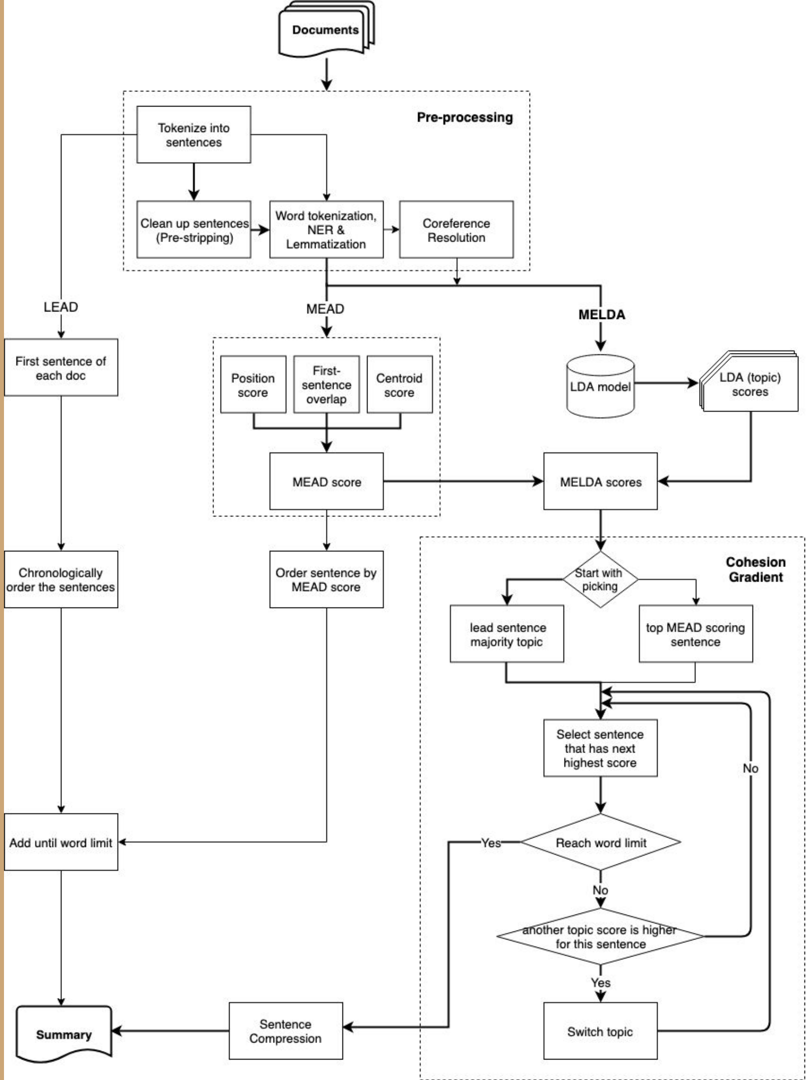
MELDA

LING 573 Text Summarization D4

Claude Zhang
Julia McAnallen
Genevieve Peaslee
Zoe Winkworth

Systems

	Lead Sentence	**MEAD System**	**MELDA System**
Content Selection	First sentence of each doc	MEAD	MELDA
Information Ordering	chronological	MEAD score	Cohesion Gradient
Content Realization	Add until next sentence exceeds 100 words	Add until next sentence exceeds 100 words	Sentence Compression


Documents
Pre-processing
Tokenize into sentences
Clean up sentences (Pre-stripping)
Word tokenization, NER & Lemmatization
Coreference Resolution
LEAD
MEAD
MELDA
First sentence of each doc
Position score
First-sentence overlap
Centroid score
MEAD score
LDA model
LDA (topic) scores
MELDA scores
Chronologically order the sentences
Order sentence by MEAD score
Cohesion Gradient
Start with picking
lead sentence majority topic
top MEAD scoring sentence
Select sentence that has next highest score
No
Add until word limit
Yes
Reach word limit
No
another topic score is higher for this sentence
Yes
Summary
Sentence Compression
Switch topic

