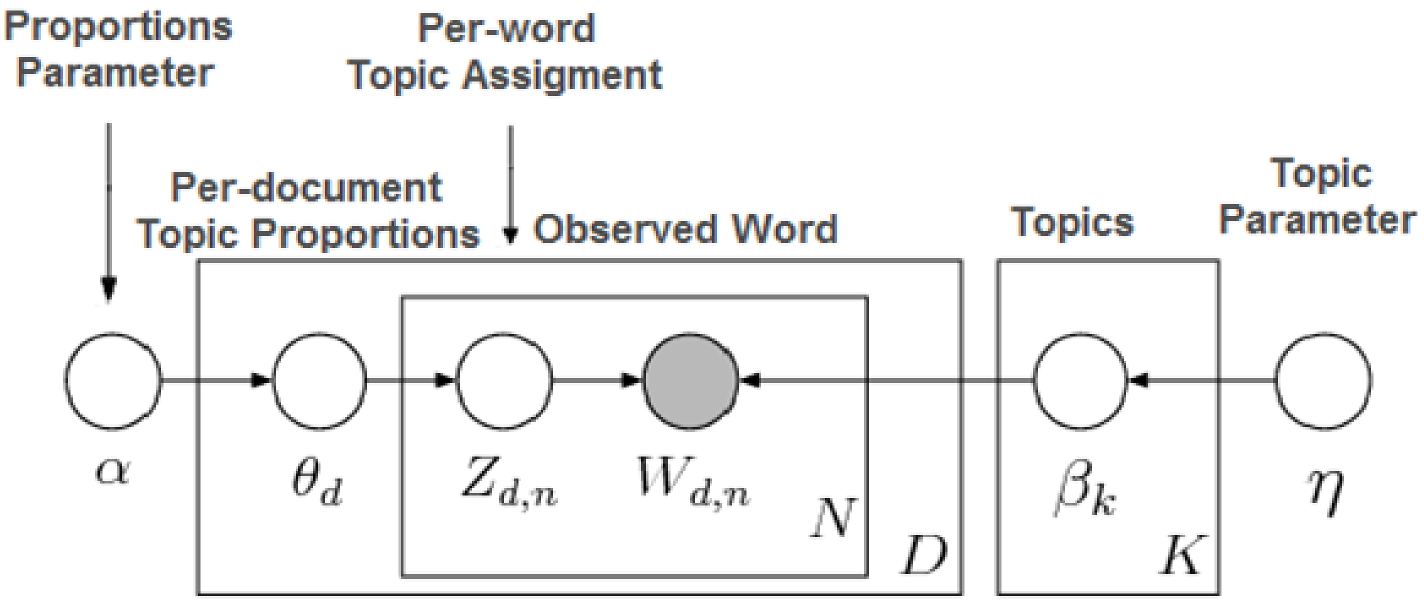
program

LDA - Corpus Generation Process



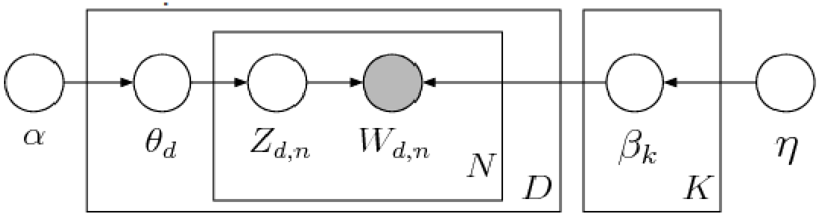
Repeat N Times → **Doc 1:** program file earth command...

LDA - Corpus Generation Process



Repeat D times ⟶ (Doc1, Doc2, Doc3….)

LDA - Inference Process



What we want/hidden : $\beta_{1:K}, \theta_{1:D}, z_{1:D}$

What we know/observed: $w_{1:D}$

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

$$\longrightarrow$$

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

$\beta_{1:K}, \theta_{1:D}, z_{1:D}$ **could not be directly computed**

LDA – Gibbs Sampling

- Initiate $\beta_{1:k}$, $\theta_{1:D}$, $z_{1:D}$ with random values

- For a specific word w in a document d:

  We assume all other assignments are correct except the current word

  $\beta_{1:k} - P(w|z_1), P(w|z_2), P(w|z_3) \ldots P(w|z_z)$
  θ_d _ $P(z_1|d), P(z_2|d), P(z_3|d) \ldots P(zz|d)$

  Mul$iply\ each\ pair\ of\ P(w|z_i)\ P(zi|d)$ ⟶ Maximum: Most suitable z_i for w in d
  Assign word w with a new topic z

- Do these steps for all words
- Repeat -- Values become stable

Implementation

a. Corpus represented in VSM
b. Dirichlet parameters

→ **LDA**
Inference →

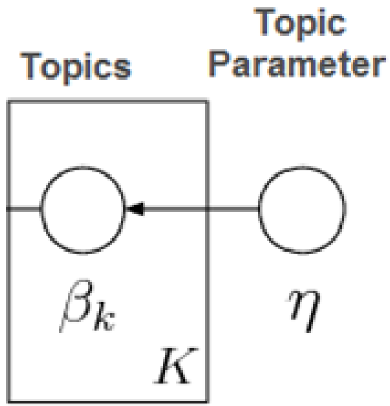
	Topic 1	Topic 2	Topic 3
computer	0.035	8.3e-5	2.3e-5
earth	3.5e-5	0.033	3.4e-7
file	0.003	3.2e-8	1.0e-8
organic	2.7e-4	0.010	9.2e-5
planet	1.9e-9	1.3e-7	0.012
...	...	...	...

	Topic 1	Topic 2	Topic 3
Doc 1	0.87	0.09	0.04
Doc 2	0.05	0.92	0.03
...	...	...	...

Research Questions

In the implementation of LDA, there are several uncertain variables.

1. K? How would the number of topics influence topic distance and coherence?



Research Questions

2. What corpus type should we use to improve topic distance and topic coherence?

- Binary – Whether a word exists in the doc or not (1:0, 2:1, 3:1, 4:0…z:0)
- Bow – How many times a word appears in a doc (1:0, 2:4, 3:10, 4:0…z:0)
- Tfidf – Whether a word is specific/important to a doc or not (1:0, 2:0.33, 3:0.52,4:0…z:0)
  - $\text{frequency(w,d)} \times \log \dfrac{Total\ number\ of\ docs}{Number\ of\ docs\ where\ w\ appears}$

Research Questions

3. To present a topic, how many words should be used to maximize coherence?

• Topic x - Program - Algorithm - Command	• Topic x - Program - Algorithm - Command - Computation	• Topic x - Program - Algorithm - Command - Computation - Software - Human - Graphics	• Topic x - Program - Algorithm - Command - Computation - Software - Human - Graphics - Mark - Template - Interface

Topic Representation

- **A vector on the vocabulary space**

  i.e. (0.034, 0.031, 0.029, 0.023…)

- **A distribution over words**

  i.e. (school:0.034, student: 0.031, assignment:0.029, education:0.023…)

- **A ranked list of words**

  i.e. (school, student, assignment, education) ranked by distribution values

- **A set of top words**

  i.e. (student, education, assignment, school)

……

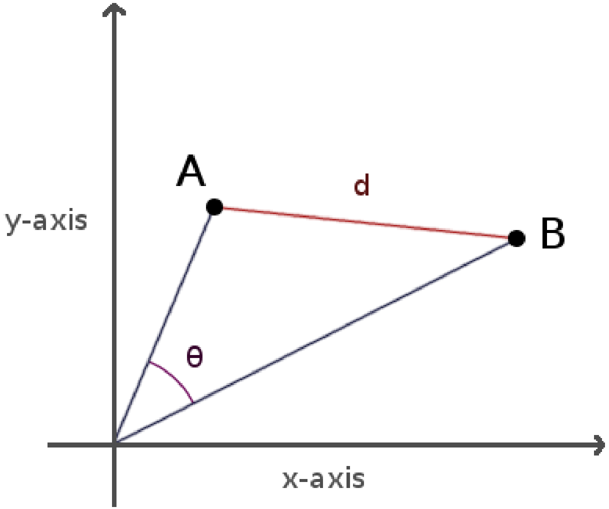
Similarity/Distance Measures

- Vector Representation: Cosine Distance
- Distribution Representation:
  - Bha.(Bhattacharyya) Distance
  - KL(Kullback–Leibler) Divergence
- Ranking Representation: Kendall’s Tau
- Set Representation: Jaccard Distance

Cosine Similarity and Distance

Calculate similarity between two vectors.
For vector A and vector B:

$$\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{|A||B|}$$



$$\text{Distance} = 1\text{- Similarity}$$

Distance Range: 0 (θ=0) – 2 (θ=180)

Bha.(Bhattacharyya) distance

Calculate distance between two distribution p and q over X
Range: 0 (same distribution) - ∞

$$D_B(p,q) = -\ln(BC(p,q))$$

Cosine Similarity between
$(\sqrt{p(x_1)}, \sqrt{p(x_2)}, \sqrt{p(x_3)} \ldots \sqrt{p(xn)})$
and
$(\sqrt{q(x_1)}, \sqrt{q(x_2)}, \sqrt{q(x_3)} \ldots \sqrt{q(xn)})$

where:

