Geographical Topic Discovery and Comparison

Zhijun Yin, Liangliang Cao, Jiawei Han,
Chengxiang Zhai, Thomas Huang

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

To appear in **WWW'11**

Presenter: Jeff Huang

Outline

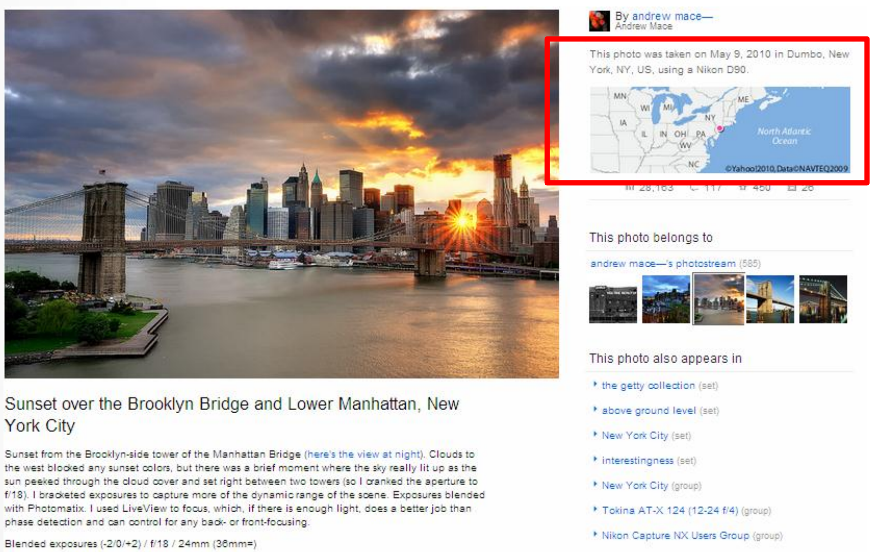
- Motivation
- Problem Formulation
- Solution Sketch
- Experiments
- Q/A

Motivation

- GPS records are popular on the Web
  - Advanced cameras with GPS receivers could record GPS locations when the photos were taken.
  - Some applications including Google Earth and Flickr provide interfaces for users to specify a location on the world map.
  - People can record their locations by GPS functions in their smart phones.

Motivation (Cont.)

- Examples of GPS-associated documents
  - Flickr: geo-tagged photos



  - Twitter: tweets from iPhone

Motivation (Cont.)

- What can we do?
  - By analyzing the geographical distribution of food and festivals, we can compare the cultural differences around the world.
  - We can also explore the hot topics regarding the candidates in presidential election in different places.
  - We can compare the popularity of specific products in different regions and help make the marketing strategy.

Motivation (Cont.)

- Discovering different topics of interests that are coherent in geographical regions.
- Comparing several topics across different geographical locations.
- Geographical topic discovery and comparison

Problem Formulation

- A **GPS-associated document** is a text document associated with a GPS location.
- A **geographical topic** is a spatially coherent theme. In other words, the words that are often close in space are clustered in a topic.
- An example of geographical topics
  - Given a collection of geo-tagged photos related to festival with tags and locations in Flickr, the desired geographical topics are the festivals in different areas, such as Cherry Blossom Festival in Washington DC and South by Southwest Festival in Austin, etc.

TEXT

GPS

Problem Formulation (Cont.)

- Given a collection of GPS-associated documents ➡ **Input**
  - Discover the geographical topics ➡ **Task I**
  - Compare the topics in different geographical locations. ➡ **Task II**

Problem Formulation (Cont.)

- An example of geographical topic discovery and comparison
  - Given a collection of geo-tagged photos related to food with tags and locations in Flickr, we would like to discover the geographical topics, i.e., what people eat in different areas. After we discover the food preferences, we would like to compare the food preference distributions in different geographical locations.

Problem Formulation (Cont.)

- A **topic distribution in geographical location** is the distribution of the topics given a specific location.
  - Formally, $p(z|l)$ is the probability of topic z given location $l = (x, y)$ where x is longitude and y is latitude.

