Topic modeling of social entrepreneur: A work note

Cheng-Jun Wang

2013 Sep 4

Outline

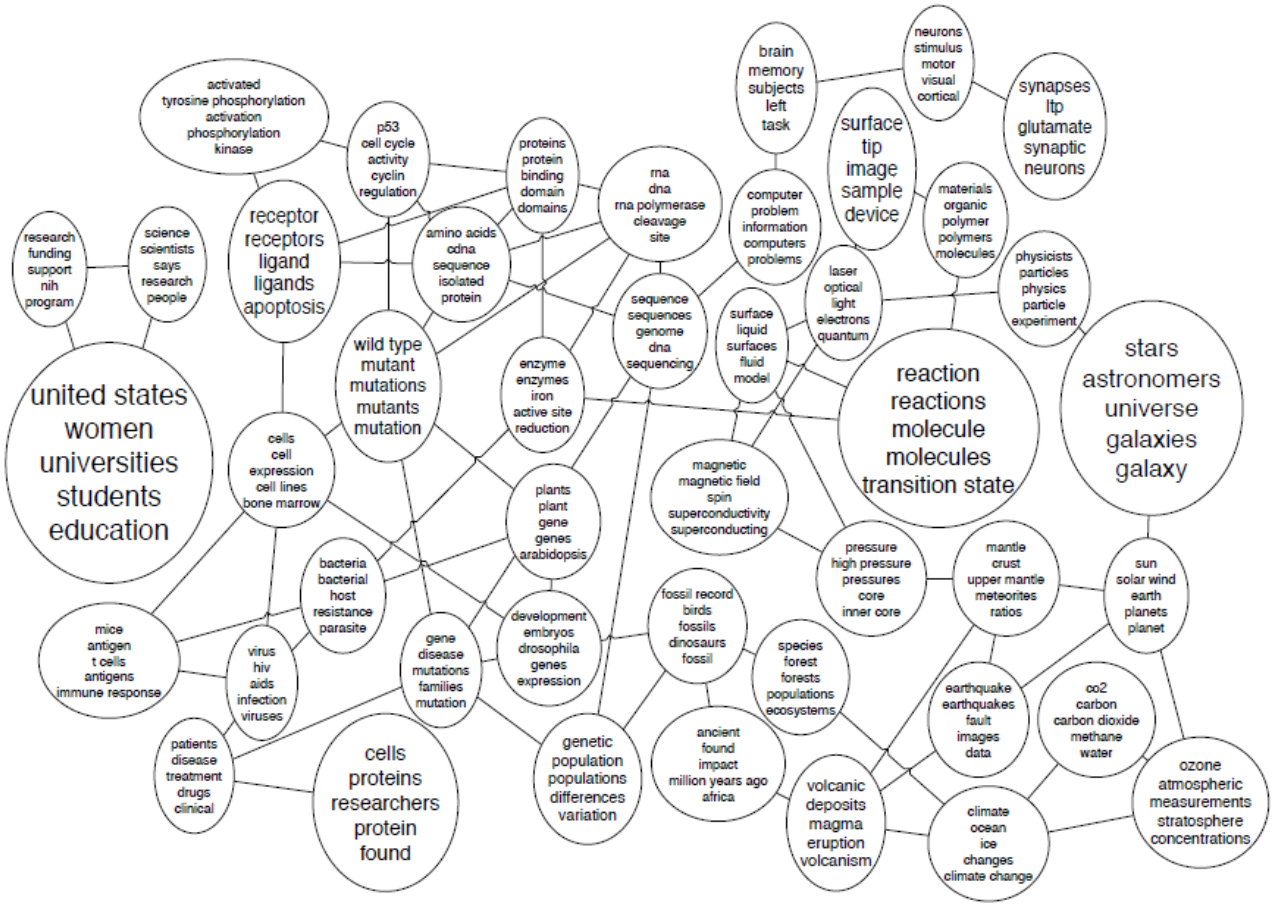
- Basic knowledge of TM
- Procedures of TM
- Results:

✓ Part 1

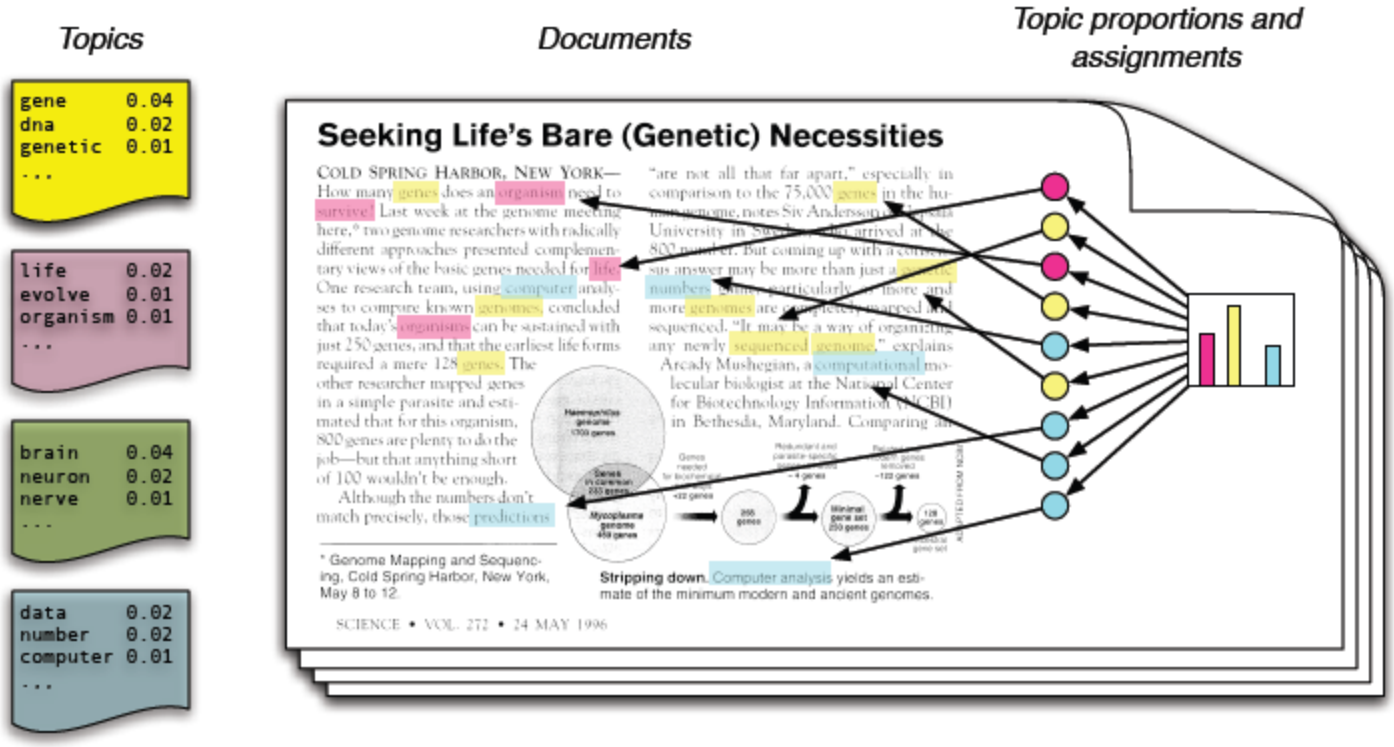
✓ Part 2

Background information

- LDA vs. CTM
- VEM vs. Gibbs sampling

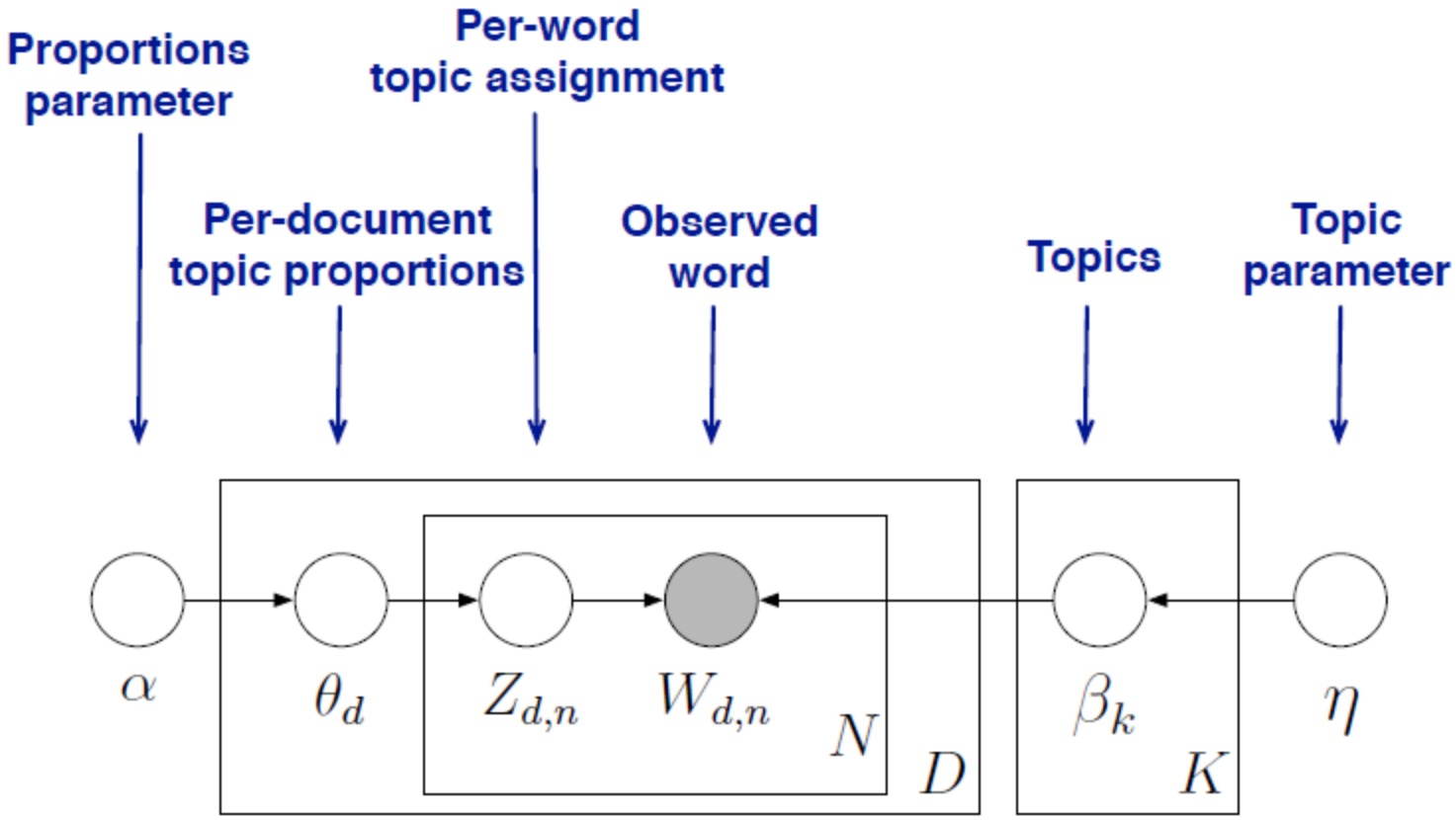


Latent Dirichlet allocation (LDA)



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

LDA as a graphical model



- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

- The posterior of the latent variables given the documents is

$$p(\beta, \theta, \mathbf{z} \mid \mathbf{w}).$$

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

- This joint defines a posterior.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, exploration, ...