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

KL Divergence

Compute similarity between two distribution P and Q over X(x_1, x_2…x_n)
Range: 0(same distribution) - ∞

$$\mathrm{D}_{\mathrm{kL}}(\mathrm{P}|\mathrm{Q}) = \sum_{i=1}^{n} P(x_i) log \frac{P(xi)}{Q(xi)}$$

$$\mathrm{D}_{\mathrm{KL}}\ \mathrm{Symmetric} = \frac{\mathrm{D}_{\mathrm{kL}}(\mathrm{P}|\mathrm{Q}) + \mathrm{Dkl}(\mathrm{Q}|\mathrm{P})}{2}$$

Kendall'Tau Rank Correlation

For ranked list X ($x_1, x_2, x_3 \ldots x_n$) and ranked list Y ($y_1, y_2, y_3 \ldots y_n$) with the same size

Pair (x_i, y_i). – ((x_1, y_1), (x_2, y_2)…(x_n, y_n))

For any two pairs of (x,y): (x_i, y_i) & (x_j, y_j)

- Concordant Pairs: $x_i > x_j$ && $y_i > y_j$ | $x_i < x_j$ && $y_i < y_j$
- Discordant Pairs: $x_i > x_j$ && $y_i < y_j$ | $x_i < x_j$ && $y_i > y_j$

Total Pairs: $\frac{n(n-1)}{2}$

i.e. Two ranked lists P(1,3,2), Q(3,2,1) -- Pair P and Q: (1,3), (3,2), (2,1)

Concordant Pairs: (3,2) & (2,1) Discordant Pairs: (1,3) & (3,2), (1,3) & (2,1)

Kendall's Tau Rank Correlation

Kendall's Tau coefficient:

Range: -1(negative correlation) - 1(positive correlation)

$$\mathrm{T} = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{n(n-1)/2}$$

Jaccard Similarity and Distance

Compute similarity between two sets of elements

We choose top 500 words for each topic.

Distance Range: 0(same) – 1(no common elements)

$$\text{Similarity(A, B)} = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{Distance} = 1 - \text{Similarity}$$

Topic Coherence Measures

Represent a topic as a set of words with top distributions

- Co-occurrence based topic coherence measure
- WordNet topic coherence measure

Co-occurrence Based Coherence Measure

For any two words v_l and v_m in the topic:

Document Frequency $D(v_l)$: the number of documents that contain v_l

Co-occurrence Document Frequency $D(v_l, v_m)$: the number of documents that contain both v_l and v_m

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

Tfidf Co-occurrence Based Coherence Measure

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{\sum_{d: w_1, w_2 \in d} \text{tf-idf}(w_1, d) \text{tf-idf}(w_2, d) + \epsilon}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)},$$

where tf-idf is computed with augmented frequency,

$$\text{tf-idf}(w, d) = \text{tf}(w, d) \times \text{idf}(w) = \left(\frac{1}{2} + \frac{f(w, d)}{\max_{w' \in d} f(w', d)} \right) \log \frac{|D|}{|\{d \in D : w \in d\}|},$$

and $f(w, d)$ is how many times term w occurs in document d.

WordNet Coherence Measure

A collection of English words. - A graph of synsets

A synset represents a specific semantical concept
Synset _____many to many_____Word

Engineer:

Noun

- S: (n) **engineer**, applied scientist, technologist (a person who uses scientific knowledge to solve practical problems)
- S: (n) **engineer**, locomotive engineer, railroad engineer, engine driver (the operator of a railway locomotive)

Verb

- S: (v) **engineer** (design as an engineer) *"He engineered the water supply project"*
- S: (v) mastermind, **engineer**, direct, organize, organise, orchestrate (plan and direct (a complex undertaking)) *"he masterminded the robbery"*

WordNet Coherence Measure

Synsets are connected by different relationships:

- Hierarchical:

a. Hypernyms: carnivore is a hypernym of dog

b. Hyponyms: dog is a hyponym of carnivore

- Horizontal:

a. Antonym: able, unable

b. Pertainym: academic, academia

c. There are many other horizontal relationships

WordNet Coherence Measure

There are diverse methods measuring relevance of two synsets.

Path: $\mathrm{sim}(s_1,s_2) = \frac{1}{sp(s_1,s_2)}$

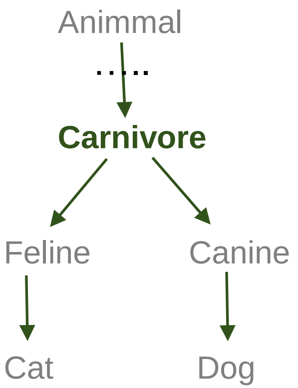
LCH: $\mathrm{sim}(s_1,s_2) = -\log\left(\frac{\mathrm{sp}\ (\mathrm{s1},\ \mathrm{s2})}{2D}\right)$

D: the maximum depth of WordNet

WuP: $\mathrm{sim}(s_1,\ s_2) = \frac{2\mathrm{depth}(\mathrm{lcs}\ _{c1,c2)})}{\mathrm{depth}\ c_1 + \mathrm{depth}\ c_2 + 2*\mathrm{depth}(\mathrm{lcs}\ _{c1,\ c2)}}$

LCS (The least common subsumer):

The most specific node in the hierarchy that subsumes both synsets



WordNet Coherence Measure

Use information about words in a corpus

Information Content $\quad \mathrm{IC}(s) = -\log p(s)$

RES: $\mathrm{sim}\,(s_1, s_2) = \max_{c \in S(s_1, s_2)} [-\log p(s)]$

LIN: $\mathrm{sim}(s_1, s_2) = \dfrac{2 \times \log p(\mathrm{lcs}_{s1,s2})}{(\log p(s_1) + \log p(s_2))}$

JCN: $\mathrm{sim}(s_1, s_2) = \dfrac{1}{\mathrm{IC}(s_1) + \mathrm{IC}(s_2) - 2 \times \mathrm{IC}(\mathrm{lcs}_{s1,s2})}$

WordNet Coherence Measure

- Topic T - (w_1, w_2, w_3....w_z)

For a pair of words (w_1, w_2) in the topic :
w_1 belongs to N synsets – w_{1}_s_1, w_{1}_s_2....w_{1}_s_n
w_2 belongs to M synsets – w_{2}_s_1, w_{2}_$s_{2...}$$w_{2}$_$s_m$

Similarity/Relevance between two words w_1, w_2

- $\mathrm{sim}(w_1, w_2) = \max(\sum_{j=1}^{M} \sum_{i=1}^{N} \mathrm{sim}(w_{1}\_sj, w_{2}\_si))$

- Coherence of a Topic
  - mean $(\sum_{m=2}^{Z} \sum_{l=1}^{m-1} \mathrm{sim}(wm, wl))$
  - median$(\sum_{m=2}^{Z} \sum_{l=1}^{m-1} \mathrm{sim}(wm, wl))$

Outline

- Background
- Distance Experiments
- How the number of topics influence distance
- How corpus types influence distance
- Correlations among results from different measures
- Coherence Experiments
- Discussion

Packages

- NLTK
- Natural Language Toolkit
- Natural Language Processing Functions
- Python

- Gensim
- Topic Modeling Algorithms
- Python

Materials

Two corpora

Reuters:

- News collection
- 90 categories: jobs, housing, coffee, gas, wheat…
- 10788 Files
- Vocabulary size: 26518

Brown:

- Literature collection
- 16 categories: romance, humor, adventure, ...
- 500 Files
- Vocabulary Size: 36450

Preprocessing

- Tokenization
- Tagging   w1/Noun w2/Verb …
- Lemmatization - WordNetLematizer
- Words Can't be lemmatized - > Stemming   removing ing, es, s, ed…
- RE Matching
  a. Tokens with letters,
  b. Numbers/Hyphens embedded in tokens   i.e. H2O, two-year
  c. Abbreviations   i.e. U.S.A
- Removing Stop Words   i.e. is then when

Distance Experiment

Corpus Name	Reuters	Brown
Topic Numbers:	10 20 30 40 50	12 14 16 18 20
Corpus Type	tfidf, bow, binary	
Measures	- Cosine Distance - Bhattacharyya Distance - KL Divergence - Jaccard Distance - Kendall’s Tau Correlation	

Distance Experiment

Topic Number & Corpus Type:

For each set of topics (t0, t1, t2…tn):

Distance = $\text{mean}(\text{distance}(t0,t1) + \text{distance}(t0, t2) \ldots \text{distance}(tn-1, tn))$