Geographical Topic Discovery and Comparison

- Given a collection of GPS-associated documents *D* and the number of topics *K*, we would like to discover *K* geographical topics, i.e., $\theta = \{\theta_z\}_{z \in Z}$ where *Z* is the topic set and a geographical topic *z* is represented by a word distribution

$$\theta_z = \{p(w|z)\}_{w \in V} \quad \text{s.t.} \quad \sum_{w \in V} p(w|z) = 1 \ .$$

- Along with the discovered geographical topics, we also would like to know the topic distribution in different geographical locations for topic comparison, i.e., $p(z|l)$ for all $z \in Z$ in location *l*.

Solution

- Location-Driven Model (**LDM**)
- Text-Driven Model (**TDM**)
- Location-Text Joint Model (Latent Geographical Topic Analysis (**LGTA**))

Location-Driven Model (LDM)

- LDM
  - Clustering based on document locations
  - One location clustering is a topic
  - Generate topic description for each cluster
- Disadvantage
  - No text guidance
  - It is possible that there is no spatial cluster patterns. A geographical topic may be from several different areas and these areas may not be close to each other.
    - In landscape dataset, mountains exists in different areas and these areas are not close to each other

Text-Driven Model (TDM)

- Discover the geographical topics using topic modeling
  - Topic modeling with network regularization [Mei et al. WWW'08]
  - Regularization based on the closeness in location between documents

$$L(D) = -(1-\lambda)\sum_{d\in D}\sum_{w\in V} c(w,d)\log\sum_{z\in Z} p(w|z)p(z|d) + \frac{\lambda}{2}\sum_{(u,v)\in E} w(u,v)\sum_{z\in Z}(p(z|d_u)-p(z|d_v))^2$$

- Disadvantage
  - How to define the document closeness *w(u, v)*?
  - How to have the topic distribution of locations *p(z | l)*?

LOCATION-TEXT JOINT MODEL

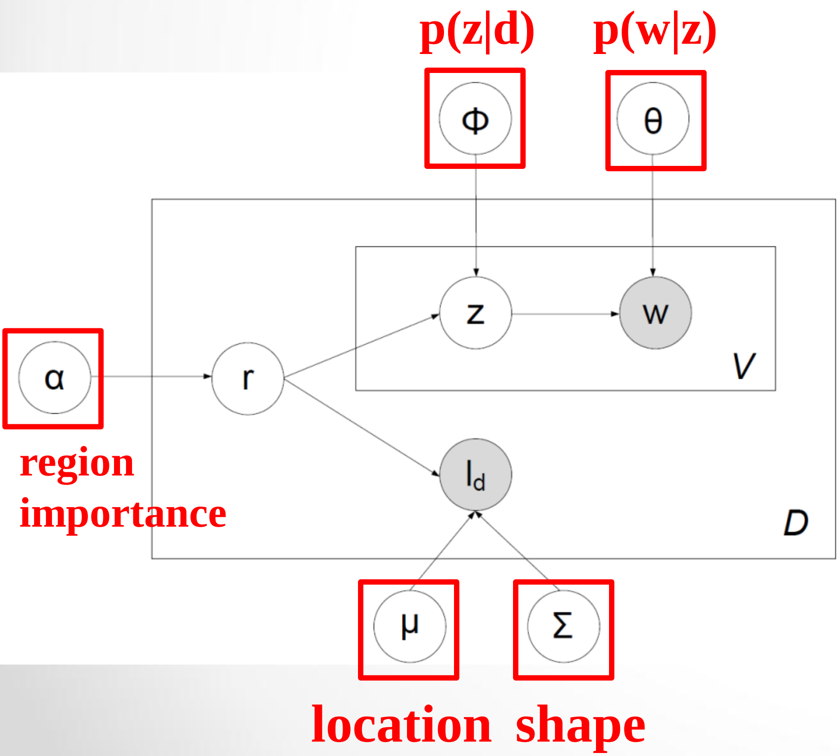
- **Main Insight:** Construct a model to encode the spatial structure of words
  - The words that are close in space are likely to be clustered into the same geographical topic.
- Assume there are a set of **regions**. The topics are generated from regions instead of documents.
  - If two words are close to each other in space, they are more likely to belong to the same region.
  - If two words are from the same region, they are more likely to be clustered into the same topic.