Architecture

Improvements

- Better preprocessing (stripping)
- Tokenization/whitespace fixes
- Lots of misc. bug fixes that improved content selection and info ordering
- Info ordering options
- Sentence Compression

MELDA

Content Selection

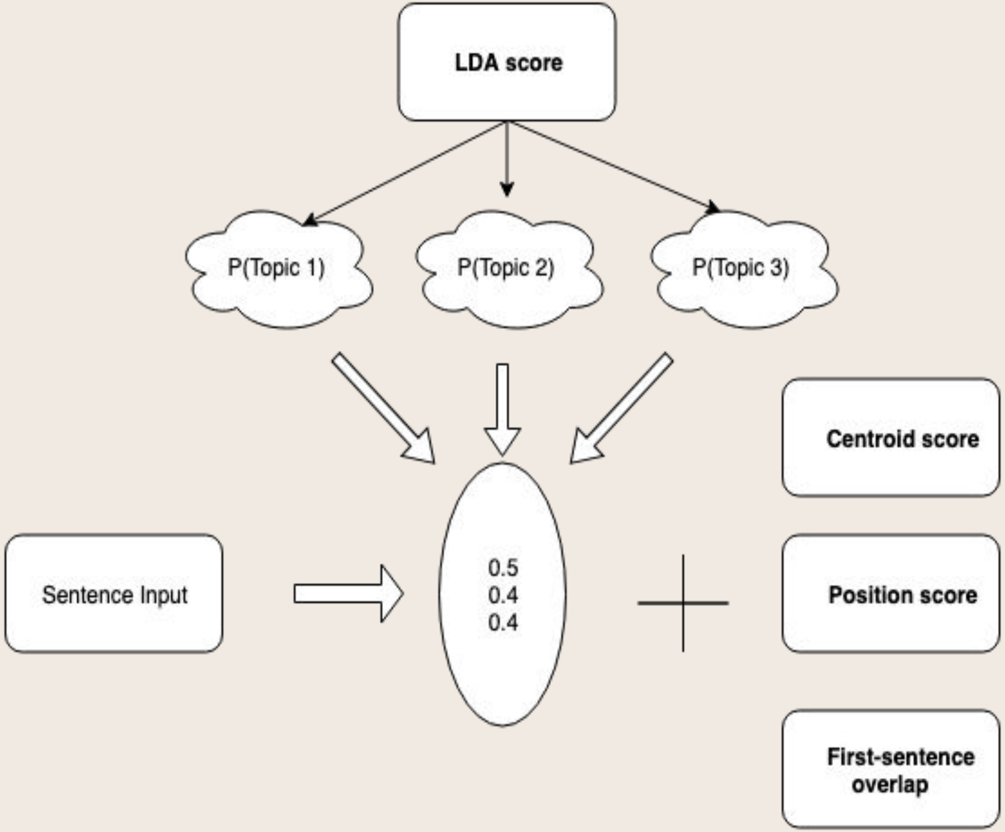
- Centroid score + Sentence position + First sentence overlap + Redundancy penalty = MEAD
- LDA probs for sentence weighted by MEAD score

Information Ordering

- Cohesion Gradient

Content Realization

- Sentence Compression



Cohesion Gradient

[**0.85** , 0.61 , 0.23 , 0.01]

[**0.74** , 0.05 , 0.15 , 0.43]

[**0.62** , 0.22 , **0.68** , 0.07]

[0.54 , 0.11 , **0.53** , 0.32]

[0.61 , 0.33 , **0.45** , **0.91**]

[0.32 , **0.88** , 0.41 , **0.76**]

[0.14 , **0.73** , 0.04 , 0.44]

