Extracting and Analyzing Semantic Relatedness between Cities Using News Articles

Yingjie Hu[1], Xinyue Ye[2], and Shih-Lung Shaw[1]

[1]Department of Geography, University of Tennessee, Knoxville
[2]Department of Geography, Kent State University, Kent

Apr. 8, 2017

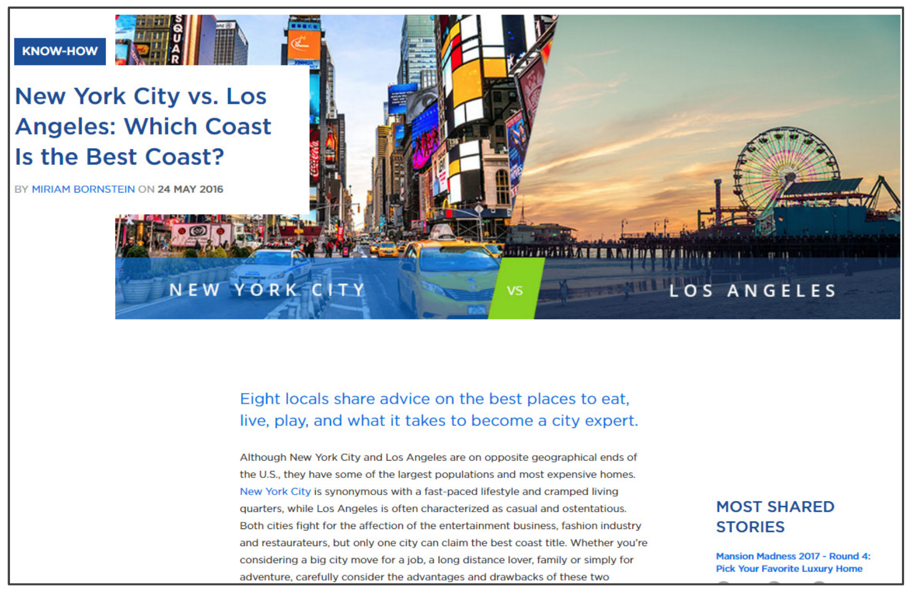
Introduction

- News articles are rich sources of information

- Diverse topics
  - Economy, politics, science, sports, …
- Various entities
  - Persons, organizations, places, …
- Timely information
  - Prompt report of latest events

Introduction

- Cities, as hubs of human activities, are frequently mentioned in news articles
- Two or more cities may co-occur in the same news article
- E.g., comparing the lifestyles of two cities


KNOW-HOW

New York City vs. Los Angeles: Which Coast Is the Best Coast?

BY MIRIAM BORNSTEIN ON 24 MAY 2016

Eight locals share advice on the best places to eat, live, play, and what it takes to become a city expert.

Although New York City and Los Angeles are on opposite geographical ends of the U.S., they have some of the largest populations and most expensive homes. New York City is synonymous with a fast-paced lifestyle and cramped living quarters, while Los Angeles is often characterized as casual and ostentatious. Both cities fight for the affection of the entertainment business, fashion industry and restaurateurs, but only one city can claim the best coast title. Whether you're considering a big city move for a job, a long distance lover, family or simply for adventure, carefully consider the advantages and drawbacks of these two