Regions ⟷ Words

(Pseudo-documents)

Latent Geographical Topic Analysis (LGTA)

- Combine text and location information
- Adapts the region discovery process according to the dataset.



To generate a geographical document d in collection D:

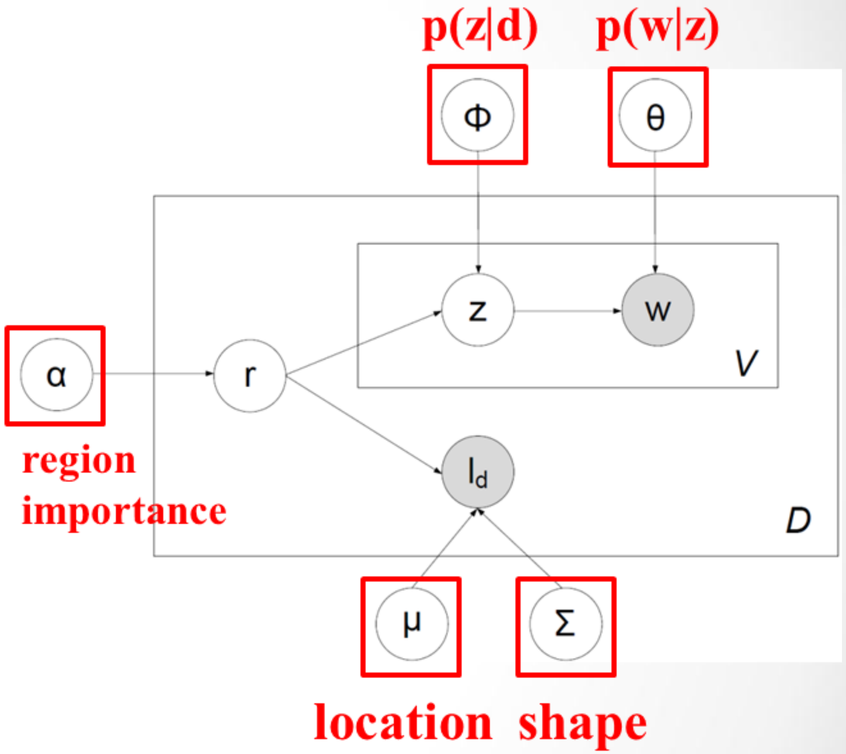
1. Sample a region r from the discrete distribution of region importance α, $r \sim Discrete(\alpha)$.
2. Sample location l_d from Gaussian distribution of μ_r and Σ_r.

$$p(l_d|\mu_r, \Sigma_r) = \frac{1}{2\pi\sqrt{|\Sigma_r|}} \exp\left(\frac{-(l_d - \mu_r)^T \Sigma_r^{-1} (l_d - \mu_r)}{2}\right)$$

3. To generate each word in document d:
   - (a) Sample a topic z from multinomial ϕ_r.
   - (b) Sample a word w from multinomial θ_z.

Parameter Estimation

- EM algorithm
- Iterations:
  - Geo-clustering (region discovery) is based on both location and topic information.
  - Topic modeling is based on the text and region information.



Data Set

- Flickr images with GPS locations
  - Flickr API supports search criteria including tag, time, GPS range, etc.