Step 1: Data pre-processing

```
dtm <- DocumentTermMatrix(corpus,
        control =  list(
                stemming = TRUE,
                stopwords = TRUE,
                wordLengths=c(4, 15),
                bounds = list(global = c(5,Inf)),
                removeNumbers = TRUE,
                removePunctuation  = list(preserve_intra_word_dashes = FALSE),
                encoding = "UTF-8"
```

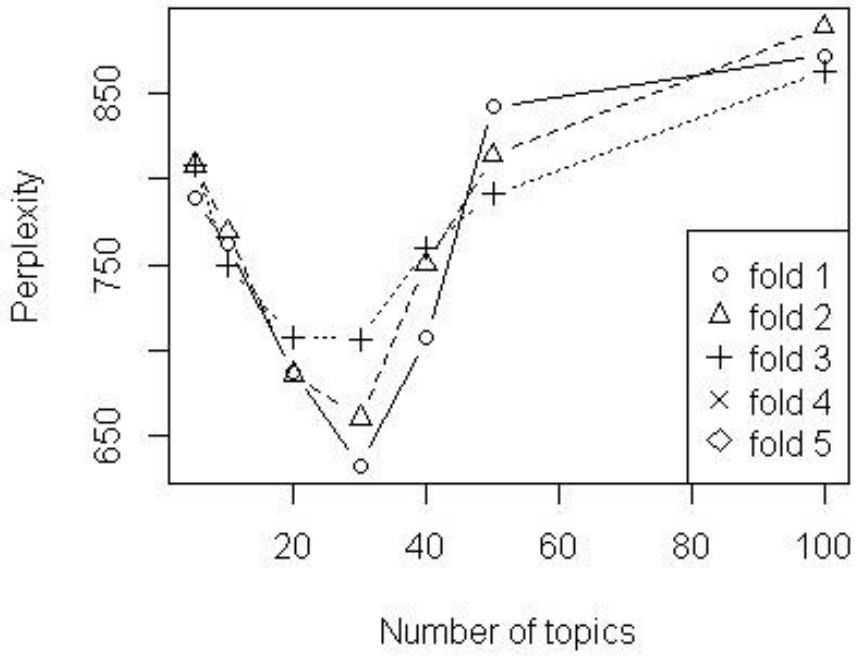
Step 2: Identify the number of topics by calculating the perplexity

- ✓ Using 10-fold cross-validation
- ✓ Train topic models with 90% data, and test with the other 10% data

- To model an unknown probability distribution *p*, based on a training sample that was drawn from *p*. Given a proposed probability model *q*, one may evaluate *q* by asking how well it predicts a separate test sample x_1, x_2, ..., x_N also drawn from *p*. The perplexity of the model *q* is defined as:

$$2^{-\sum_{i=1}^{N} \frac{1}{N} \log_2 q(x_i)}$$

- Better models q of the unknown distribution p will tend to assign higher probabilities $q(x_i)$ to the test events. Thus, they have lower perplexity: they are less surprised by the test sample.

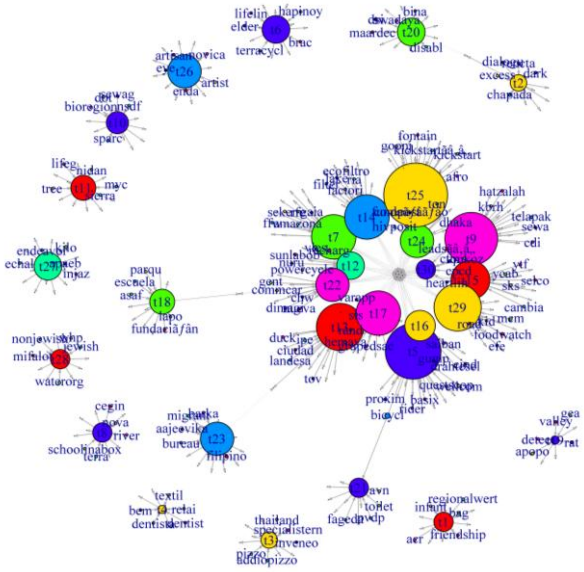


Step 3: Get the topic-term matrix and topic-document matrix

rs = posterior(jss_TM$CTM, dtm)