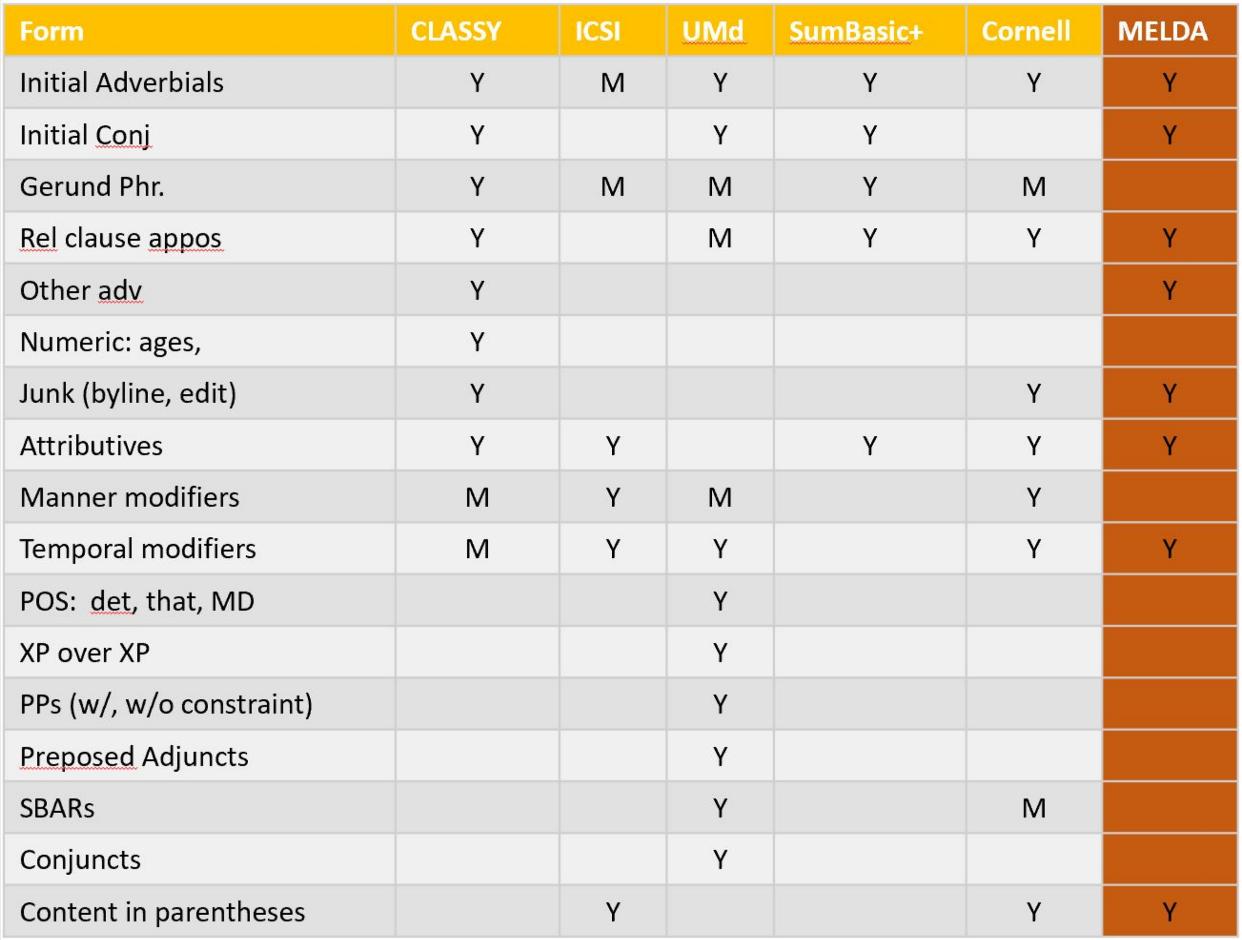
[0.02 , **0.68** , 0.17 , 0.39]

- Two options:
  - Start with lead sentence majority topic
  - Start with top MEAD scoring sentence*
- Select next highest topic until another topic score is higher
- Continue for new topic

*Thanks, Kevin, for the suggestion!