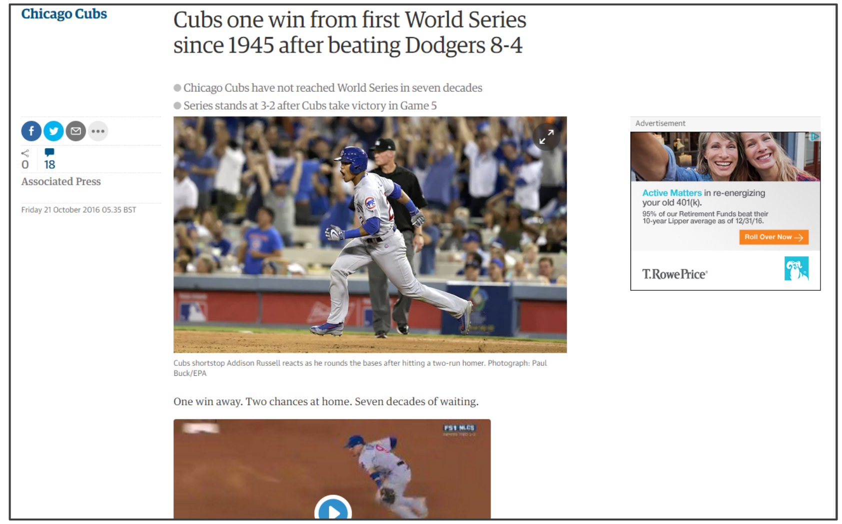
MOST SHARED STORIES

Mansion Madness 2017 - Round 4: Pick Your Favorite Luxury Home

Los Angeles & New York City

Introduction

- E.g., sports may draw teams from two cities together



Los Angeles & Chicago

- E.g., cities may address environmental issues collaboratively

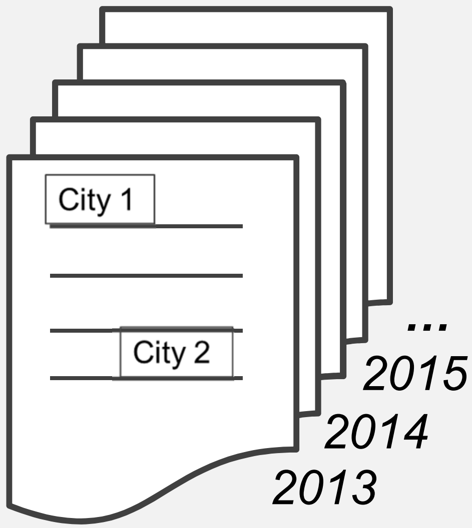
Los Angeles & New Orleans

Introduction

- Cities can be related under a variety of topics (semantic relatedness)
- Such semantic relatedness is partially captured in news articles
- Objective: to develop a computational framework that can automatically process a large number of news articles and extract semantic relatedness

Problem Formalization

Input

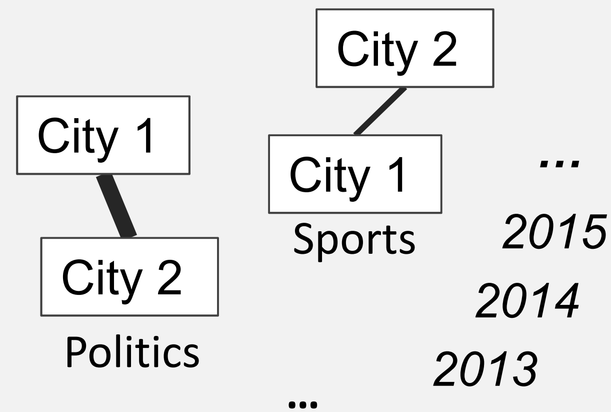


Culture
Business
Environment
Politics
Education
…
Sports

News Articles based on:
- A set of cities $\{c_i\}$
- A set of time periods $\{y_k\}$

Semantic Topics defined by:
- A set of topic terms $\{t_m\}$

Output

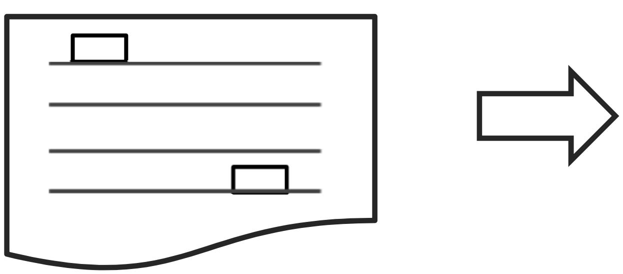


Semantic Relatedness:
- A set of relatedness values $\{r_{ijkm}\}$

Problem Formalization

- Core idea:
  1) Identify the topics of news articles
  2) Assign the topics to the cities
  3) Quantify the semantic relatedness