rs$topics # topics and documents

rs$terms # topic and terms



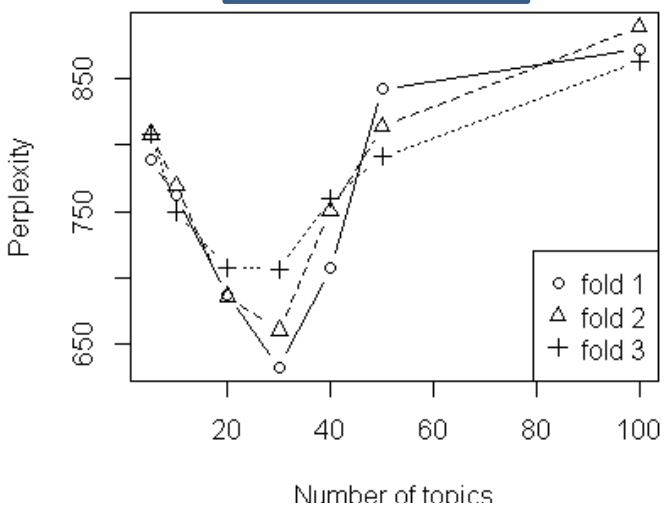
1	2	3	4	5
dna	protein	water	says	mantle
gene	cell	climate	researchers	high
sequence	cells	atmospheric	new	earth
genes	proteins	temperature	university	pressure
sequences	receptor	global	just	seismic
human	fig	surface	science	crust
genome	binding	ocean	like	temperature
genetic	activity	carbon	work	earths
analysis	activation	atmosphere	first	lower
two	kinase	changes	years	earthquakes

The Schwab Foundation for Social Entrepreneurship provides unparalleled platforms at the regional and global level to highlight and advance leading models of sustainable social innovation.