Sentence Compression

Using Spacy Dependency Parser, developed pattern based rules



Form	CLASSY	ICSI	UMd	SumBasic+	Cornell	MELDA
Initial Adverbials	Y	M	Y	Y	Y	Y
Initial Conj	Y		Y	Y		Y
Gerund Phr.	Y	M	M	Y	M	
Rel clause appos	Y		M	Y	Y	Y
Other adv	Y					Y
Numeric: ages,	Y					
Junk (byline, edit)	Y				Y	Y
Attributives	Y	Y		Y	Y	Y
Manner modifiers	M	Y	M		Y	
Temporal modifiers	M	Y	Y		Y	Y
POS: det, that, MD			Y			
XP over XP			Y			
PPs (w/, w/o constraint)			Y			
Preposed Adjuncts			Y			
SBARs			Y		M	
Conjuncts			Y			
Content in parentheses		Y			Y	Y

Sentence compression examples

Good

Damage from the Indian Ocean tsunami could have been reduced if coastal areas had maintained their protective shields of mangrove swamps and coral reefs.

(Damage from the Indian Ocean tsunami could have been significantly reduced if more coastal areas had maintained their protective shields of mangrove swamps and coral reefs, conservation groups said Thursday.)

The University of Maryland's Horn Point Laboratory in Cambridge is home to research on the Chesapeake Bay's native oyster.

(The University of Maryland's Horn Point Laboratory in Cambridge, Md., is home to research on the Chesapeake Bay's native oyster, Crassostrea virginica, and an Asian oyster that some officials want to introduce into the bay.)

???

Actor Robert Black was found not guilty Wednesday of shooting his wife.

(After a 12-week trial that ended with jurors saying they did not believe two Hollywood stuntmen central to the prosecution's case, actor Robert Blake was found not guilty Wednesday of fatally shooting his wife.)

Not-so-good

The concentration of nitrobenzene, kept falling.

(The concentration of nitrobenzene, the prime pollutant in the contaminated Songhua River following explosion of a petrochemical plant near the river earlier this month, has kept falling, said local environmental protection bureau on Saturday.)

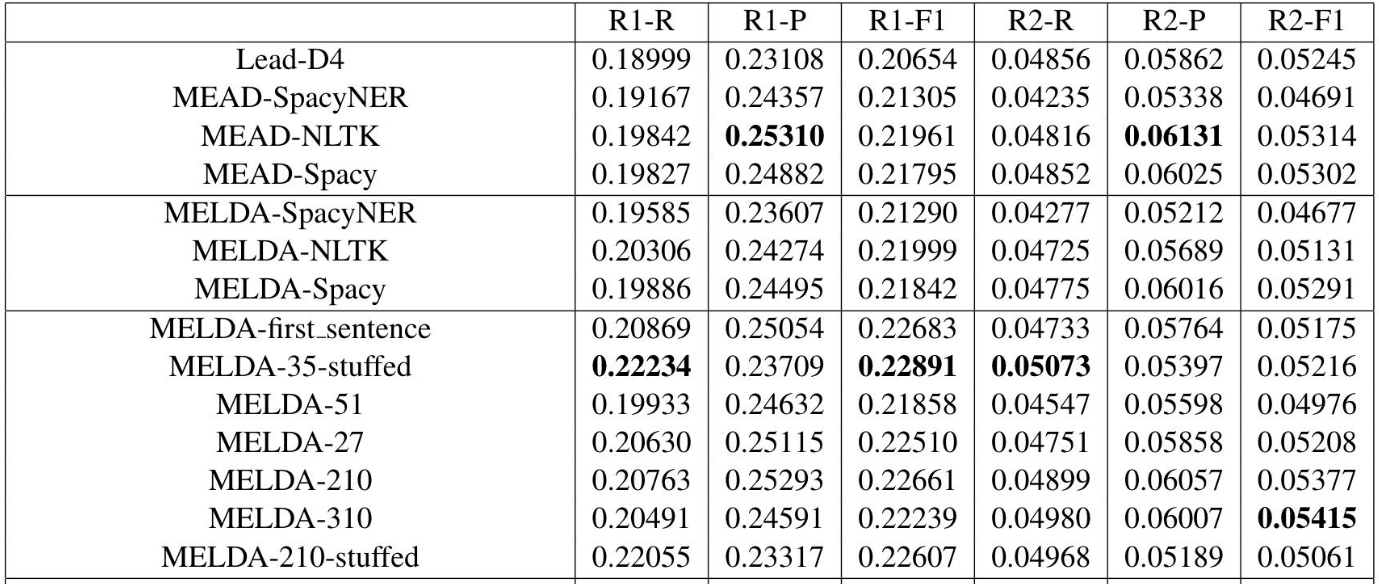
Two - thirds of Swiss voters.

