Distant-supervised Heterogeneous multitask learning

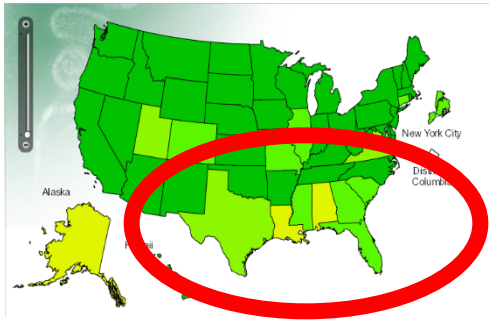
for social event forecasting with multilingual indicators

Liang Zhao

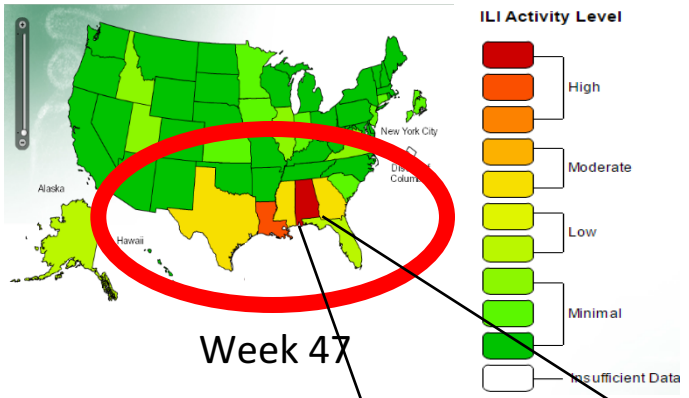
George Mason University

What are Spatiotemporal Events?