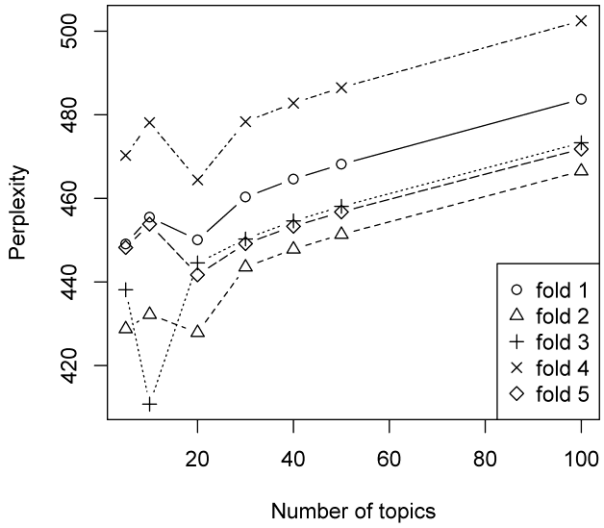
PART 1: SCHWAB

SCHWAB FOUNDATION FOR
SOCIAL ENTREPRENEURSHIP

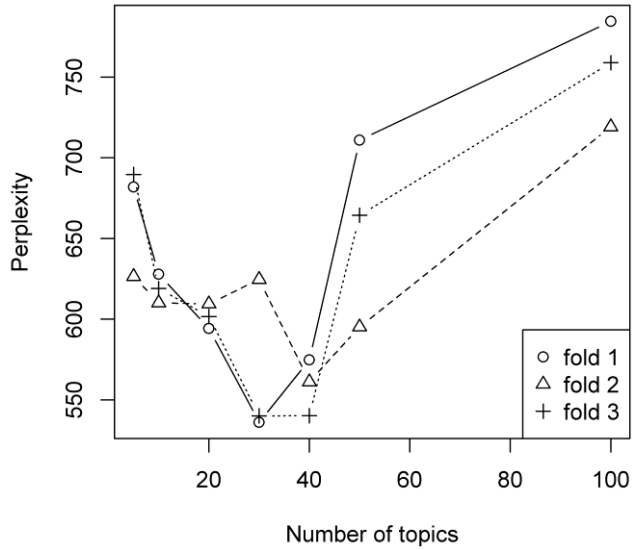
Background



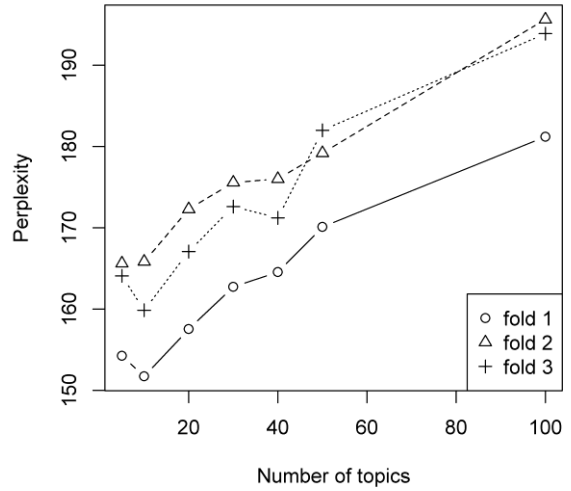
Entrepreneur



Innovation



Short Introduction



Background

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	rural	educ	disabl	servic	educ	train	develop	health	school	peopl
2	land	school	peopl	develop	student	construct	peopl	medic	children	vision
3	women	youth	activ	rural	river	afford	children	servic	educ	million
4	electr	student	servic	client	teacher	local	medic	patient	student	glass
5	famili	teacher	social	farmer	learn	villag	activ	provid	youth	develop
	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
1	poor	water	peopl	servic	provid	vehicl	sustain	work	work	entrepreneur
2	communiti	villag	disabl	worker	develop	bank	resourc	employ	artisan	econom
3	rural	children	access	social	communiti	world	work	product	mother	world
4	activ	provid	children	bank	million	organ	bioregion	social	famili	countri
5	servic	famili	book	migrant	activ	manag	live	peopl	hiv	rural
	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
1	product	hous	wast	communiti	women	social	communiti	sustain	develop	area
2	rural	poor	develop	health	use	famili	local	develop	innov	social
3	entrepreneur	live	activ	mani	equip	help	citi	organ	new	peopl
4	communiti	slum	organ	activ	activ	communiti	urban	activ	product	live
5	need	communiti	recycl	live	communiti	creat	develop	food	communiti	program

Entrepreneur

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	socil	yer	socil	tht	awrd	develop	work	awrd	socil	work
2	work	indigen	awrd	develop	socil	work	tht	socil	compni	jne
3	develop	togeth	work	educt	univers	project	yer	work	tht	mngement
4	progrmm	cquir	univers	work	work	school	found	peopl	yer	fellow
5	yer	first	develop	yer	communiti	yer	slum	citi	orgnizt	world
6	busi	socil	right	socil	degre	technolog	sinc	tht	provid	public
7	orgnizt	mni	foundtion	busi	ateneo	experi	interntionl	develop	build	posit
8	entrepreneur	right	found	children	mnil	environmentl	orgnizt	educt	serv	young
9	estblish	club	sekem	peopl	yer	engin	first	help	develop	busi
10	entrepreneurship	friend	recogn	school	institut	receiv	need	disbl	work	studi
	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
1	develop	busi	work	socil	school	sustinbl	tht	tht	univers	innovt
2	tht	univers	nigeri	ledership	develop	develop	awrd	univers	awrd	tech
3	mngement	develop	socil	peopl	educt	work	world	work	servic	communiti
4	awrd	indi	awrd	serv	work	univers	fellow	develop	serv	univers
5	indi	time	interntionl	work	univers	compni	helth	women	receiv	ceo
6	helth	socil	lso	globl	socil	one	interntionl	educt	tht	work
7	found	tht	specil	busi	fellow	new	univers	peopl	lern	amer
8	public	world	energi	sector	entrepreneurship	led	globl	degre	school	cofound
9	univers	engin	telpk	vision	first	oliveir	econom	afric	yer	provid
10	bngldesh	peopl	disbl	south	busi	nercessin	work	unit	microfinnc	wht