(Two-thirds of Swiss voters on Sunday endorsed government-proposed legislation to allow limited stem cell research, local media reported.)

Researchers at Florida's Harbor Branch Oceanographic Institution have discovered that this year's rash of hurricanes scoured damaging seaweed, known as macroalgae, from coral reefs off the Florida coast, though they warned it is likely to come.

(Researchers at Florida's Harbor Branch Oceanographic Institution have discovered that this year's rash of hurricanes scoured damaging seaweed, known scientifically as macroalgae, from coral reefs off the Florida coast, though they warned it is likely to come back.)

Results - devtest



	R1-R	R1-P	R1-F1	R2-R	R2-P	R2-F1
Lead-D4	0.18999	0.23108	0.20654	0.04856	0.05862	0.05245
MEAD-SpacyNER	0.19167	0.24357	0.21305	0.04235	0.05338	0.04691
MEAD-NLTK	0.19842	**0.25310**	0.21961	0.04816	**0.06131**	0.05314
MEAD-Spacy	0.19827	0.24882	0.21795	0.04852	0.06025	0.05302
MELDA-SpacyNER	0.19585	0.23607	0.21290	0.04277	0.05212	0.04677
MELDA-NLTK	0.20306	0.24274	0.21999	0.04725	0.05689	0.05131
MELDA-Spacy	0.19886	0.24495	0.21842	0.04775	0.06016	0.05291
MELDA-first_sentence	0.20869	0.25054	0.22683	0.04733	0.05764	0.05175
MELDA-35-stuffed	**0.22234**	0.23709	**0.22891**	**0.05073**	0.05397	0.05216
MELDA-51	0.19933	0.24632	0.21858	0.04547	0.05598	0.04976
MELDA-27	0.20630	0.25115	0.22510	0.04751	0.05858	0.05208
MELDA-210	0.20763	0.25293	0.22661	0.04899	0.06057	0.05377
MELDA-310	0.20491	0.24591	0.22239	0.04980	0.06007	**0.05415**
MELDA-210-stuffed	0.22055	0.23317	0.22607	0.04968	0.05189	0.05061

Key

Spacy - Spacy package for tok.
NLTK - NLTK package for tok.

First_sentence - info ordering by topic first sentence topic (vs. by mead centroid score)
Stuffed - summaries filled to capacity with sentences (despite cohesion)

KN
K - number of LDA topics
N - number of sentences per topic

History of the devset results

Each version, with default parameters

	ROUGE-1 F score	ROUGE-2 F score
D2 (LEAD) D2 (MEAD)	0.13792 0.14540	0.02417 0.03078
D3 (LEAD) D3 (MEAD) D3 (MELDA)	0.20364 0.21247 0.18818	0.05247 0.04717 0.04474
D4 (LEAD) D4 (MEAD) D4 (MELDA)	0.20654 0.21961 0.22891	0.05247 0.05314 0.05216

Bad

D1005

They had treated monkeys with Parkinson's disease through a stem cell transplant, paving the way for an ideal remedy to the intractable disease.
Exercise was enough to prevent the degeneration of brain cells in rats with Parkinson's disease, University of Pittsburgh researchers report.
A new medical study helps explain the puzzling effect of the hormone dopamine on patients suffering from the degenerative nerve disorder **out**.
Scientists believe they have pinpointed a tiny genetic flaw that is to blame for around four percent of all cases of Parkinson's **,**
South Korea will allow stem - cell research to find cures.