Week 45



Week 46



Week 47

Epidemics outbreak on Week 47 ending Nov 22, 2014 in southern region

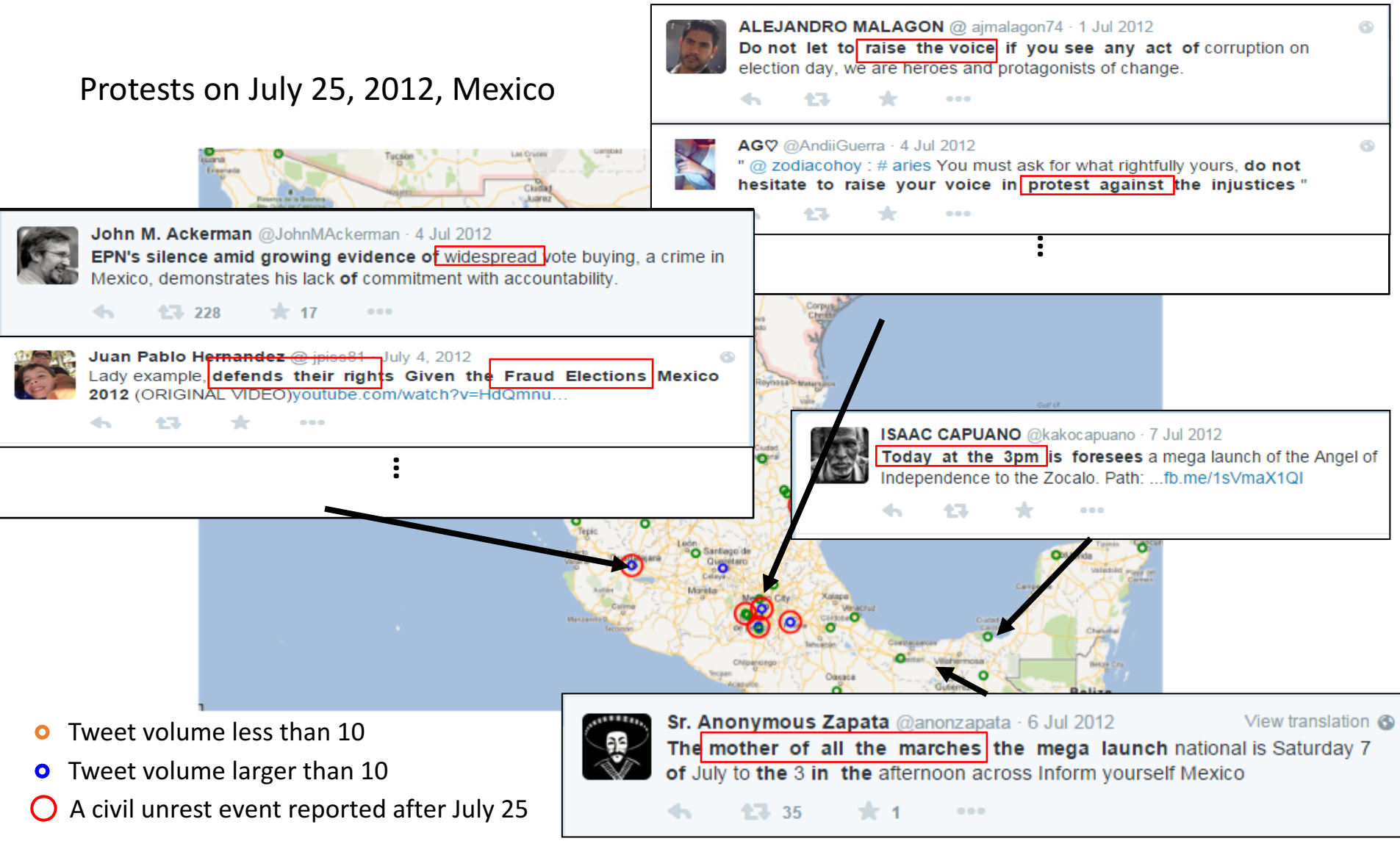
influenza

Civil unrest events on Mar 17, 2013 in Brazil

Protests

Open Source Indicators as the Social Sensor

Protests on July 25, 2012, Mexico

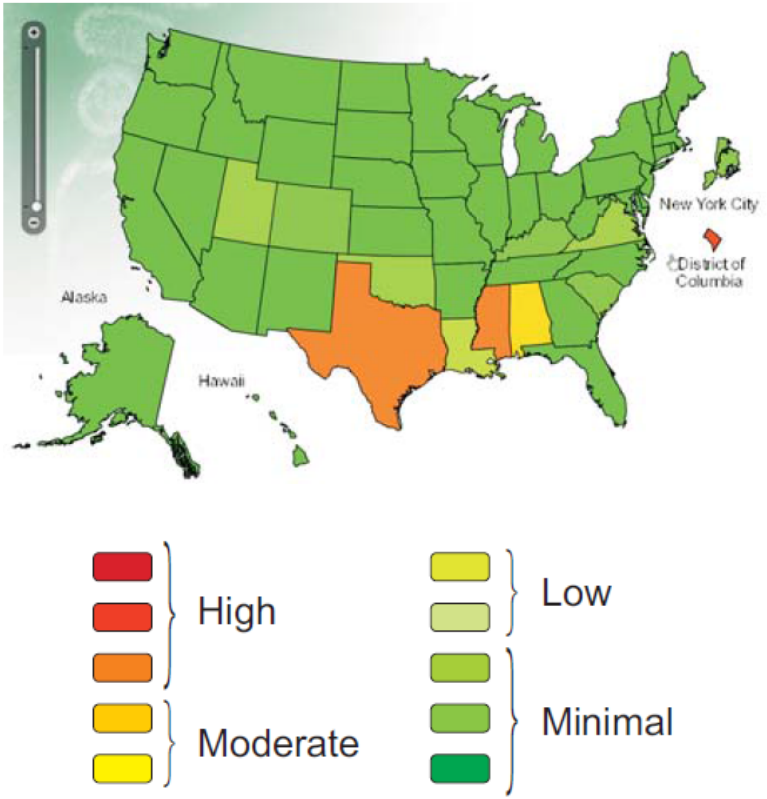


- Tweet volume less than 10
- Tweet volume larger than 10
- A civil unrest event reported after July 25