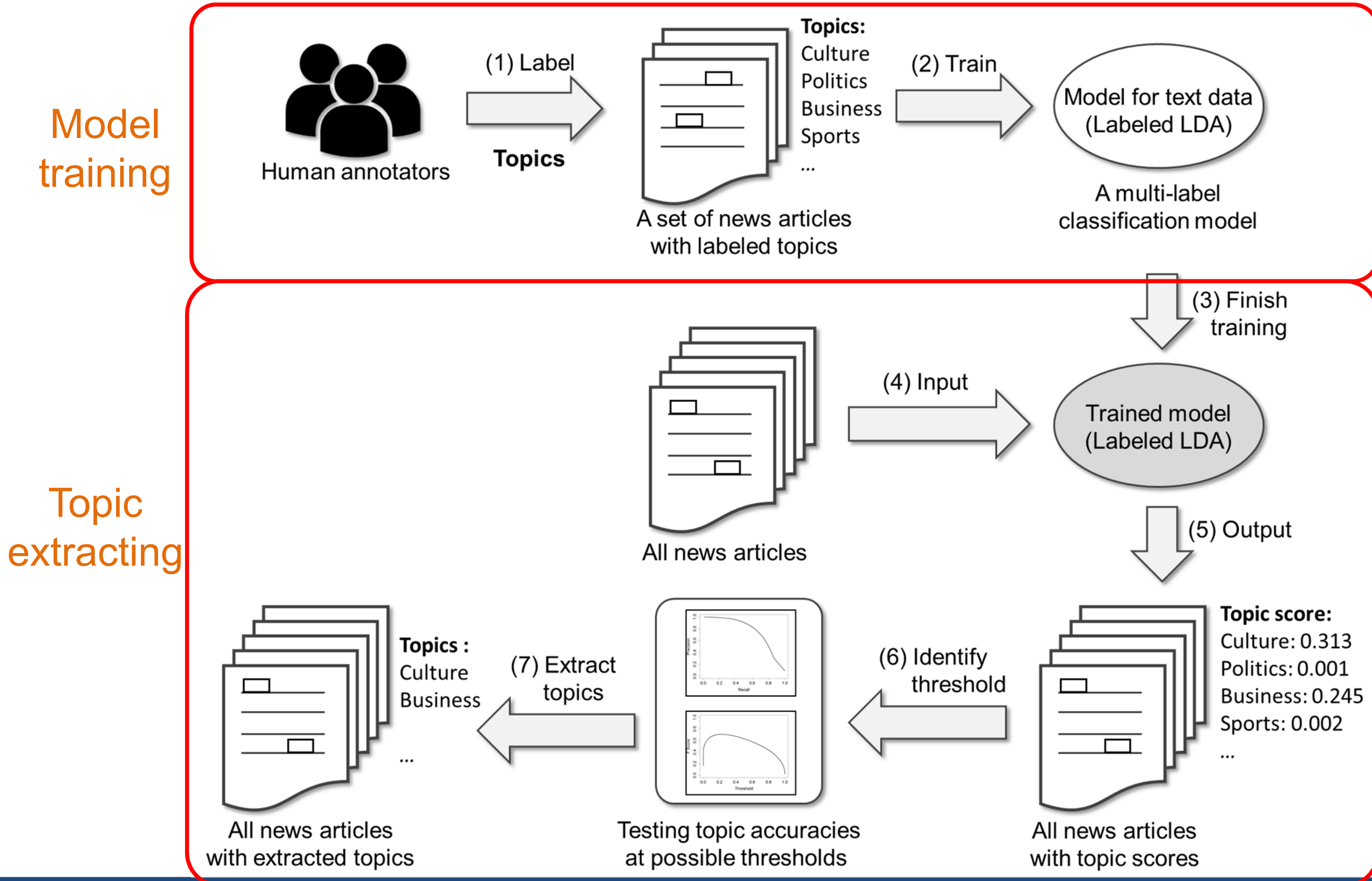
- Key question: given a news article, which topics is it talking about?