They had treated monkeys with **Parkinson's** disease through a stem cell transplant, paving the way for an ideal remedy to the intractable disease.
Californians voted Tuesday to become the first state to fund controversial **embryonic stem cell research** that was banned by President George W. Bush.
A team of Egyptian doctors examined **Yasser Arafat** after he suffered from fever, nausea and a stuffy nose.
South Korea will allow stem - cell research to find cures.
Do you know any **kids who have diabetes?**
Two - thirds of **Swiss voters.**

Issues

- Imperfect sentence compression
- Preprocessing doesn’t help - NER, coreference resolution, lemmatization
- Nondeterministic lemmatization
- Spacy tokenizer is worse than NLTK
- Runtime - 20-40 minutes!

Future Work

- Smarter sentence compression (discourse parser?)
- Move sentence compression before info ordering or before content selection
- Building IDF array takes a long time (could pickle during dev)
- Coreference resolution (pre or post processing)
- Parameter optimization

Related Readings

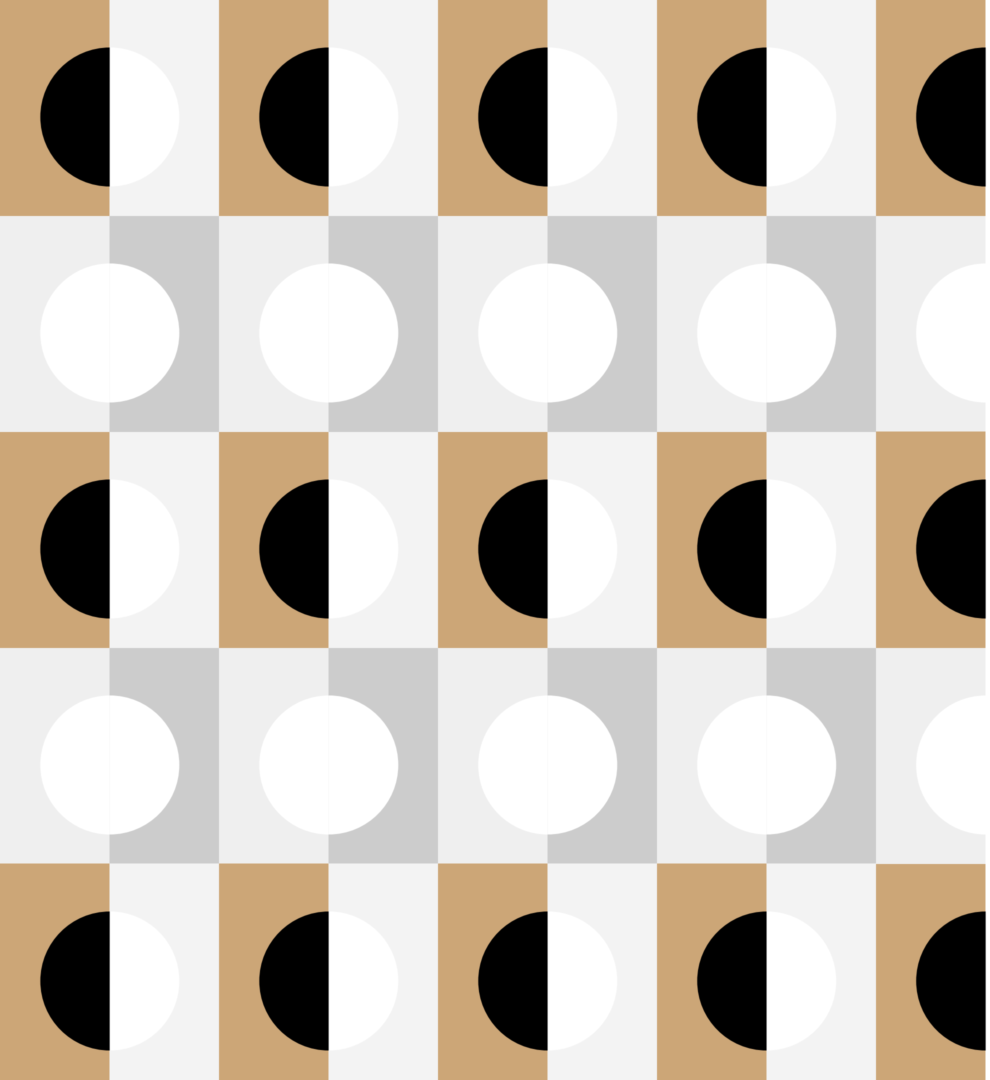
Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” Journal of machine Learning research 3.Jan (2003): 993-1022.