Data set	Time span	# image	# words
Landscape	09/01/09 - 09/01/10	5791	1143
Activity	09/01/09 - 09/01/10	1931	408
Manhattan	09/01/09 - 09/01/10	28922	868
Festival	09/01/09 - 09/01/10	1751	421
National Park	09/01/09 - 09/01/10	2384	351
Car	01/01/06 - 09/01/10	34707	12
Food	01/01/06 - 09/01/10	151747	278

Compared Methods

- LDM: Location-driven model
- TDM: Text-driven model
- GeoFolk [Sizov WSDM'10]:
  - A topic modeling method that uses both text and spatial information.
  - Model each region as an isolated topic
  - Assume the geographical distribution of each topic is Gaussian
- LGTA: Latent Geographical Topic Analysis

Topic Discovery Comparison

- Festival dataset
  - Topics related to South By Southwest Festival

TDM	GeoFolk	LGTA
sxsw 0.124	sxsw 0.173	sxsw 0.163
brooklyn 0.082	austin 0.136	austin 0.149
southbysouthwest 0.061	southbysouthwest 0.127	texas 0.142
south 0.055	texas 0.125	southbysouthwest 0.085
streetfestival 0.050	south 0.121	south 0.070
southwest 0.049	southwest 0.103	funfunfunfest 0.061
funfunfunfest 0.044	downtown 0.093	southwest 0.060
atlanticavenue 0.044	musicfestival 0.074	musicfestival 0.057
atlanticantic 0.041	live 0.034	downtown 0.040
streetfair 0.040	stage 0.010	music 0.034