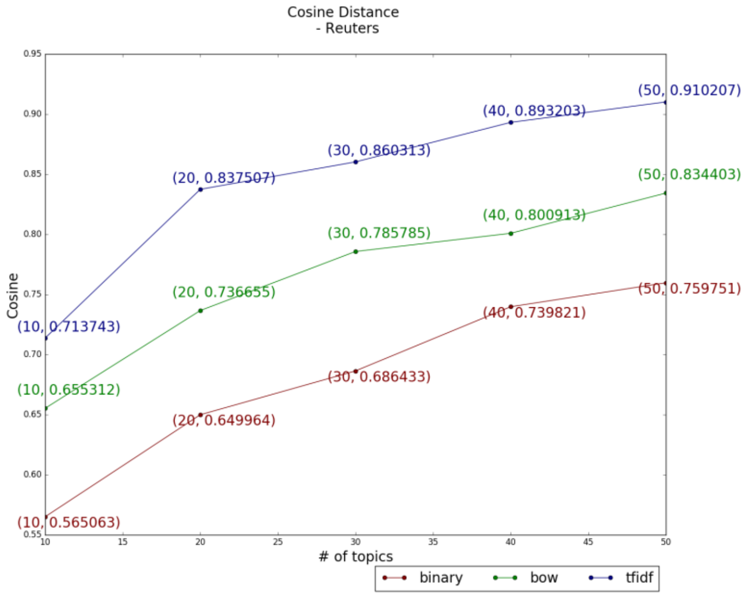
Measures Correlation

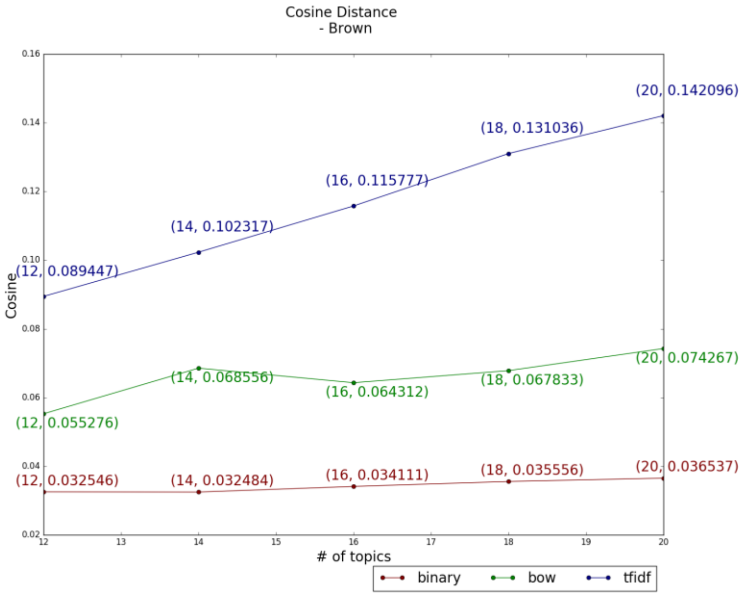
- Pearson Correlation
- Kendall’s Tau Ranking Correlation

Distance Results

Topic as a vector over the vocabulary space – Cosine Distance

Distance	Reuters			Brown		
# of topics increases	Increase	Increase	**Increase**	Steady	Almost Steady	**Increase**

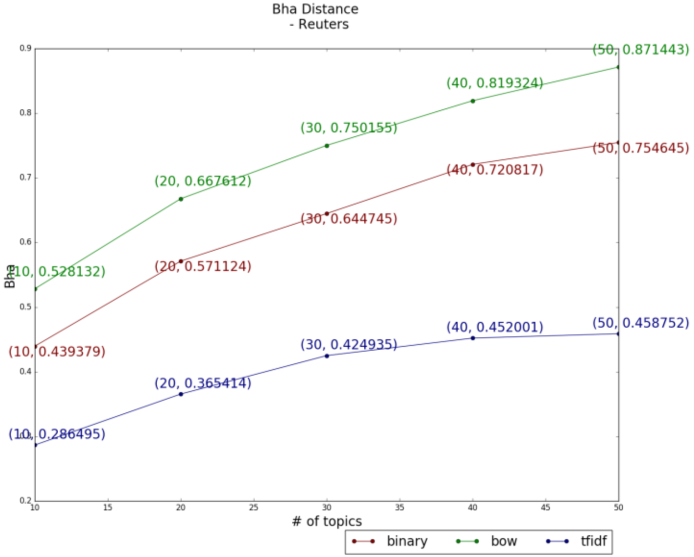


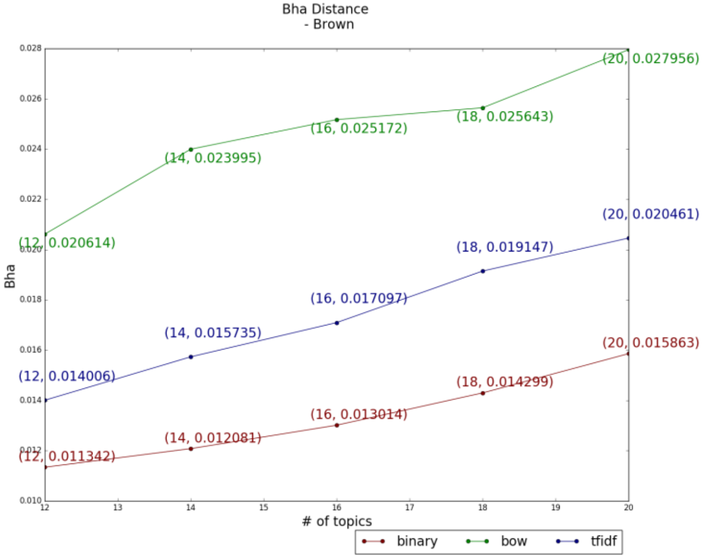


Distance Results

Topic as a distribution over words – Bha. Distance

Distance	Reuters			Brown		
# of topics increases	Increase	Increase	Increase	Increase	Increase	Increase

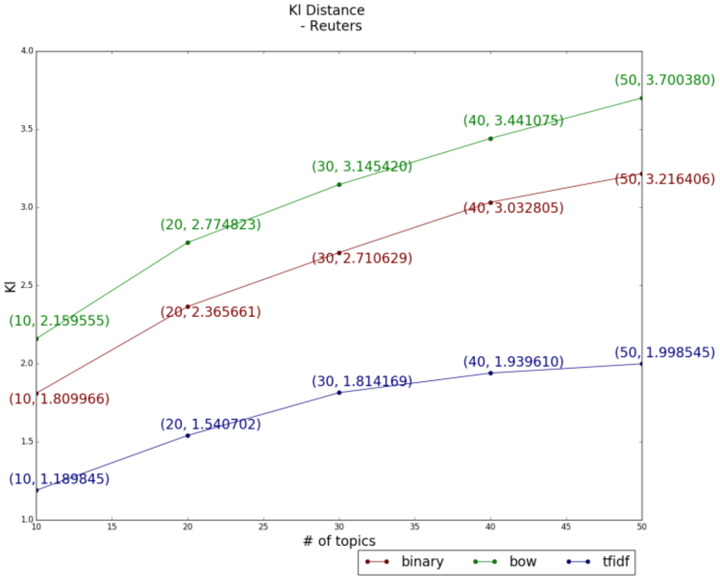


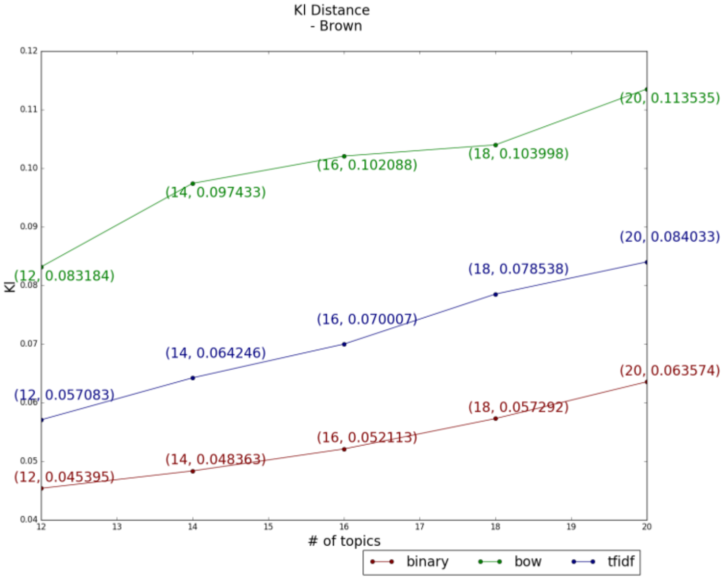


Distance Results

Topic as a distribution over words – KL Divergence

Distance	Reuters			Brown		
# of topics increases	Increase	Increase	Increase	Increase	Increase	Increase

