



AWS  
re:Invent

**AIM 425 - R**

# Understanding a large amount of text by modeling & visualizing topics

**Angela Wang**

R&D engineer

Amazon Web Services











# Set up

<https://bit.ly/large-text-understanding>


# THE AWS ML STACK

Broadest and deepest set of capabilities















## Application Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
 REKOGNITION IMAGE	 REKOGNITION VIDEO	 TEXTRACT	 POLLY	 TRANSCRIBE	 TRANSLATE	 COMPREHEND & COMPREHEND MEDICAL	 LEX	 FORECAST	 PERSONALIZE

## Platform Services

 <b>Amazon SageMaker</b>	Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting
--	--------------	-----------	--------------------------	------------------------	----------	--------------	------------	---------

## ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES	INFRASTRUCTURE								
 TensorFlow  mxnet  PYTORCH	 GLUON  K Keras	 EC2 P3 & P3DN	 EC2 G4 EC2 C5	 FPGAS	 DL CONTAINERS & AMIs	 ELASTIC CONTAINER SERVICE	 ELASTIC KUBERNETES SERVICE	 GREENGRASS	 ELASTIC INFERENCE	 INFERENCE

# Topic Detection

## Topics

 Government

 Information Technology

 Politics

The New York Times

U.S.

## High-Tech Industry, Long Shy of Politics, Is Now Belle of Ball

By LIZETTE ALVAREZ DEC. 26, 1999

### Correction Appended

At a time when Congress is bitterly divided and unable to reach consensus on issues like gun control and health care, Democrats and Republicans are happily reaching across party lines to pass legislation backed by high-tech companies.

The high-tech industry, at the same moment, is lavishing new attention on Washington and changing its once-alloof posture toward the federal government.

Republicans and Democrats are both eager to win the loyalties of high-tech companies and executives, knowing that they represent untold jobs, wealth and ultimately votes and campaign contributions.

For its part, the industry has realized that the federal government can do its members as much harm as good. Microsoft, and its battle with the Justice Department, along with a spate of other threatened legal problems, drilled this point home.

"Microsoft was a poster child for our industry," said Connie Correll, director of communications for the Information Technology Industry Council, a trade organization that represents America Online, Dell and I.B.M., among others.

In the House, Representative David Dreier, Republican of California, Robert W. Goodlatte, Republican of Virginia, and Thomas M. Davis III, a Virginia Republican whose district includes a growing number of high-tech companies, among others, are viewed as industry experts.

The New Democrats are often willing to buck their own leadership to live up to the label as technology boosters. In the past two years, their numbers have grown, to 64 this year from 41 in 1997.

These New Democrats have breakfast with high-tech executives every week and visit Silicon Valley routinely. "I think they are trying to create a mini high-tech party in a way," said Wade Randlett, a co-founder of TechNet and now an executive at Red Gorilla, an Internet company. "It's a smart political approach."

The bulging docket of high-tech issues had led to dramatic growth in the industry's lobbying on Capitol Hill. Internet and software companies have snagged some of Capitol Hill's best talent this year -- former Congressional aides whose expertise blends technology and politics. While Microsoft built a large battery of lobbyists, beginning two years ago, few other high-tech companies had Washington lobbying operations. Suddenly, companies, including Yahoo and Gateway, are busily opening Washington outposts, hiring lobbyists and starting trade associations. And in yet another coming-of-age gesture, executives are beginning to dip into their coffers.

"It's so critical that Silicon Valley be involved in Washington right now," said Chris Larsen, the 39-year-old founder of E-Loan who has made six trips here this year to meet with members. "The stakes are really high."

High-tech lobbying has become so profitable that some high-powered lobbyists are carving out specialty niches and crossing party lines to do it. Edward W. Gillespie, a former policy and communications director for Representative Dick Arme of Texas, the House majority leader, and Jack Quinn, Vice President Al Gore's former chief of staff, have teamed up to start a firm. The two have worked together on legislation to ease encryption export restrictions, a high priority for the high-tech industry. Tony Podesta, the brother of John D. Podesta, White House chief of staff,