Innovation

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	develop	villag	busi	disabl	busi	artisan	develop	electr	sustain	wast
2	social	rural	pizzo	peopl	unit	work	servic	develop	project	collect
3	children	construct	regionalwe rt	employ	rat	artist	busi	famili	communiti	recycl
4	creat	product	region	provid	programm	one	also	communiti	carbon	compani
5	servic	train	pay	aid	cooper	addit	group	system	compani	manag
	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
1	educ	ytf	organ	peopl	women	communiti	innov	farmer	children	world
2	school	work	compani	work	includ	centr	health	support	communiti	toilet
3	develop	children	peopl	provid	social	local	mobil	communiti	project	sanit
4	student	academi	volunt	communiti	product	cooper	new	train	urban	global
5	land	communiti	product	employ	train	develop	open	local	busi	comic
	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
1	servic	program	care	health	servic	health	bank	social	communiti	youth
2	peopl	dot	provid	peopl	provid	commcar	provid	develop	product	peopl
3	young	local	eye	develop	loan	chw	servic	communiti	rural	offer
4	emerg	network	patient	bicycl	migrant	communiti	client	organ	hapinoy	market
5	provid	train	servic	healthcar	develop	improv	poor	manag	villag	environme nt
	Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
1	light	communiti	sustain	efe	educ	social	conserv	busi	health	social
2	solar	farmer	develop	educ	school	busi	environme nt	social	servic	develop
3	energi	econom	sekem	market	student	develop	ipe	manag	communiti	peopl
4	member	develop	project	peopl	teacher	communiti	lifeg	train	provid	provid
5	servic	organ	organ	chang	train	organ	sustain	communiti	work	communiti

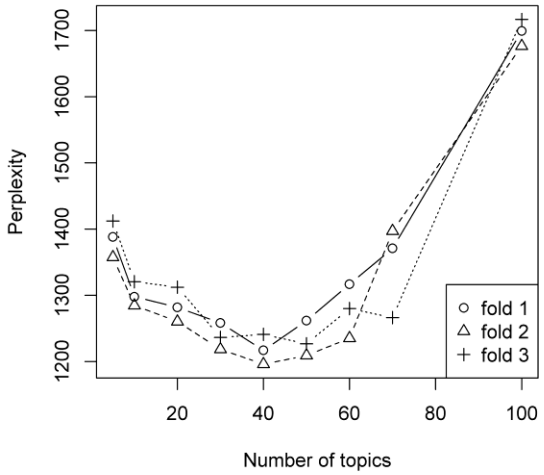
Short Introduction

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	social	develop	econom	poor	sustain
2	compani	social	organ	help	world
3	develop	support	educ	develop	busi
4	product	sustain	communiti	urban	live
5	farm	work	opportun	women	provid
6	peopl	educ	lowincom	promot	benefit
7	communiti	improv	develop	cultur	product
8	valu	help	social	work	condit
9	farmer	individu	nation	new	million
10	organ	econom	improv	rural	practic
	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	water	rural	health	provid	technolog
2	solar	villag	servic	servic	peopl
3	energi	servic	qualiti	train	promot
4	innov	provid	provid	job	develop
5	world	communiti	access	youth	use
6	area	afford	medic	busi	educ
7	develop	sustain	care	creat	programm
8	technolog	train	work	societi	improv
9	system	work	improv	programm	young
10	empow	develop	children	opportun	social

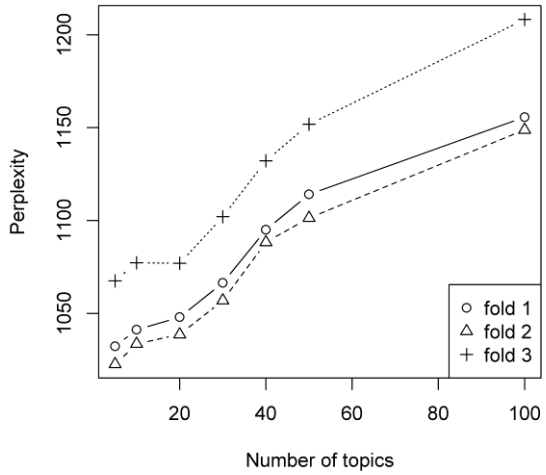
Ashoka is the largest network of social entrepreneurs worldwide, with nearly 3,000 Ashoka Fellows in 70 countries putting their system changing ideas into practice on a global scale.