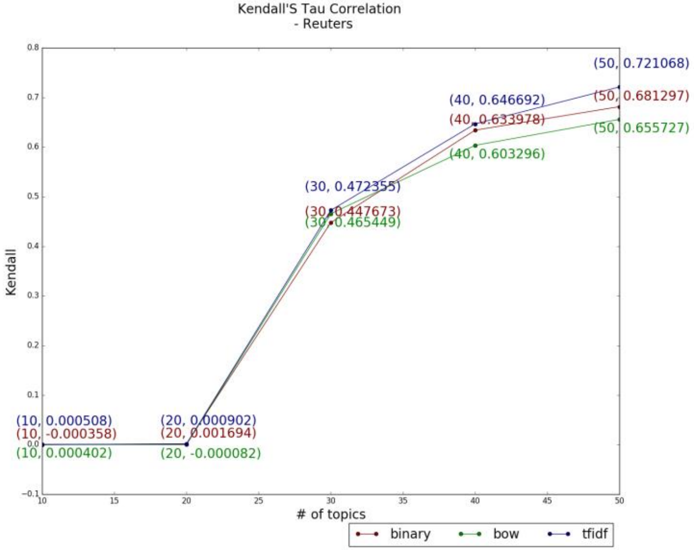


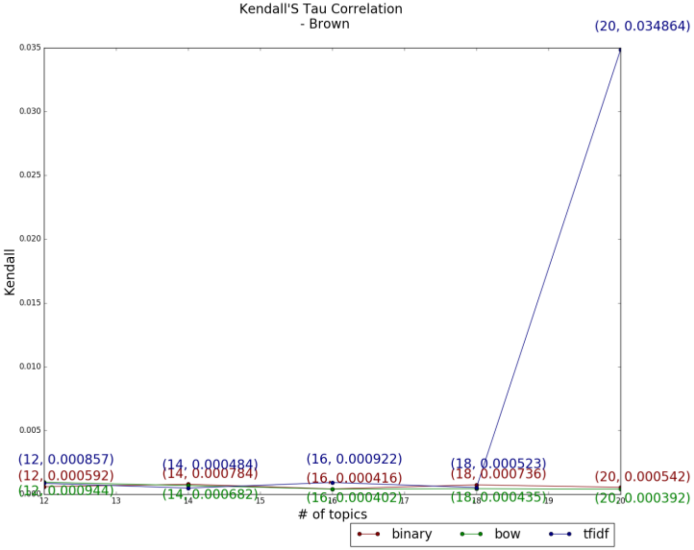


Distance Results

Topic as a ranked list of words – Kendall's Tau Correlation (-1 to 1)

Similarity	Reuters			Brown		
# of topics increases	Increase	Increase	Increase	Steady	Steady	Increase

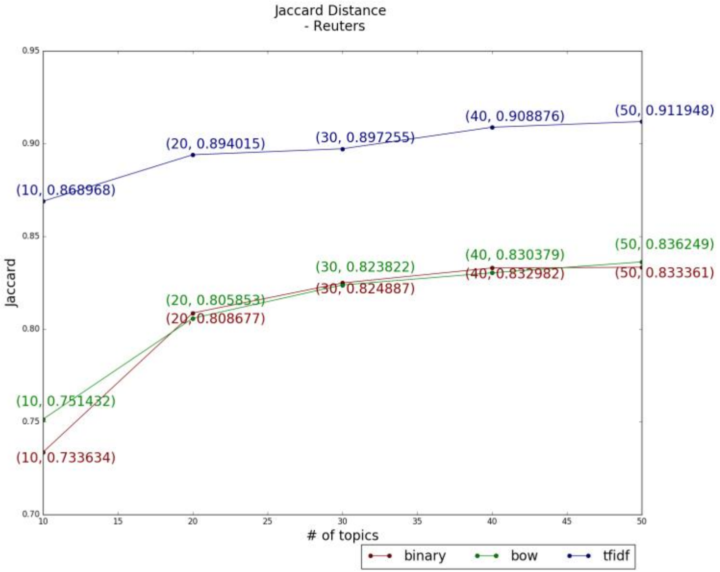


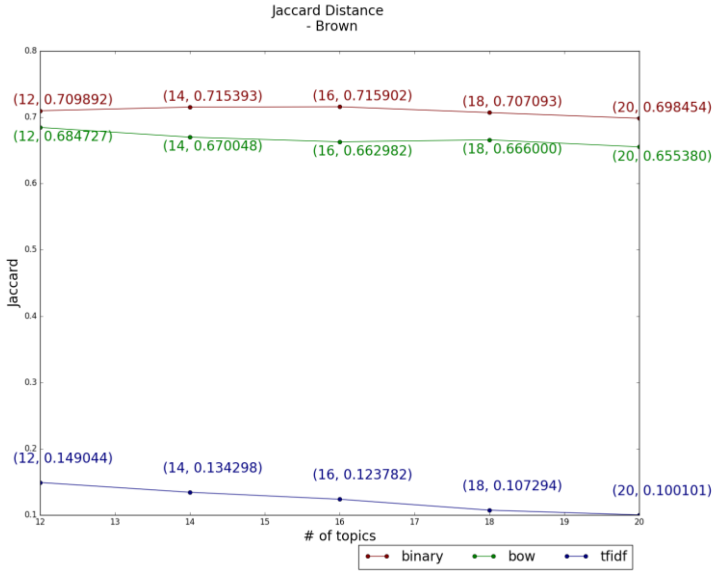


Distance Results

Topic as a set of words – Jaccard Distance

Distance	Reuters			Brown		
# of topics increases	Increase	Increase	**Increase**	**Steady**	Steady	Decrease





Distance Results- Correlations among measures

Pearson correlation(all $p<0.05$) - Reuters

	Cosine	**Bha**	**KL**	**Jaccard**	**Kendall**
Cosine		0.167	0.131	**0.760**	0.331
Bha	0.167		**0.990**	-0.070	0.127
KL	0.131	**0.990**		-0.123	0.171
Jaccard	**0.760**	-0.070	-0.123		0.304
Kendall	0.331	0.127	0.171	0.304	

Similarity Results - Correlations among measures

Kendall's Tau Ranking Correlation(all p<0.05) - Reuters

	Cosine	**Bha**	**KL**	**Jaccard**	**Kendall**
Cosine		-0.128	-0.139	0.258	0.134
Bha	-0.128		-0.191	-0.121	-0.056
KL	-0.139	-0.191		-0.117	-0.044
Jaccard	0.258	-0.121	-0.117		0.141
Kendall	0.134	-0.056	-0.044	0.141	

Similarity Results- Correlations among measures

Pearson correlation - Brown

	Cosine	Bha	KL	Jaccard	Kendall
Cosine		**0.374	**0.397	**** -0.875**	**0.188
Bha	**0.374		****1.000**	**0.046	0.038
KL	**0.397	****1.000**		0.022	0.042
Jaccard	**** -0.875**	**0.046	0.022		** -0.152
Kendall	**0.188	0.038	0.042	** -0.152	

Note:
** : $p < 0.05$.
Bold : strong correlation

Similarity Results - Correlations among measures

Kendall’s Tau Ranking Correlation - Brown

	Cosine	Bha	KL	Jaccard	Kendall
Cosine		**0.325	**0.328	**-0.658	0.019
Bha	**0.325		**0.349	**-0.333	**-0.121
KL	**0.328	**0.349		**-0.341	**-0.127
Jaccard	**-0.658	**-0.333	**-0.341		-0.005
Kendall	0.019	**-0.121	**-0.127	-0.005	

Note:
** : $p < 0.05$.
Bold : strong correlation

Outline

- Background
- Distance Experiments
- Coherence Experiments
  - How number of top words influence coherence
  - How number of topics influence coherence
  - How corpus types influence coherence
  - Correlations among different measures
- Discussion

Coherence Experiment

Corpora	Reuters, Brown
Numbers of Topics:	5,10,15,20
Numbers of Top Words	5, 10, 15, … 150
Corpus Type:	Tfidf, Bow, Binary
Measures:	- Co-occurrence based coherence Measure - Tfidf Co-occurrence based coherence Measure - WordNet coherence measures

Coherence Experiment

Number of Top Words:

For each set of topics (t1, t2…tn):

Topic-set Coherence Value = $\frac{\mathrm{coh(t1)} + \mathrm{coh(t2)} \ldots \mathrm{coh(tn)}}{n}$

Baseline – Random words

Number of Topics & Corpus Type:

For each set of topics (t1, t2…tn):

Topic-set Coherence Value = $\frac{\mathrm{coh(t1)} + \mathrm{coh(t2)} \ldots \mathrm{coh(tn)}}{n}$