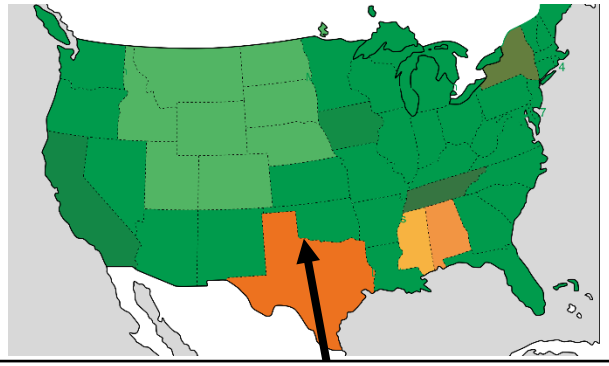
Open Source Indicators as the Social Sensor

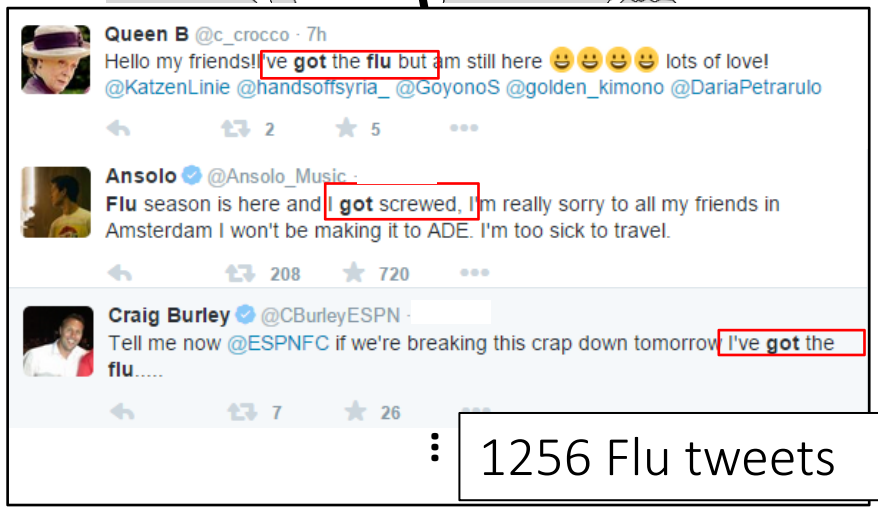
2013-14 Influenza Season Week 46

CDC flu activity map
(reported on Week 47)



Flu tweets geographical distribution
(reported on Week 46)





1256 Flu tweets

Challenge 1: Multilingual features

1. Must consider multilingual, because

Dataset	#Tweets	SPA (%)	ENG (%)	POR (%)
Argentina	160,564,890	91.6	7.3	1.1
Brazil	185,286,958	10.1	16.0	73.9
Chile	97,781,414	82.8	16.4	0.8
Colombia	158,332,002	89.8	9.4	0.8
Ecuador	50,289,195	91.1	8.1	0.8
El Salvador	21,992,962	91.5	7.8	0.7
Mexico	197,550,208	83.7	15.4	0.9
Paraguay	30,891,602	92.2	6.4	1.4
Uruguay	10,310,514	89.7	8.8	1.4
Venezuela	167,411,358	92.3	6.9	0.8

Moreover..

- Countries with hundreds of languages
- Omit a language → omit a group of people
- Cannot omit, even small ones

Social events can be triggered by any people

2. Too large dimension, too sparse feature vector

Imagine a feature vector of a tweet: 10 nonzeros with 1M zeros..