# Topic Modeling

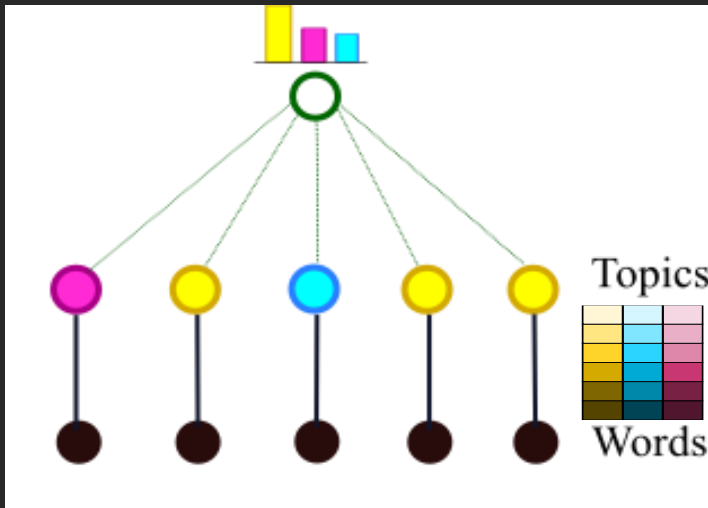
- Use case:
  - Understand trends and patterns of historical documents
  - document classification and automatic content tagging
  - document summarization
  - content recommendation
- Unsupervised learning
  - labeled sample documents hard to obtain
  - discover topics automatically

# Topic Modeling in Amazon Comprehend

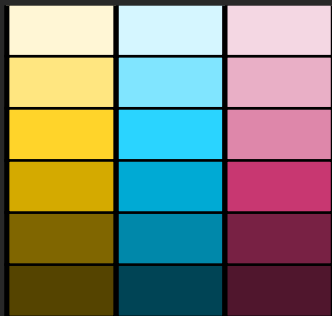
Document corpus



LDA Model  
(Latent Dirichlet Allocation)