Topic Discovery Comparison

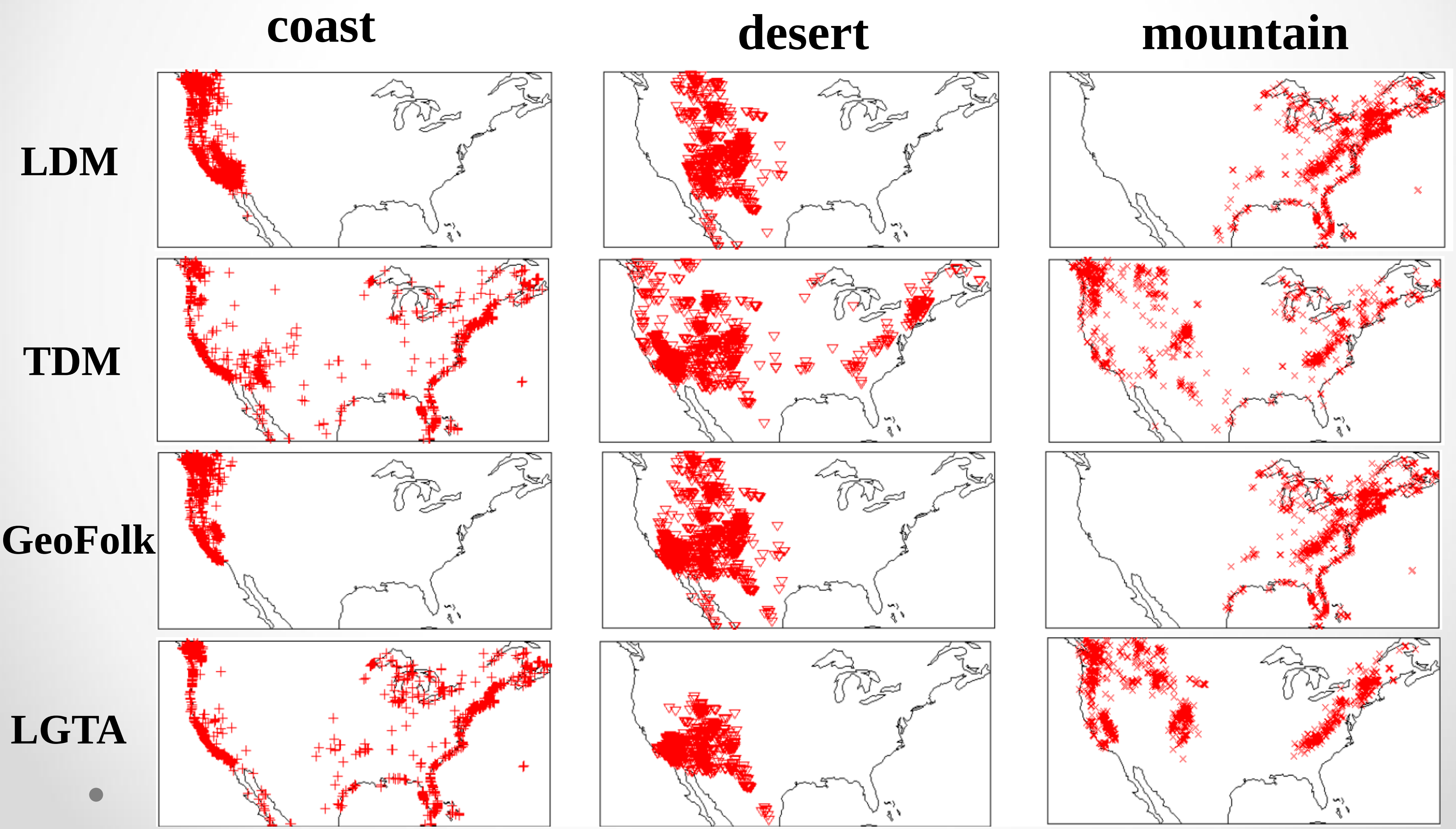
- Activity dataset

GeoFolk		LGTA	
Topic 1	Topic 2	Topic 1(surfing)	Topic 2(hiking)
hiking 0.077	hiking 0.095	surfing 0.070	hiking 0.109
mountains 0.037	mountains 0.050	beach 0.065	mountains 0.059
mountain 0.027	mountain 0.041	california 0.059	mountain 0.042
california 0.027	surfing 0.032	ocean 0.053	nature 0.027
surfing 0.024	beach 0.030	surf 0.031	trail 0.019
beach 0.023	[nh] 0.029	hiking 0.031	hike 0.017
nature 0.020	white[mtn]s 0.022	waves 0.028	desert 0.017
ocean 0.019	trail 0.021	water 0.025	washington 0.014
trail 0.015	ocean 0.021	surfer 0.022	lake 0.013
hike 0.015	nature 0.019	pacific 0.018	camping 0.013

*[mtn] is mountain. [nh] is newhampshire.

Topic Discovery Comparison

- Landscape dataset



Topic Quality Qualitative Comparison

- Average distance of word distributions of all pairs of topics by KL-divergence

Data set	LDM	TDM	GeoFolk	LGTA
Landscape	0.159	**0.311**	0.141	0.281
Activity	0.164	0.402	0.164	**0.491**
Manhattan	0.908	**1.091**	0.965	1.020
National Park	2.576	2.325	2.474	**2.598**
Festival	2.206	2.109	2.080	**2.258**
Car	2.518	**3.745**	2.365	3.731

Topic Quality Qualitative Comparison

- Text Perplexity

$$perplexity_{text}(D_{test}) = exp\{-\frac{\sum_{d \in D_{test}} \log p(\mathbf{w}_d)}{\sum_{d \in D_{test}} N_d}\}$$

Data set	LDM	TDM	GeoFolk	LGTA
Landscape	394.680	444.676	384.411	**366.546**
Activity	184.970	176.234	184.979	**157.775**
Manhattan	193.823	201.042	193.001	**192.010**
National Park	118.159	120.100	117.238	**117.077**
Festival	177.978	214.975	173.621	**170.033**
Car	9.936	9.926	9.937	**9.924**