3. Few data for small language

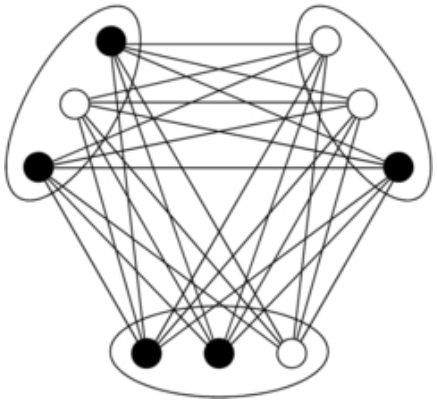
Challenge 2: Cross-lingual semantic correlation

1. Features are highly semantically redundant

One feature

(https://www.profluentplus.com/blog/)

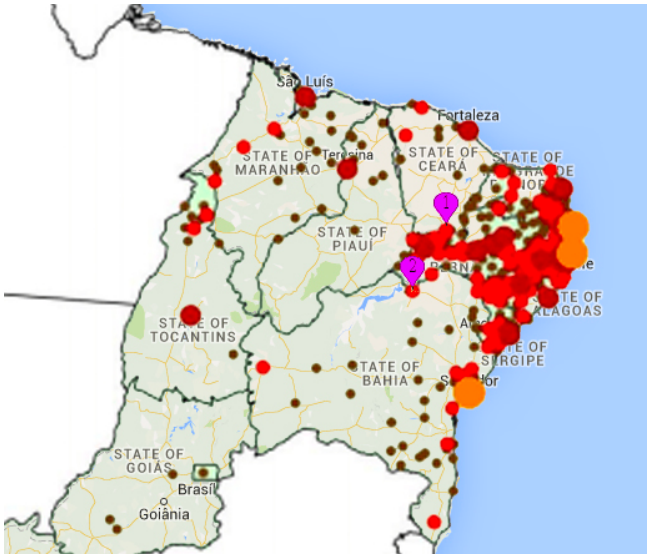
2. Features are correlated via multi-partite relationship



(http://www.writeopinions.com/complete-multipartite-graph)

Challenge 3: Lack of language-wise supervision

Zika outbreaks in Brazil



(http://blogs.discovermagazine.com/science-sushi/2016/01/31/genetically-modified-mosquitoes-didnt-start-zika-ourbreak/)

No label on how much each group of language-speakers contribute

Mass occupation underscores Brazil's poverty, creates angst

By MAURICIO SAVARESE | Associated Press

Image 1 of 2

SAO BERNARDO DO CAMPO, Brazil – Luciano Oliveira, a bricklayer, gazes at the floor of his tiny wood shack, which is one of thousands of makeshift settlements that comprise a massive squat in this suburb of Sao Paulo.

Oliveira was fired from his job at a restaurant a few months ago, shortly after arriving from the northeastern state of Bahia.