Measures Correlation

- Pearson Correlation
- Kendall’s Tau Ranking Correlation

Coherence Experiment

Coherent Topic

TC: -70.089

bank : 0.0279657465152

rate : 0.023811830955

pct : 0.0156586442573

dollar : 0.0134510791657

market : 0.0132709087771

currency : 0.00921212587181

u.s. : 0.00859911745669

exchange : 0.00802242679236

cut : 0.00696400103261

mark : 0.00672299580672

Incoherence Topic

TC: -149.210028238

three-for-two : 0.00510543190802

pct : 0.00441486771091

rise : 0.00413571122315

february : 0.00403229866795

january : 0.00330294977387

writedown : 0.00293321731826

africa : 0.00290027441263

mthly : 0.00289415081164

rand : 0.00283650905673

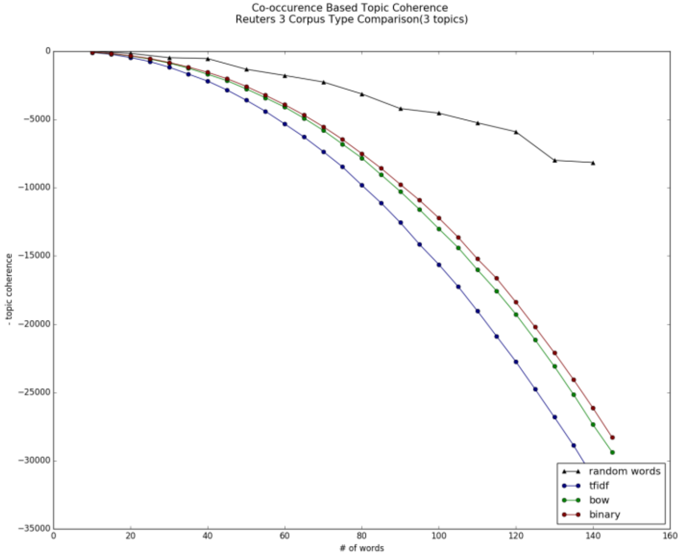
year-on-year : 0.00256394193525

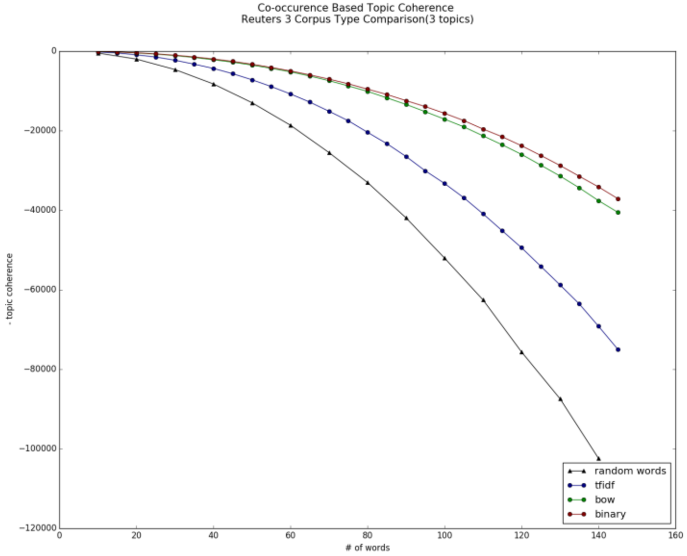
Outline

- Background
- Distance Experiments
- Coherence Experiments
  - How number of top words influence coherence
  - How number of topics influence coherence
  - How corpus types influence coherence
  - Correlations among different measures
- Discussion

Coherence Results – Co-occurence

The number of top words does not influence coherence





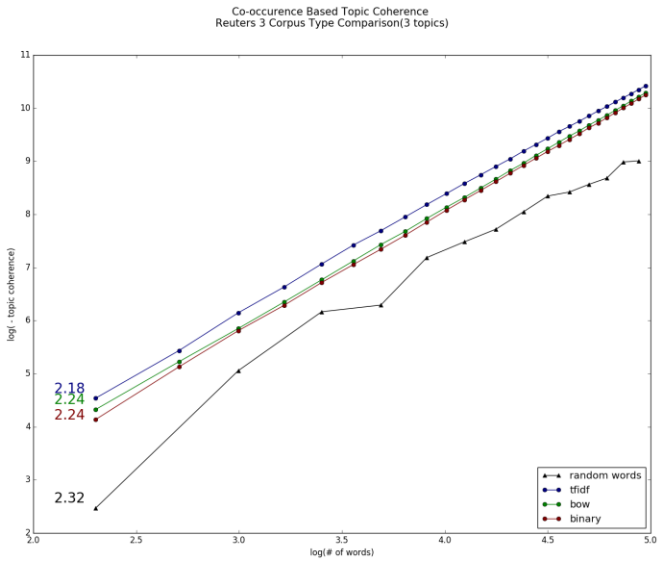
$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

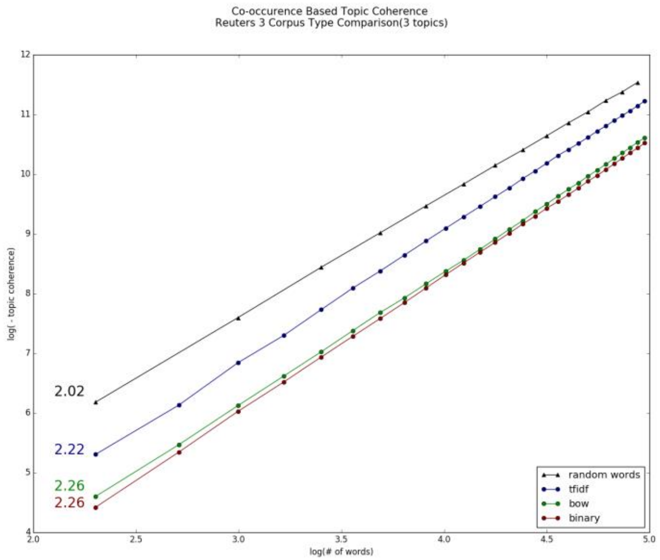
Word pair co-occurrence contribution = $\log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$

Most time this is negative

Coherence Results – Co-occurence

Coherence value gradually becomes negative



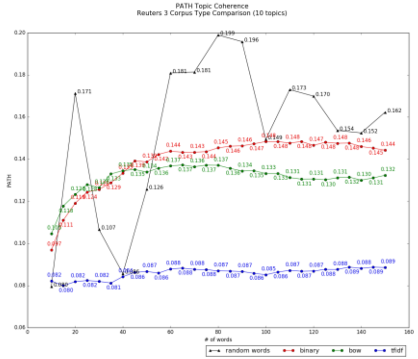


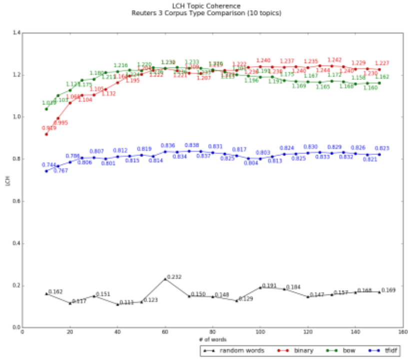
$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D(v_m^{(t)},v_l^{(t)})+1}{D(v_l^{(t)})}$$

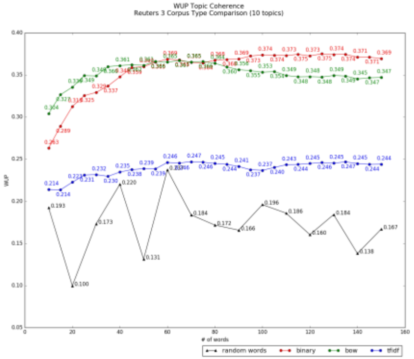
Word pair co-occurrence contribution = $\log\frac{D(v_m^{(t)},v_l^{(t)})+1}{D(v_l^{(t)})}$

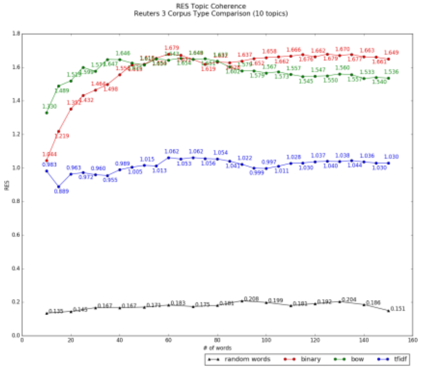
Coherence Results – WordNet (mean)

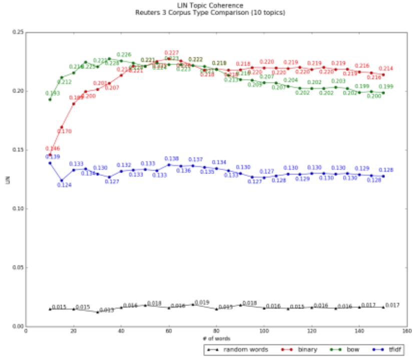
Binary & Bow Corpus: After 40 - 60 top words – steady

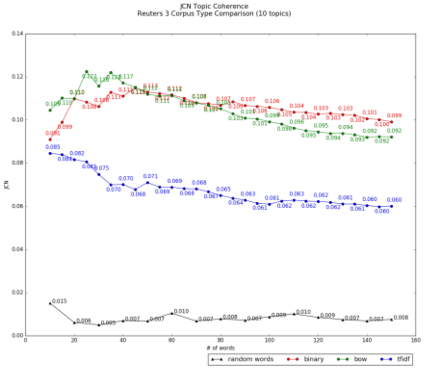






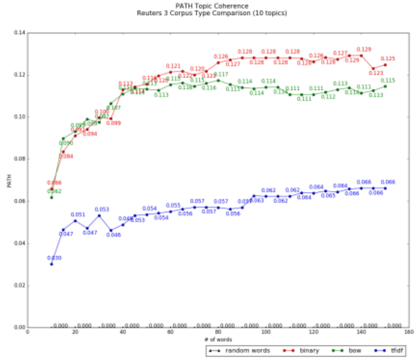


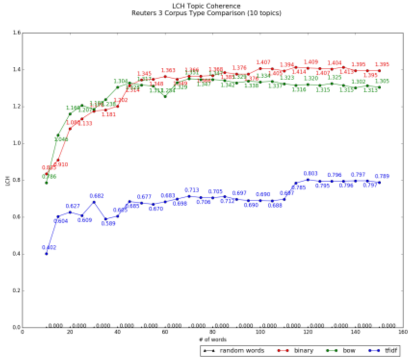


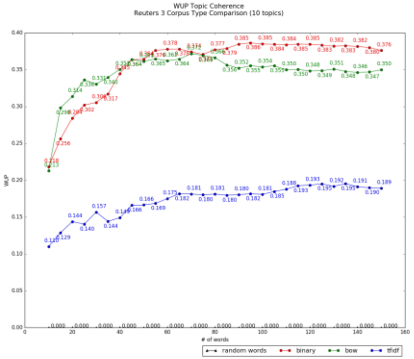


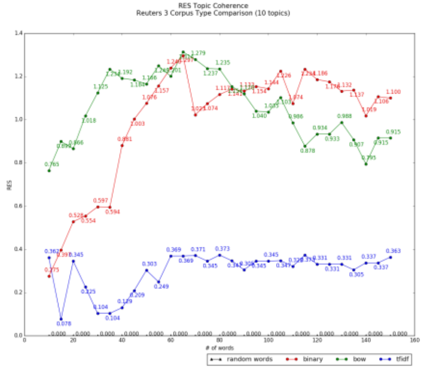
Coherence Results – WordNet (median)