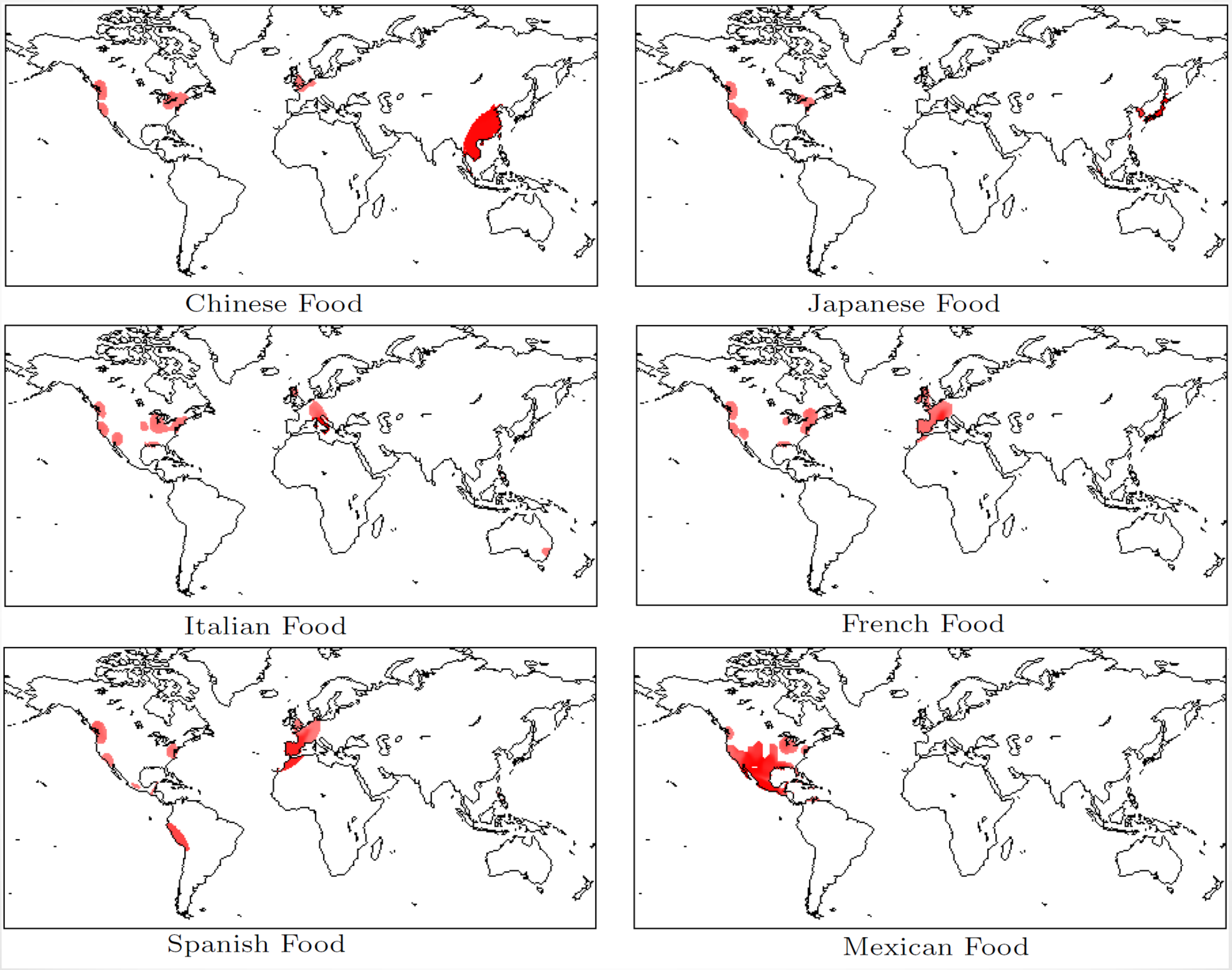
Topic Quality Qualitative Comparison

- Location/Text Perplexity

$$perplexity_{location/text}(D_{test}) = exp\{-\frac{\sum_{d \in D_{test}} \log p(\mathbf{w}_d, l_d)}{\sum_{d \in D_{test}} N_d}\}$$

Data set	LDM	GeoFolk	LGTA
Landscape	688.628	672.967	**569.047**
Activity	358.559	358.577	**257.086**
Manhattan	109.103	107.620	**105.684**
National Park	136.435	112.973	**103.853**
Festival	99.308	94.604	**91.230**
Car	40242.767	40348.974	**8718.927**

Geographical Topic Comparison



Thanks!
Questions?

- Complicated model and parameter estimation
- How to set the number of regions and the number of topics?

- How about estimating geographical locations for images that are without geo information?
    - Generating representative photos for the landmarks