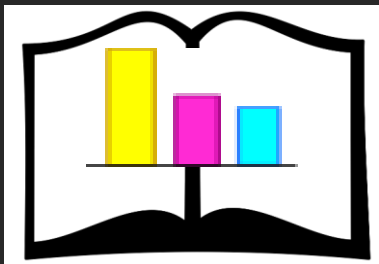


Keywords Topic Groups



+

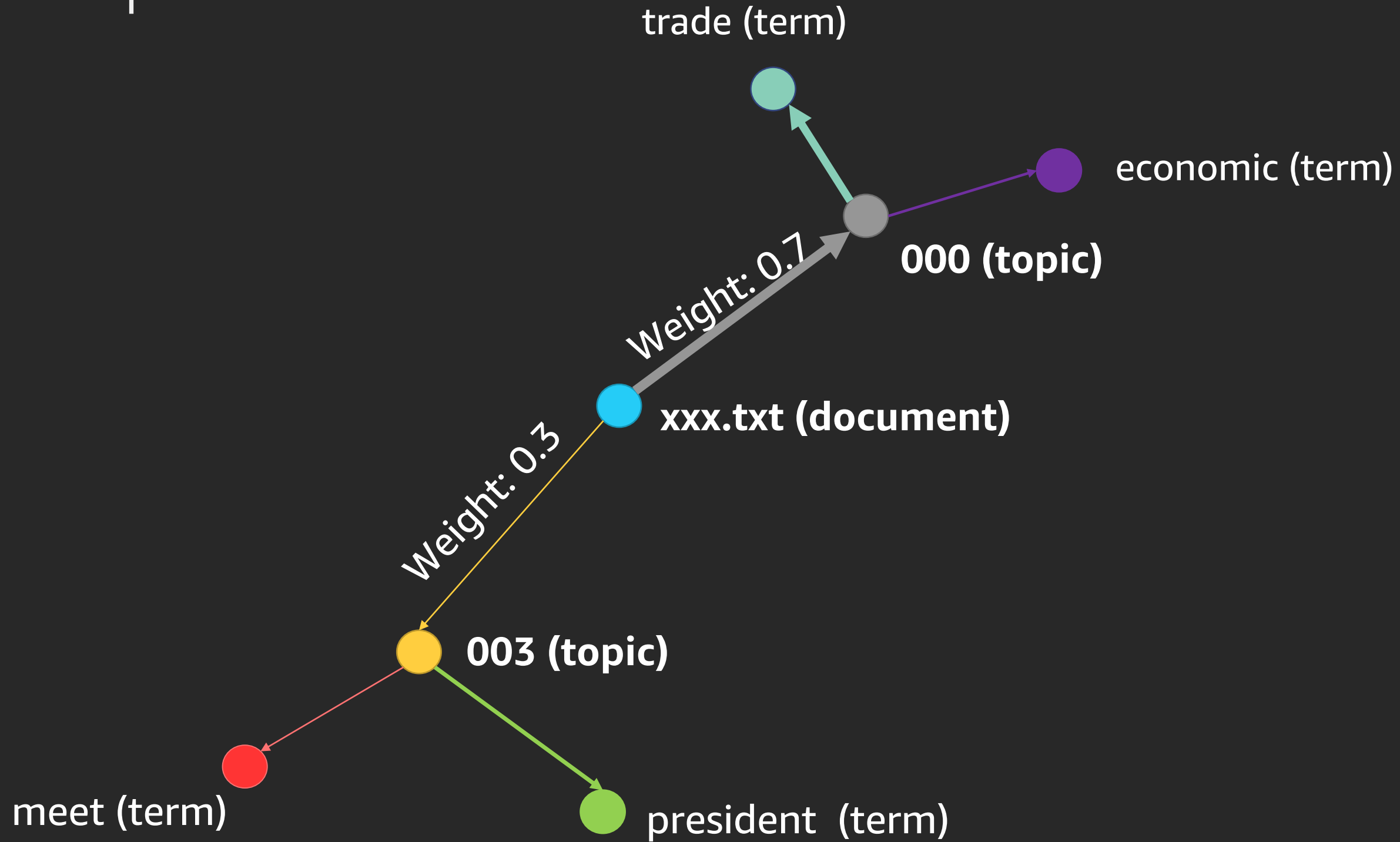
Document Relationship  
to Topics



Topic	Term	Weight
0	team	0.1185
0	game	0.1061
0	player	0.0316
1	cup	0.2052
1	food	0.0407
1	minutes	0.0361
1	add	0.0297
1	tablespoon	0.0288

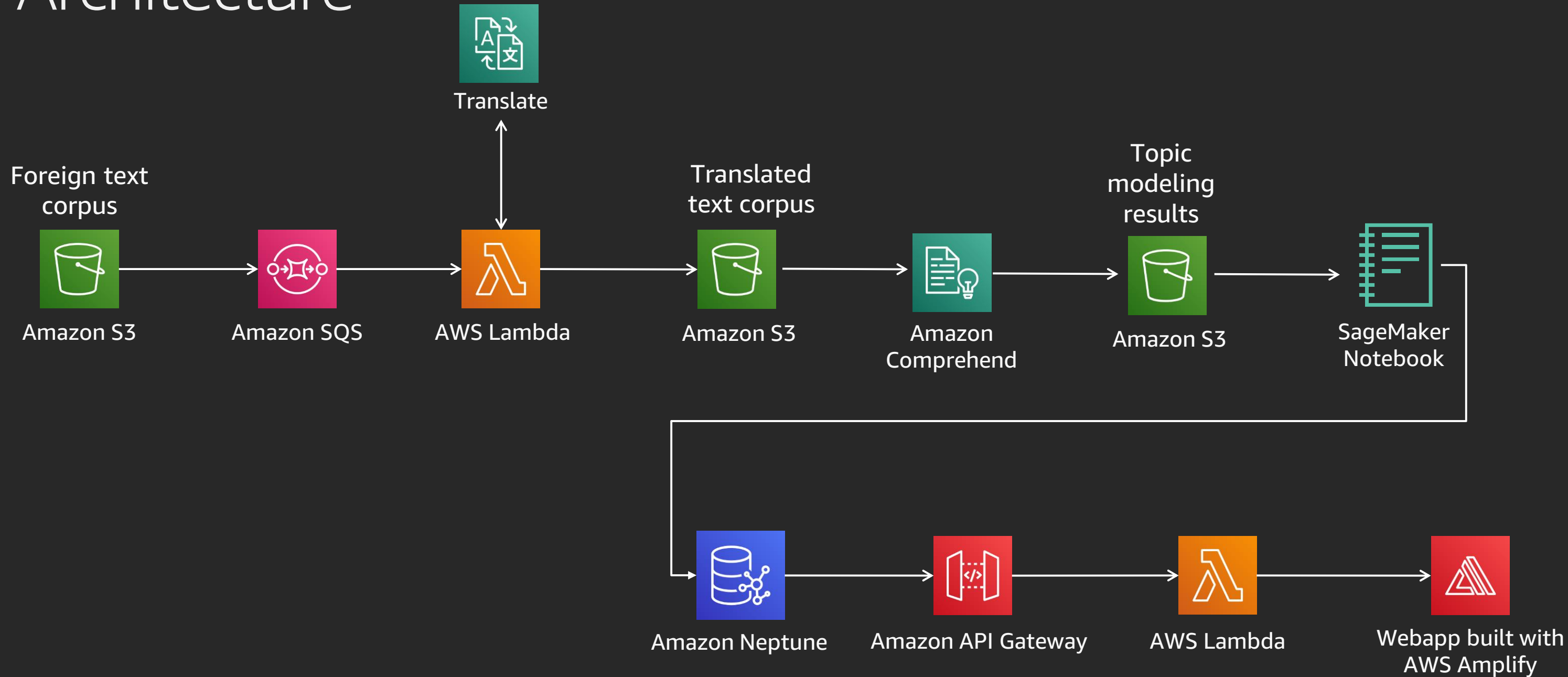
Document	Topic	Proportion
sample-doc1	0	0.5234
sample-doc1	1	0.3043
sample-doc2	2	0.9984