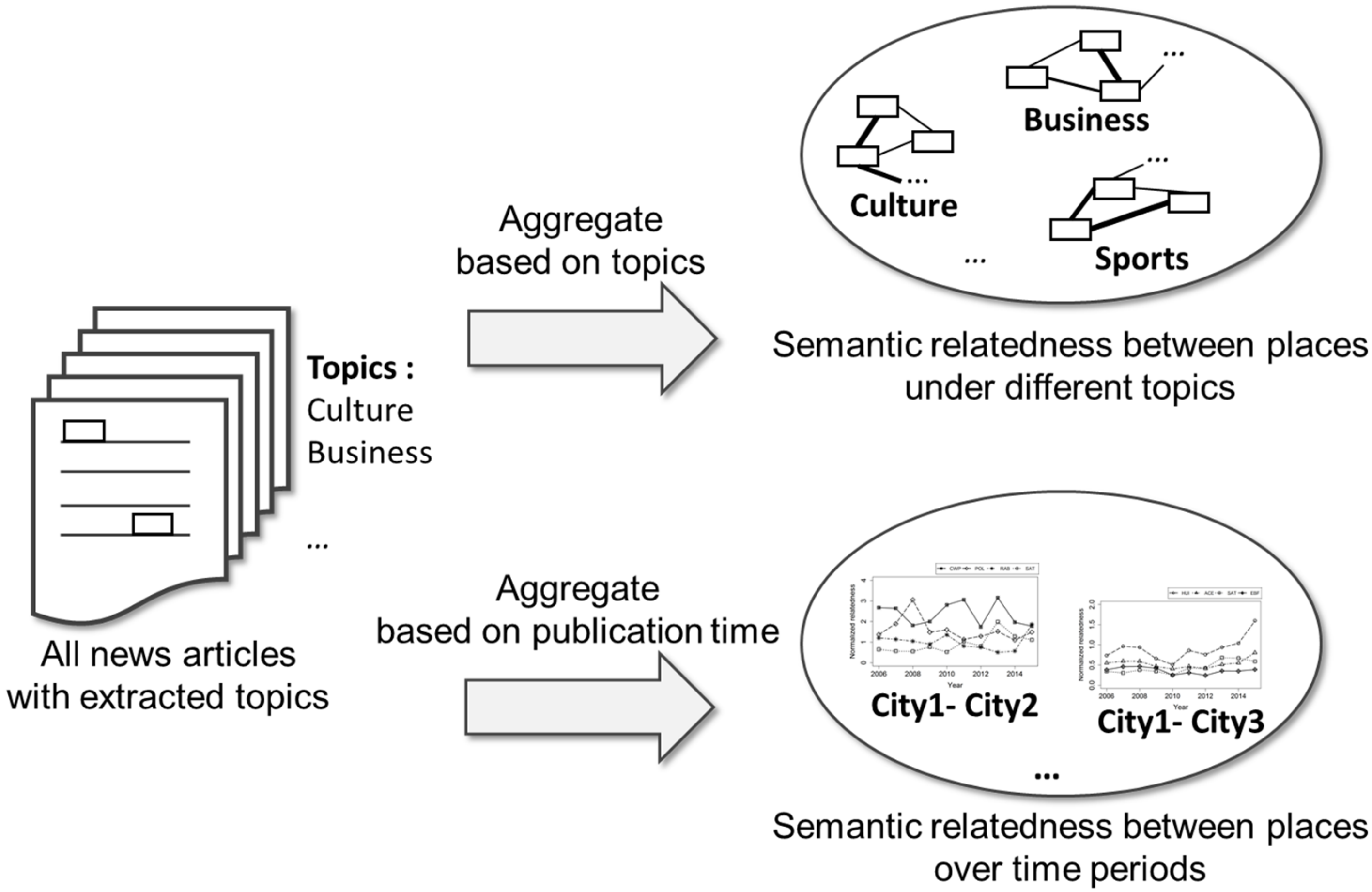
Culture?
Business?
Environment?
…

Multi-label classification problem

Framework


Model training
Human annotators
(1) Label
Topics
Topics:
Culture
Politics
Business
Sports
...
A set of news articles with labeled topics
(2) Train
Model for text data (Labeled LDA)
A multi-label classification model
(3) Finish training
Topic extracting
All news articles
(4) Input
Trained model (Labeled LDA)
(5) Output
Topic score:
Culture: 0.313
Politics: 0.001
Business: 0.245
Sports: 0.002
...
All news articles with topic scores
(6) Identify threshold
Testing topic accuracies at possible thresholds
(7) Extract topics
Topics :
Culture
Business
...
All news articles with extracted topics