More from Fox News

Amy Schumer on sexual misconduct...
Fox News Entertainment

Prince Harry's billionaire friend...
Fox News Entertainment

Evidence that young victims endured...
Fox News US

'Dukes of Hazzard's' John Schneider...
Fox News Entertainment

Kathy Griffin talks life after Trump...
Fox News Entertainment

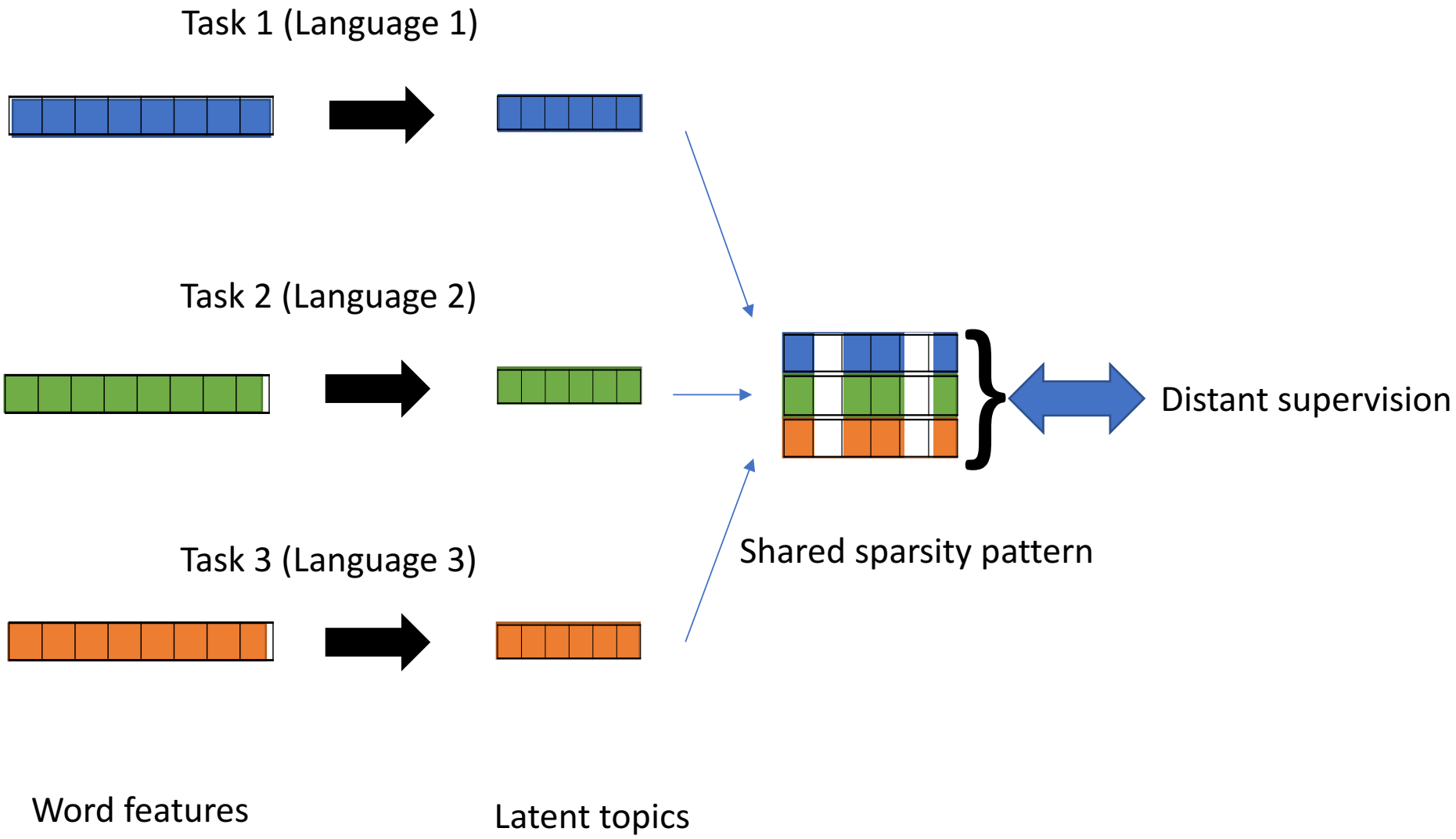
Report: Russian pilot captured, killed
Fox News World - Video

Russia retaliates after pilot is killed...
Fox News World - Video

Democrat hits Pelosi over 'make...
Fox News Politics

(http://www.foxnews.com/world/2017/12/18/mass-occupation-underscores-brazils-poverty-creates-angst.html)

Heterogeneous Multitask learning under distant supervision



Objective function

Distant supervision: any language triggers, the whole triggers
none language triggers, the whole not triggers

Higher-level topic representation
and transition matrix

$$\min_{U_l,\Theta_l\geq 0,l\in L} \frac{1}{|S||T|}\sum_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l^T\Theta_l X_{s,t,l}^T), Y_{s,\tau}) + \lambda_1\|U\|_{2,1} + \lambda_2\sum_l^L\|\Theta_l\|_1 \quad s.t. \Theta_l\Theta_l^T = I,\ l\in L$$