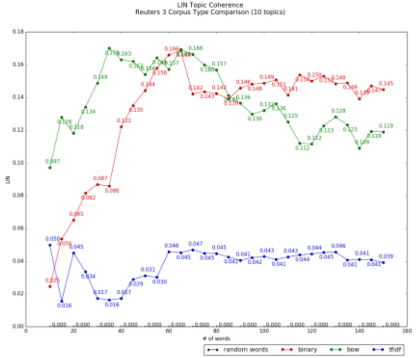
Binary & Bow Corpus: 40 - 60 top words have the best coherence

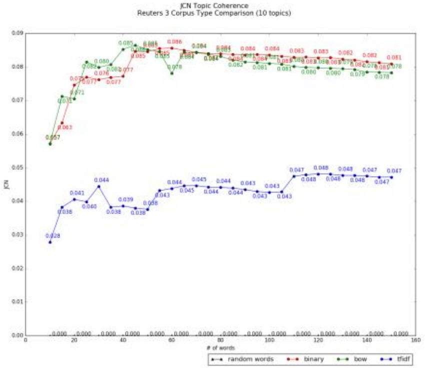










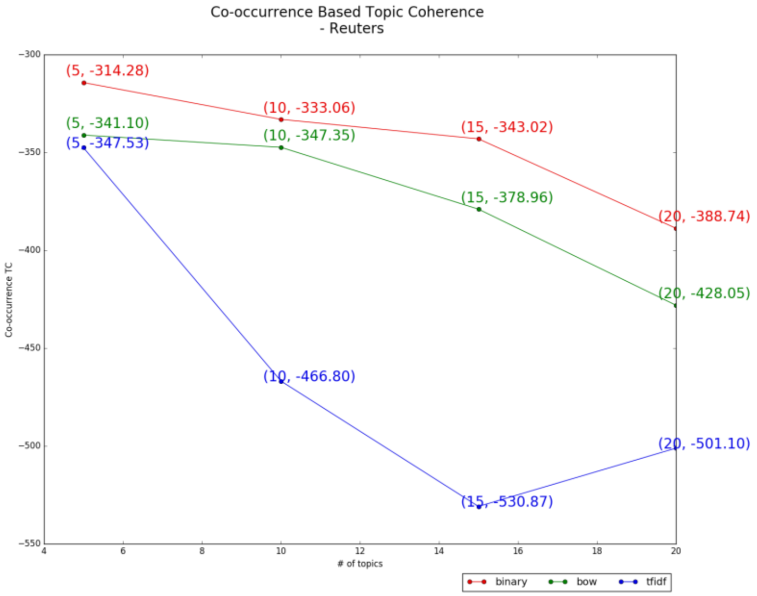


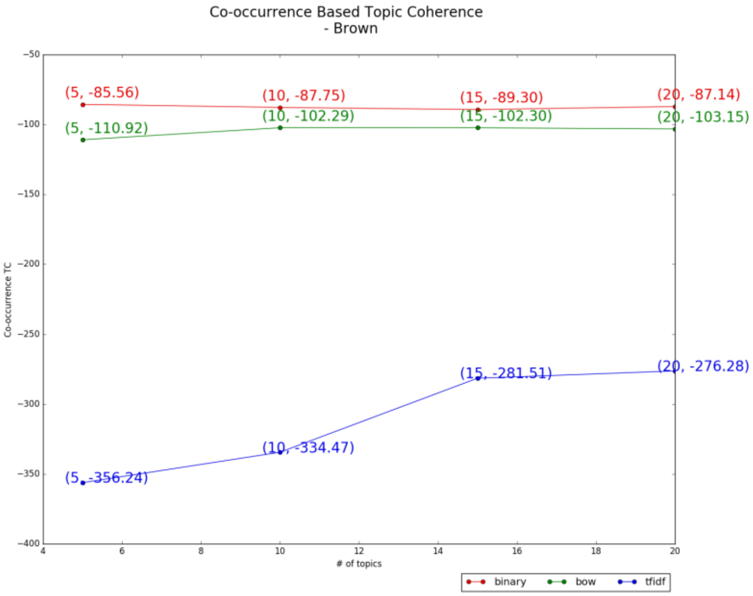
Outline

- Background
- Distance Experiments
- Coherence Experiments
  - How number of top words influence coherence
  - How number of topics influence coherence
  - How corpus types influence coherence
  - Correlations among different measures
- Discussion

Coherence Results - Co-occurrence

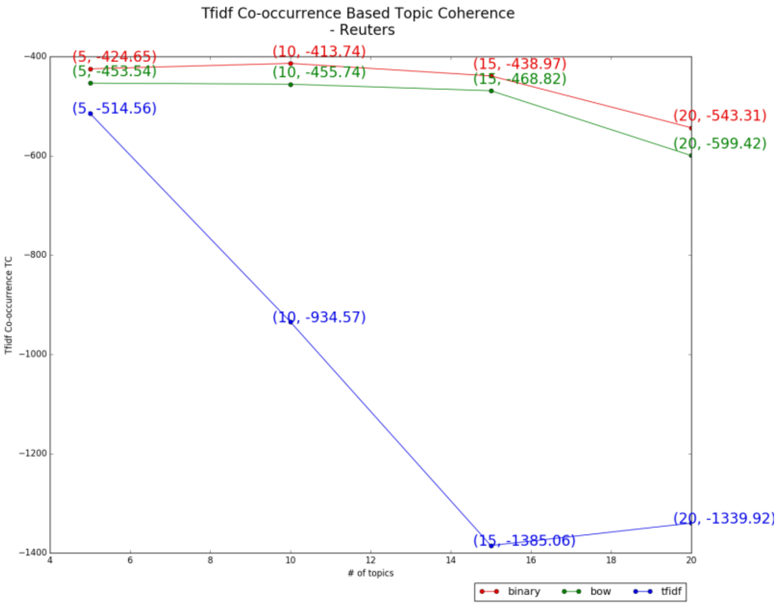
	Reuters			Brown		
# of topics increases	Decrease	Decrease	Decrease then Increase	Steady	Steady	Increase