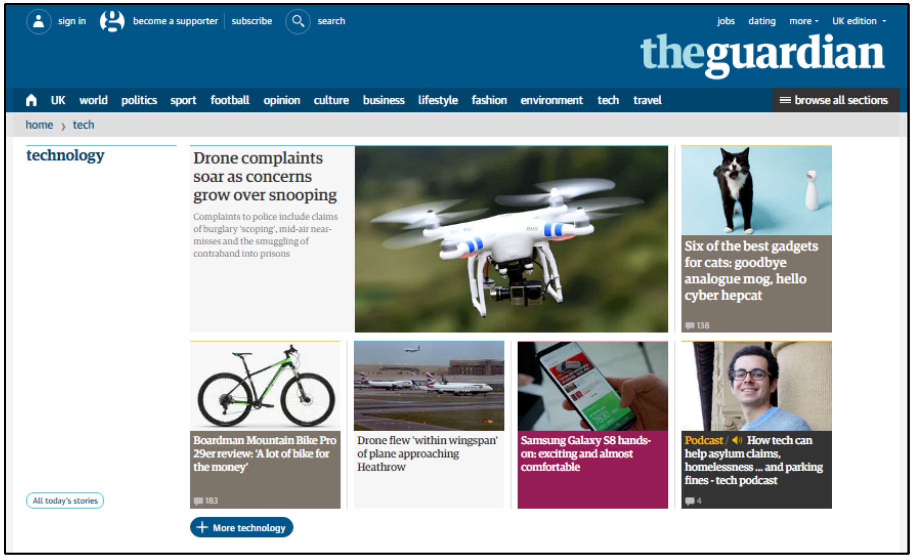
Framework


Business
Culture
Sports
Aggregate
based on topics
Semantic relatedness between places
under different topics
Topics :
Culture
Business
All news articles
with extracted topics
Aggregate
based on publication time
City1- City2
City1- City3
Semantic relatedness between places
over time periods

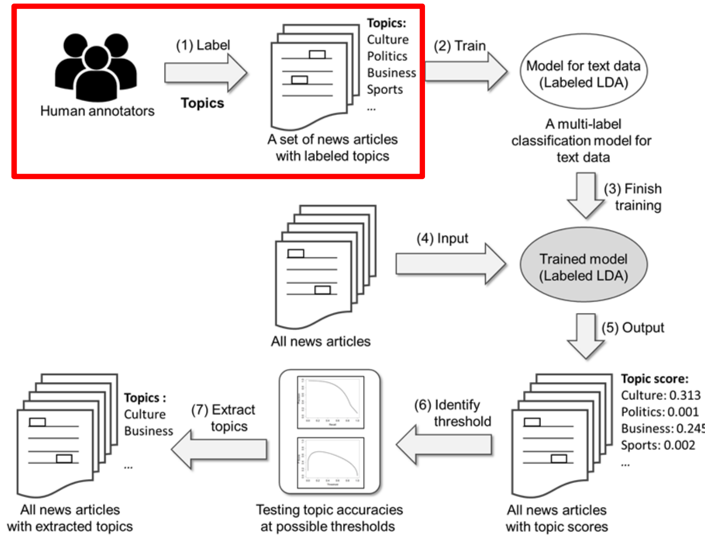
Experiments

- Cities: top 100 cities in the contiguous U.S.
- Time: 1/1/2006 and 12/31/2015

- News articles from The Guardian
  - 543,824 news articles



Experiments

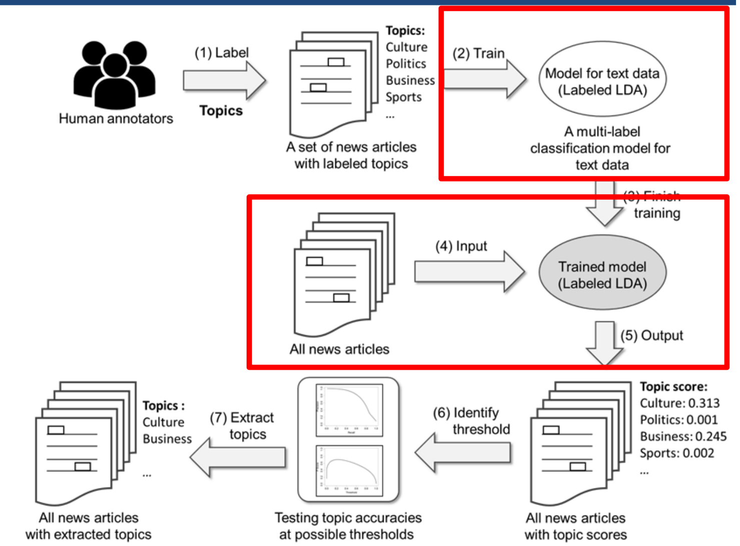


- 17 semantic topics from IPTC
  - E.g., Culture, Politics, Sports, Disaster, Crime, …

- Obtaining training data
  - Existing news tags
  - Mapping some tags to topics
  - 141,765 training data records