Conroy, John M, and Judith D Schlesinger. “Back to Basics: CLASSY 2006,” n.d., 9.

Radev, Dragomir R., Sasha Blair-Goldensohn, and Zhu Zhang. ”Experiments in single and multidocument summarization using MEAD.” First document understanding conference. 2001.

Radev, Dragomir, Adam Winkel, and Michael Topper. ”Multi document centroid-based text summarization.” ACL 2002. 2002.

Radev, Dragomir R., et al. ”Centroid-based summarization of multiple documents.” Information Processing & Management 40.6 (2004): 919-938.



Thank you

Sample Outputs

D1002

With the indictments unsealed against four police officers in the Amadou Diallo shooting, a battle is taking shape over physical evidence in the case, as lawyers and experts seek to buttress their own versions of what happened based on entrance wounds, bullet trajectories and other forensic details.
Albany questioning the change in venue from the Bronx to Albany, lawmakers and activists called on the U.S. Justice Department Monday to monitor the upcoming murder trial of four police officers.
An appellate court ordered the trial.
The cheering began as the parents of Amadou Diallo left the prosecutor’s office.
The Rev.
(LDA topics: 3; sentences per topic: 5; “stuff” summary until full)

With the indictments unsealed against four police officers in the Amadou Diallo shooting, a battle is taking shape over physical evidence in the case, as lawyers and experts seek to buttress their own versions of what happened based on entrance wounds, bullet trajectories and other forensic details.
Albany questioning the change in venue from the Bronx to Albany, lawmakers and activists called on the U.S. Justice Department Monday to monitor the upcoming murder trial of four police officers.
An appellate court ordered the trial.
On the seesaw that is political life in New York, the Rev.
The Rev.
(LDA topics: 2; sentences per topic: 10; no stuffing until full)

Sample Outputs

D1002

With the indictments unsealed against four police officers in the Amadou Diallo shooting, a battle is taking shape over physical evidence in the case, as lawyers and experts seek to buttress their own versions of what happened based on entrance wounds, bullet trajectories and other forensic details.

Albany questioning the change in venue from the Bronx to Albany, lawmakers and activists called on the U.S. Justice Department Monday to monitor the upcoming murder trial of four police officers.

An appellate court ordered the trial.

On the seesaw that is political life in New York, the Rev.

The Rev.

(LDA topics: 2; sentences per topic: 10; “stuff” until full)

With the indictments unsealed against four police officers in the Amadou Diallo shooting, a battle is taking shape over physical evidence in the case, as lawyers and experts seek to buttress their own versions of what happened based on entrance wounds, bullet trajectories and other forensic details.

Albany questioning the change in venue from the Bronx to Albany, lawmakers and activists called on the U.S. Justice Department Monday to monitor the upcoming murder trial of four police officers.

An appellate court ordered the trial.

A shooting that has led.

The Rev.

Valenzuela is free.

The Rev.

(LDA topics: 2; sentences per topic: 10; no stuffing until full)