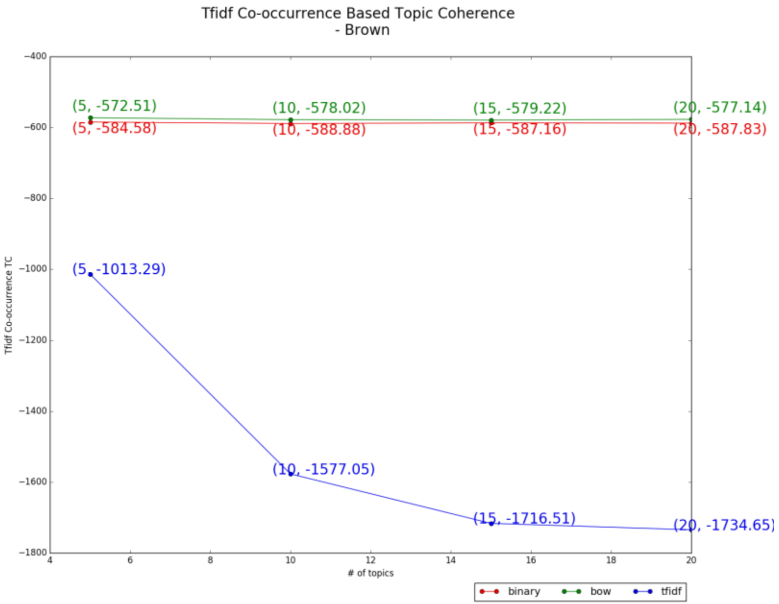




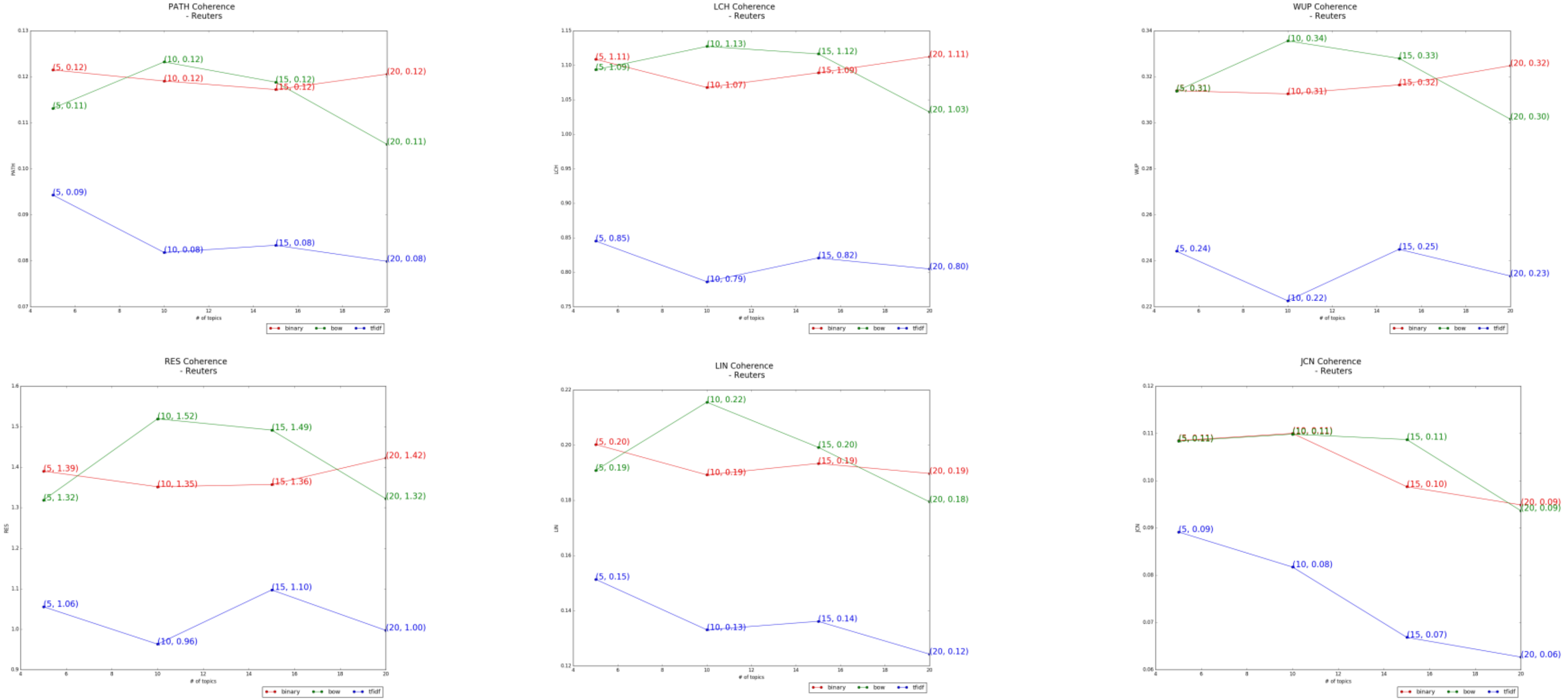
Coherence Results – Tfidf Co-occurrence

	Reuters			Brown		
# of topics increases	Decrease	Decrease	Decrease then Increase	Steady	Steady	Decrease

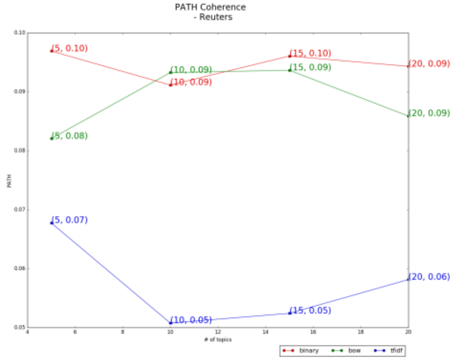


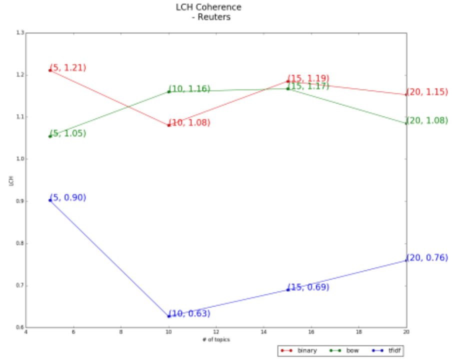


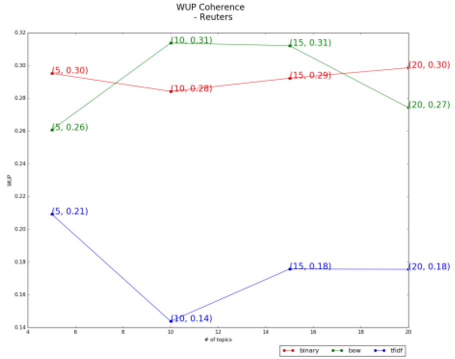
Coherence Results – WordNet (Reuters Mean)

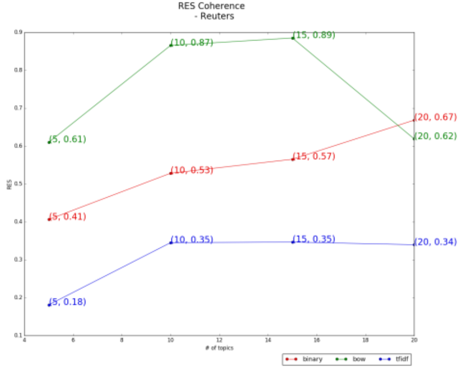


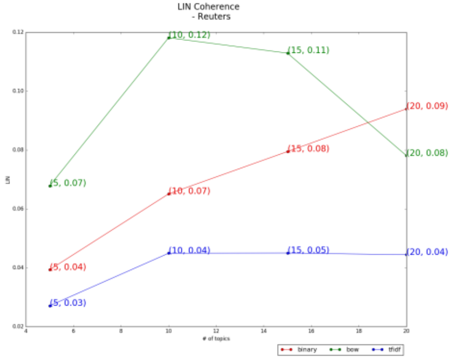
Coherence Results – WordNet (Reuters Median)

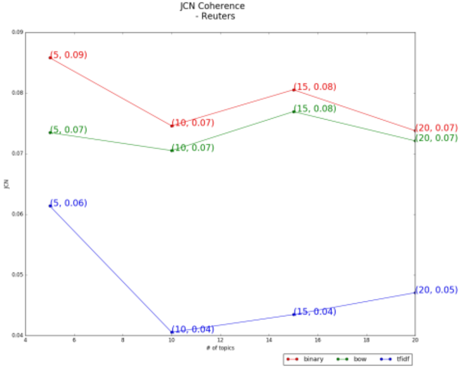






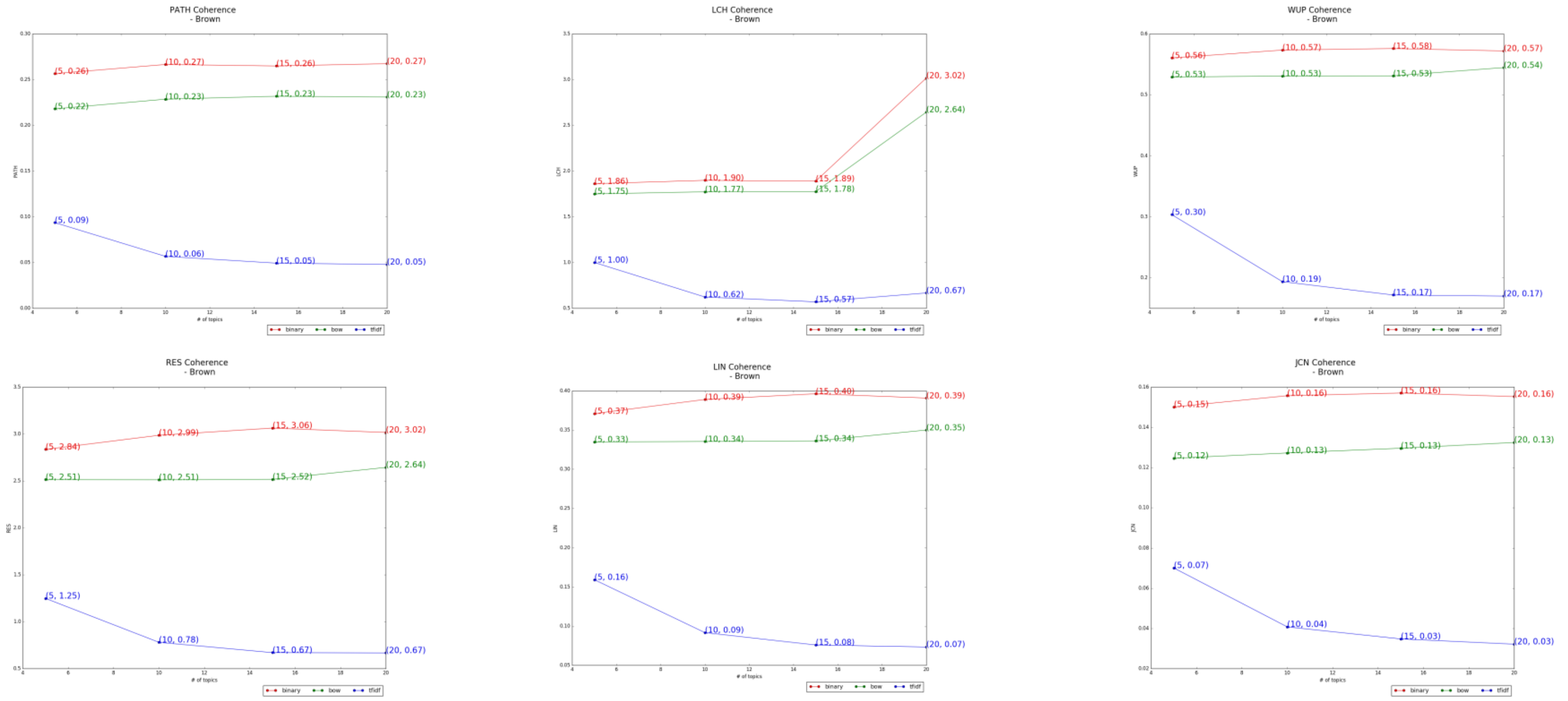






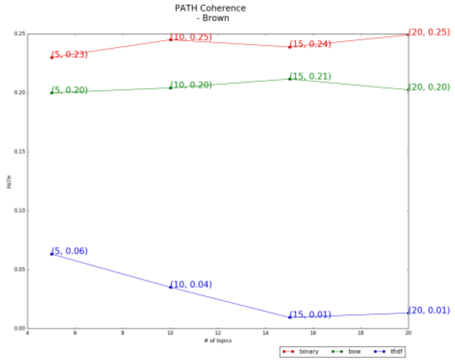
Coherence Results – WordNet (Brown Mean)