PART 2: ASHOKA

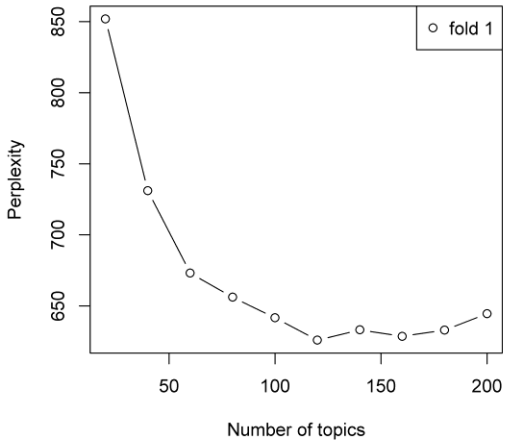
idea



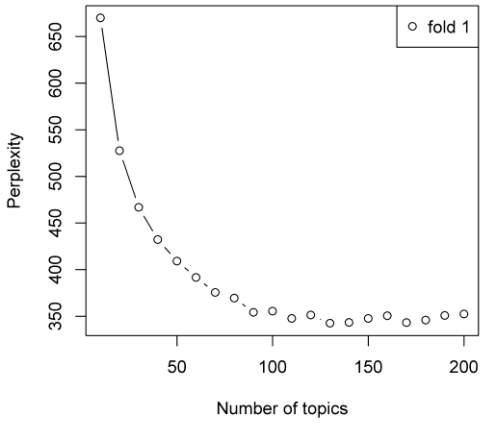
Intro



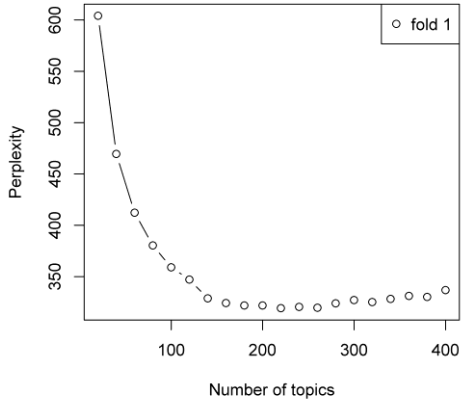
Person



Problem



Strategy



Intro

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	women	women	environment	busi	women
2	environment	center	disabl	communic	protect
3	human	conserv	women	produc	student
4	farmer	violenc	busi	learn	legal
5	worker	farmer	field	hous	land
6	foster	legal	prevent	univers	citi
7	mental	villag	livelihood	valu	environment
8	student	human	strategi	polit	center
9	industri	food	offer	teach	campaign
10	disabl	forest	technolog	technolog	altern

Strategy

Number of topics for Strategy = 180

Topic 1	women	michal	sport	girl	highway	sister	athlet	fenc	ambul	villag
Topic 2	forest	javier	campus	energi	amazon	clinic	wood	wildlif	speci	bandung
Topic 3	mike	mexican	user	emiss	carbon	mexico	bioga	fertil	nepal	csos
Topic 4	export	consum	mexico	coffe	union	bird	investor	breed	mari	grower
Topic 5	station	broadcast	veteran	victoria	afterschool	ciudadano	news	journalist	teen	weaver
Topic 6	medicin	mexico	camp	mauricio	sibl	clinic	treatment	healer	indigen	bengal
Topic 7	farmer	crop	farm	villag	plant	water	tree	seed	speci	irrig
Topic 8	dream	farm	villag	farmer	anim	rabbit	livestock	breed	clinic	agribusi
Topic 9	hospit	plant	medicin	tortur	soul	township	conserv	cell	amazon	anemia
Topic 10	migrant	burmes	thai	william	thailand	francisco	paul	journalist	solar	daniel

To do list