IPTC Topic	News Tags
Arts, Culture and Entertainment	culture, music, film, media, books, artanddesign, television, art, fashion, festivals, history, comedy, museums, opera, drama, poetry, documentary, painting, theatre, sculpture

Experiments

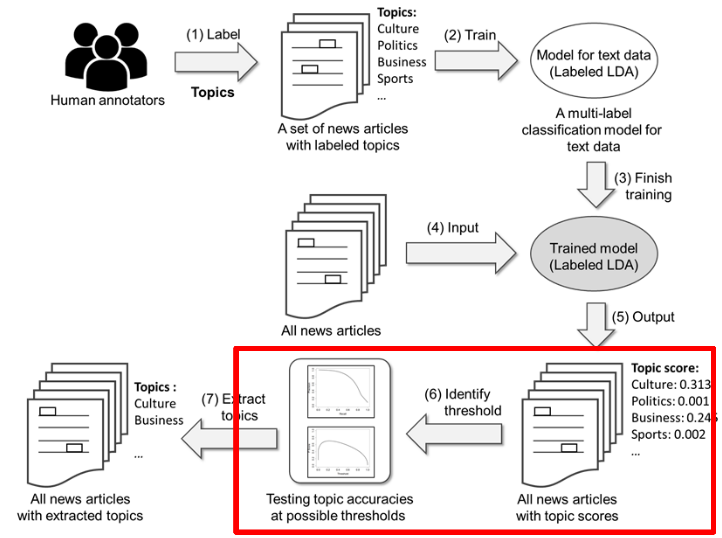


- Training the LLDA model

- Topic extracting
  - Applying the trained LLDA model to all news articles

A training data record

IPTC labels	Processed text
Artscultureandentertainment Lifestyleandleisure	la hard city accept reality suffer personal setback illness sick gorgeous setting move santa monica beach...



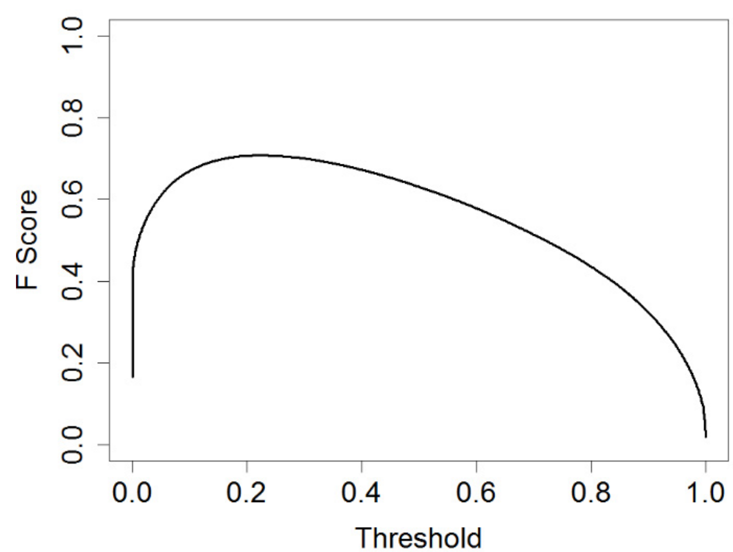
Experiments

- Identifying suitable threshold

$$Precision = \frac{|Extracted\ Relevant\ Topics|}{|All\ Extracted\ Topics|}$$

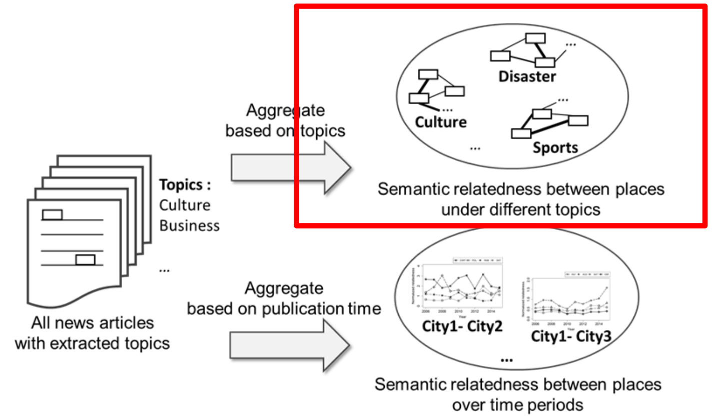
$$Recall = \frac{|Extracted\ Relevant\ Topics|}{|All\ Relevant\ Topics|}$$