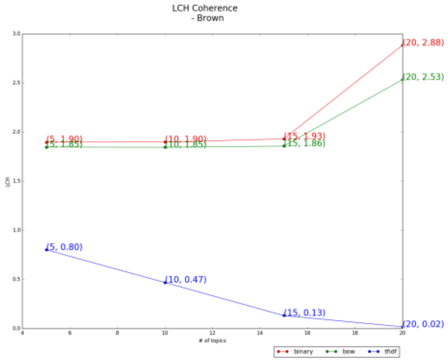
For tfidf, as topic number increases, coherence decreases

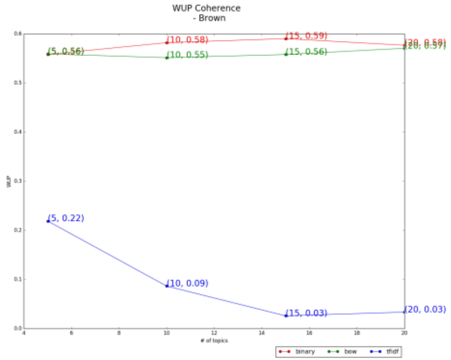


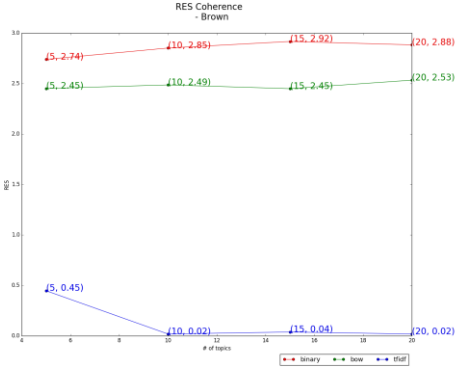
Coherence Results – WordNet (Brown Median)

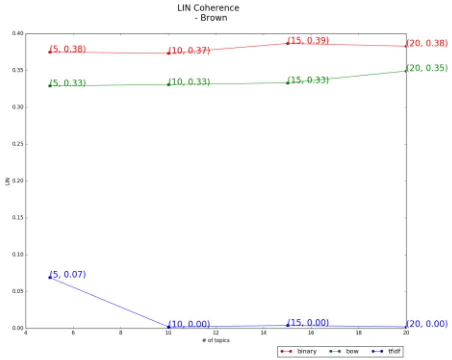
For tfidf, as topic number increases, coherence decreases

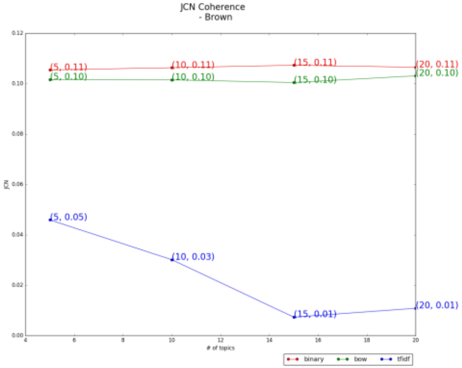












Outline

- Background
- Distance Experiments
- Coherence Experiments
  - How number of top words influence coherence
  - How number of topics influence coherence
  - How corpus type influence coherence
  - Correlations among different measures
- Discussion

Coherence Results- Correlation among Measures

Pearson Correlation

	TC	Tfidf-TC	PATH	LCH	WUP	RES	LIN	JCN
TC		**0.381	0.000	0.081	0.006	0.003	-0.006	-0.003
Tfidf-TC	**0.381		0.026	0.0625	0.047	0.018	0.019	0.023
PATH	0.000	0.026		****0.910**	****0.978**	****0.975**	****0.978**	****0.879**
LCH	0.081	0.0625	****0.910**		****0.912**	****0.913**	****0.907**	****0.795**
WUP	0.006	0.047	****0.978**	****0.912**		****0.982**	****0.983**	****0.885**
RES	0.003	0.018	****0.975**	****0.913**	****0.982**		****0.994**	****0.887**
LIN	-0.006	0.019	****0.978**	****0.907**	****0.983**	****0.994**		****0.914**
JCN	-0.003	0.023	****0.879**	****0.795**	****0.885**	****0.887**	****0.914**	

Note:
** : $p < 0.05$.
Bold : strong correlation

Coherence Results- Correlation among Measures

Kendall's Tau Correlation (all $p < 0.05$)

	TC	**Tfidf-TC**	**PATH**	**LCH**	**WUP**	**RES**	**LIN**	**JCN**
TC		0.165	0.445	0.427	0.422	0.421	0.437	0.348
Tfidf-TC	0.165		-0.327	0.314	0.320	0.288	0.310	0.159
PATH	0.445	-0.327		0.473	0.460	0.510	0.483	0.382
LCH	0.427	0.314	0.473		0.462	0.499	0.487	0.289
WUP	0.422	0.320	0.460	0.462		0.496	0.447	0.318
RES	0.421	0.288	0.510	0.499	0.496		0.480	0.337
LIN	0.437	0.310	0.483	0.487	0.447	0.480		0.289
JCN	0.348	0.159	0.382	0.289	0.318	0.337	0.289	

Outline

- Background
- Distance Experiments
- Coherence Experiments
- Discussion
  - Summarization
  - Limitations
  - Future Direction

Summarization

	Distance	**Coherence**
Corpus Type	Bow	Binary, Bow
Increase Number of Topics	Increase(Vector, Distribution) Inconsistent(Set, Ranking)	Overall: Less increase Tfidf: Decrase > Increase
Increase Number of Top Words		WordNet Binary & Bow: 40-60 words

Limitations

- 1. Topic numbers are small. Some past research tested topic numbers 100+
- 2. The range of topic numbers is small.
- 3. For WordNet coherence measures,

  - some words in the corpus are not in WordNet
  - some word pairs are not related

Pairs of words	Reuters	Brown
One word is not in WordNet	20% pairs	12% pairs
No distance	1% pairs	1% pairs

Future Directions

Overall

- Use larger number of topics and a larger range of topics

Distance

- Why would Jaccard Distance and Cosine Distance have a strong positive correlation in reuters and a strong negative correlation in brown

Future Directions

Coherence

1. Tfidf performs poorly

- Is tfidf a good choice for LDA?
- Would topic coherence measures prefer common/frequent words?
- More coherence measures

2. Instead of coherence measures, topic models have many applications

classification, relevance judgment, summarization …

- How would the number of top words, topic numbers and corpus types influence these applications?

3. For co-occurrence based coherence measure, why would random words have better coherence values?

References

- Latent Dirichlet Allocation http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- Latent Dirichlet Allocation David Blei Lectures http://videolectures.net/mlss09uk_blei_tm/
- Introduction to Latent Dirichlet Allocation http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/
- Probabilistic Topic Model https://www.cs.princeton.edu/~blei/papers/Blei2012.pdf
- Automatic Evaluation of Topic Coherence https://mimno.infosci.cornell.edu/info6150/readings/N10-1012.pdf
- Optimizing Semantic Coherence in Topic Models http://dirichlet.net/pdf/mimno11optimizing.pdf
- Topic Chains for Understanding a New Corpus http://uilab.kaist.ac.kr/research/CICLING2011/paper.pdf
- WordNet https://wordnet.princeton.edu/
- Measuring Topic Qualities in Latent Dirichlet Allocation http://logic.pdmi.ras.ru/~sergey/slides/N14_PhMLtalk.pdf
- Topic Model Image http://lca.epfl.ch/student-projects/projects/2013-09-autumn/topic_models_geotagged_tweets.html
- LDA Graphics Model Image https://filebox.ece.vt.edu/~s14ece6504/projects/alfadda_topic/index.html
- Vector Space Model Image https://en.wikipedia.org/wiki/Vector_space_model
- Cosine Similarity Image https://alexn.org/blog/2012/01/16/cosine-similarity-euclidean-distance.html

Questions?