Shared sparsity patterns of
latent topics in different tasks

Orthogonal constraint

Upper-bounded generalization error:

$$\begin{aligned}
&\mathbb{E}(\Theta^*_{(M)}, U^*_{(M)}) - \mathbb{E}(\Theta^*, U^*) \\
&= \mathbb{E}_{M\sim\mu}[\frac{1}{|S||T|}\sum_{s,t}^{S,T}\mathcal{L}(\max_l F_l([U^*_{(M)}]_l^T[\Theta^*_{(M)}]_l X^T_{s,t,l}), Y_{s,\tau})] \\
&- \inf_{\Theta\in\mathcal{F}_2, U\in\mathcal{F}_1}\mathbb{E}_{M\sim\mu}[\frac{1}{|S||T|}\sum_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l^T\Theta_l X^T_{s,t,l}), Y_{s,\tau})] \\
&\leq 2C\alpha\sqrt{\frac{2\mathcal{C}_1(X)|L|(d+12)}{|S|\cdot|T|}} + 2C|L|\alpha\sqrt{\frac{8\mathcal{C}_\infty(X)\ln(2d)}{|S|\cdot|T|}} + 2\sqrt{\frac{2\ln 2/\epsilon}{|S||T|}}
\end{aligned}$$

Optimization

Equivalent problem:

$$\min_{\Theta \geq 0, U, Z, Q} \frac{1}{|S||T|} \sum_{s,t}^{S,T} \mathcal{L}(f(Z_{s,t}), Y_{s,\tau}) + \lambda_1 \|U\|_{2,1} + \lambda_2 \sum_{l}^{L} \|\Theta_l\|_1$$

$$s.t. \Theta_l \Theta_l^T = I,\ Z_{s,t} = \max_i Q_{s,t,i},$$

$$Q_{s,t,l} = U_l^T \Theta_l X_{s,t,l}^T,\ s \in S, t \in T, l \in L$$

Alternating Direction Methods of Multipliers (ADMM)

Solve Q: Dynamic programming

Solve Θ and U: non-monotone spectral projected gradient descent

Solve Z: second-order methods

Experiments

	AR	BR	CL	CO	EC	EL	MX	PY	UY	VE
LogReg	0.594	0.686	0.677	0.644	0.599	0.618	0.661	0.6162	0.628	0.667
LASSO	0.596	0.685	0.677	0.648	0.603	0.636	0.665	0.6151	0.666	0.669
MTL	**0.733**	0.722	0.669	0.810	0.617	0.772	**0.795**	0.600	0.811	0.771
MREF	0.706	0.714	0.563	0.515	0.784	0.612	0.693	0.658	0.6812	0.588
DHML	0.704	**0.845**	**0.683**	**0.846**	**0.839**	**0.780**	0.793	**0.737**	**0.835**	**0.835**

Languages	Topics	Keywords							
Spanish	Topic 8	conflict	farmer	rancher	pit	insecurity	agrarian	whistle	deforest
	Topic 5	picket	mobilize	arrest	cooperative	impoverish	Zapatista	#cnte	upsurge
	Topic 1	class	criminalize	suppress	riot	moderate	barricade	protest	teacher
	Topic 9	#cgtp	#cofecay	#snte	@eloisago.	@chertor.	@dionisio.	@morenaj.	@unt_mx
	Topic 10	power	match	Energy	warn	food	town	defending	torture
English	Topic 7	agency	community	maintain	charge	reform	discuss	loss	legalize
	Topic 1	smash	effort	proposal	invitation	arm	university	class	fight
	Topic 5	medium	report	popular	paralyze	tax	affect	danger	payment
	Topic 4	gringo	crime	investment	attack	capture	victim	protagonist	boycott
	Topic 6	mandate	striker	confront	assembly	parliament	mandatory	freedom	parade
Portuguese	Topic 2	person	president	time	class	opportunity	deputy	election	alternative
	Topic 4	ambush	plunder	warn	police	gun	convention	agreement	officer
	Topic 6	lead	national	together	change	authority	congress	labor	violence
	Topic 5	pocket	mine	shot	catch	criminal	control	enemy	upsurge
	Topic 8	hunger	hassle	fire	treatment	defeat	Medical	groom	root