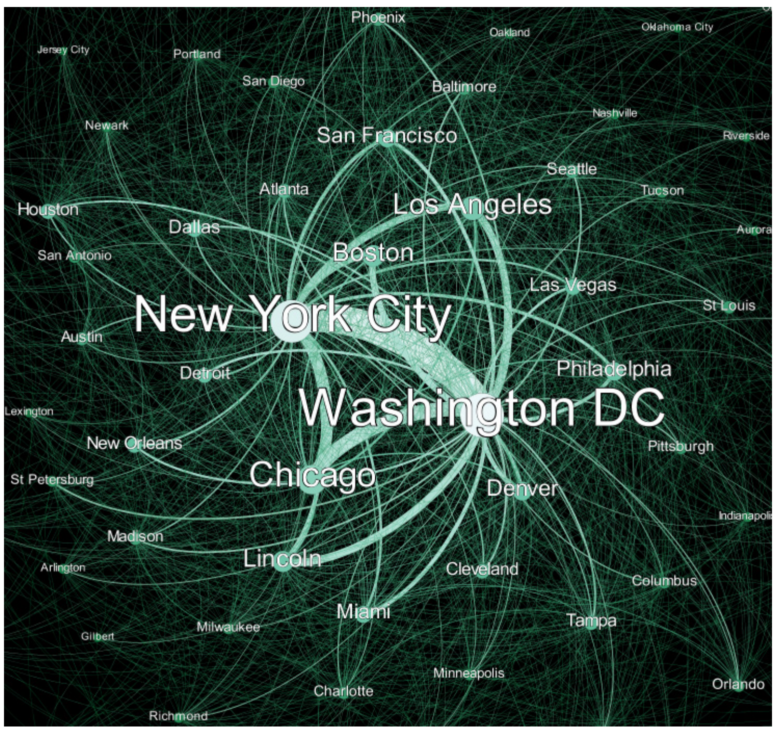
$$F\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$