# Graph database

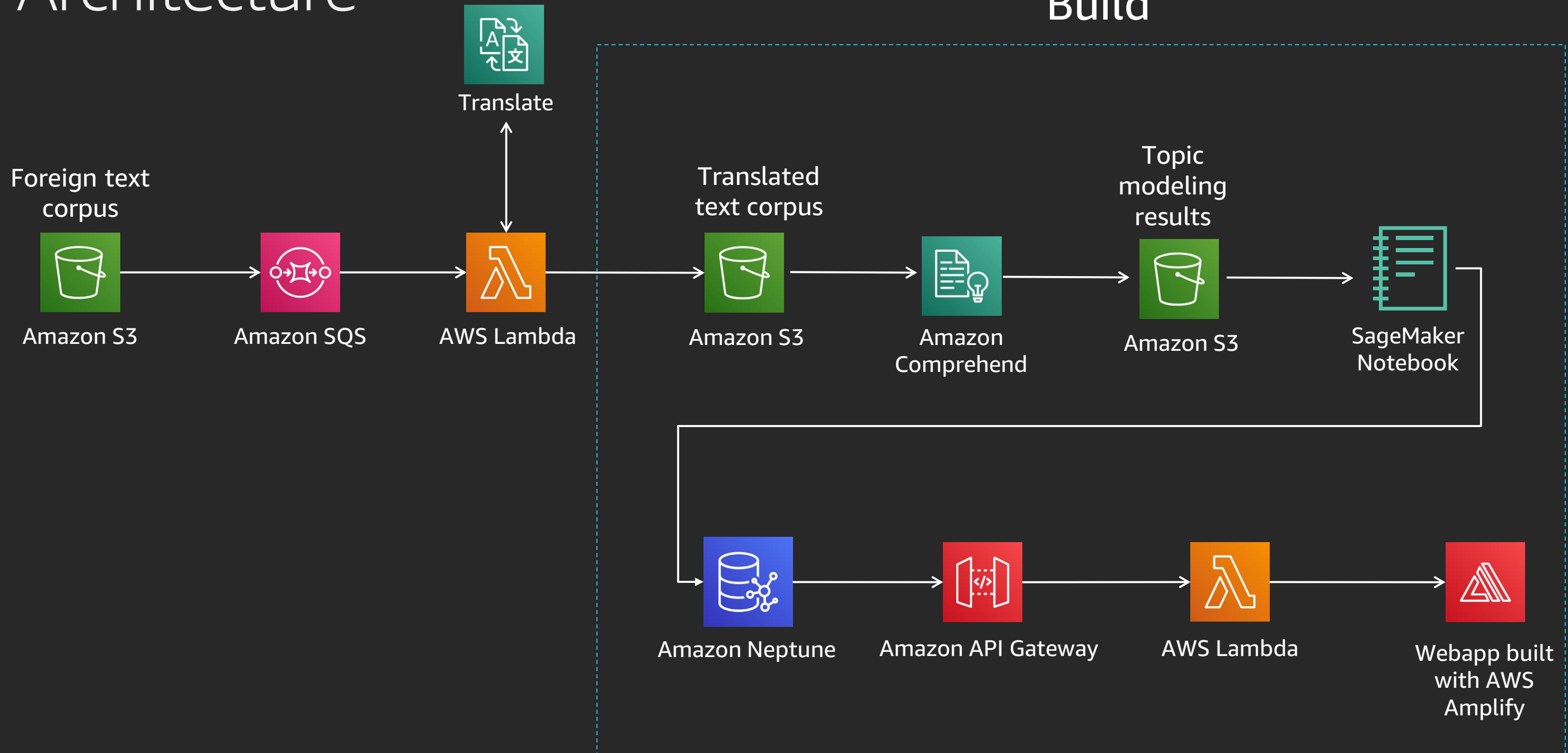




# Architecture



# Architecture



# Data source

- THUCNews (THU Chinese News)
  - Natural Language Processing and Computational Social Science Lab, Tsinghua University
  - <http://thuctc.thunlp.org/>
  - Sina (sina.com.cn) news RSS feed 2005~2011
  - 740,000 articles (2,000 for the builder session)
  - 2.19GB total
  - 14 subcategories



<https://bit.ly/large-text-understanding>

# Thank you!

**Angela Wang**

angelaw@amazon.com



Please complete the session  
survey in the mobile app.