Experiments

- Visualize city relatedness
  - Based on semantic topics



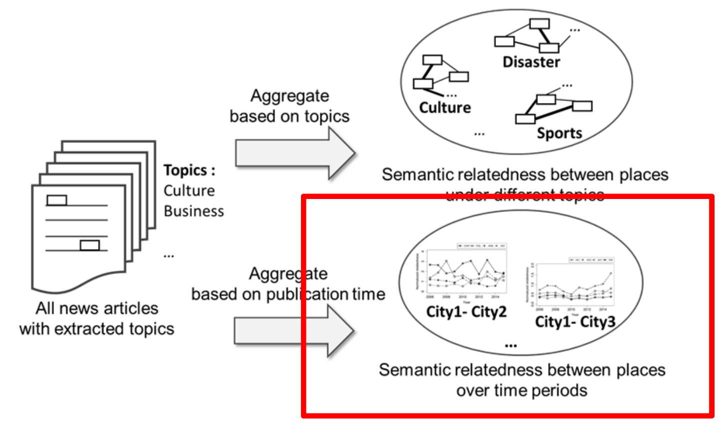


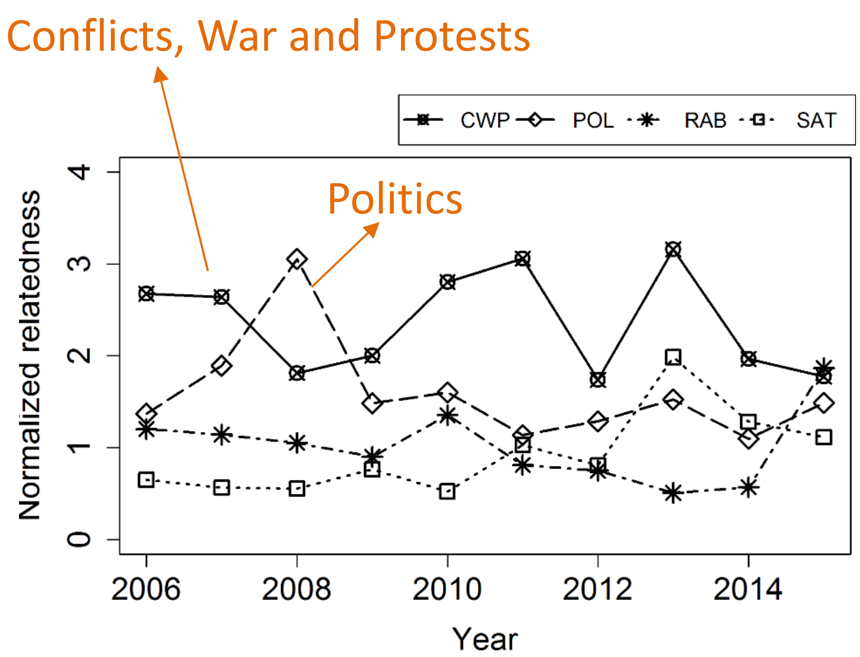
Politics

Science and Technology

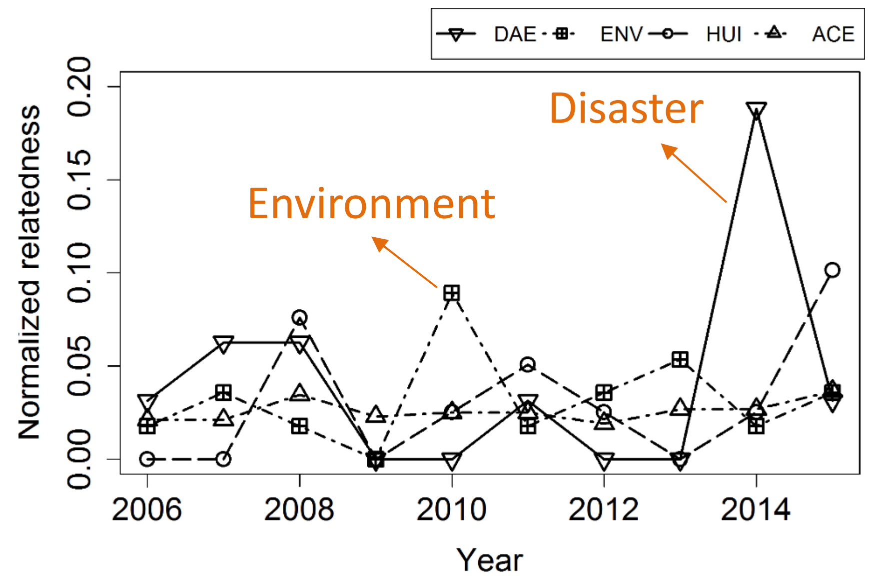
Experiments

- Visualize city relatedness
  - Based on publication time





NYC and Washington DC



Los Angeles and New Orleans

Experiments

- Both Los Angeles and New Orleans were enrolled in the program in 2014

100 RESILIENT CITIES   ABOUT US   NEWS   OUR CITIES   OUR PARTNERS

How Cities Recover from Natural Disasters

11.19.14 | BY DAVID SCHREINER

What can cities do to plan for and recover more effectively from disasters, especially when those disasters are continuing to increase in frequency and severity?

The *Financial Times* spoke with experts from the U.S. Department of Housing and Urban Development's (HUD) Rebuild by Design and 100 Resilient Cities, and suggests three broad steps, whether cities face too much water or too little, extreme heat or record cold:

We began working with our first group of 32 cities in December of 2013. In 2014, we received 330 applications from 94 countries for our second cohort, and we announced the 35 cities of round 2 in December. The third 100 Resilient Cities Challenge closed in November of 2015 and we announced our final group of cities in May 2016.

Distance Decay Analysis

- A weak distance decay effect was found in a previous research based on place co-occurrence in news articles *(Liu et al. 2014, Transactions in GIS)*

$$c_{ij} \propto \frac{c_i c_j}{d_{ij}^{\beta}}$$

- β is the friction coefficient; $\beta = 0.2$ in *Liu et al. 2014*

- City relatedness under different topics might have different distance decay effects

Distance Decay Analysis

All news: $\beta = 0.23$

Topic	β
Arts, Culture and Entertainment	0.21
Sport	0.08
Crime, Law and Justice	0.37
Science and Technology	0.19
Politics	0.32

Conclusions

- News articles partially capture the semantic relatedness between cities
- A computational framework is developed to “read” a large number of news articles and extract semantic relatedness
- An experiment based on more than 500,000 news articles shows different network structures and temporal variations
- Varied distance decay effects were observed for the different semantic relatedness

Thank You!

Questions?

Yingjie Hu
yhu21@utk.edu