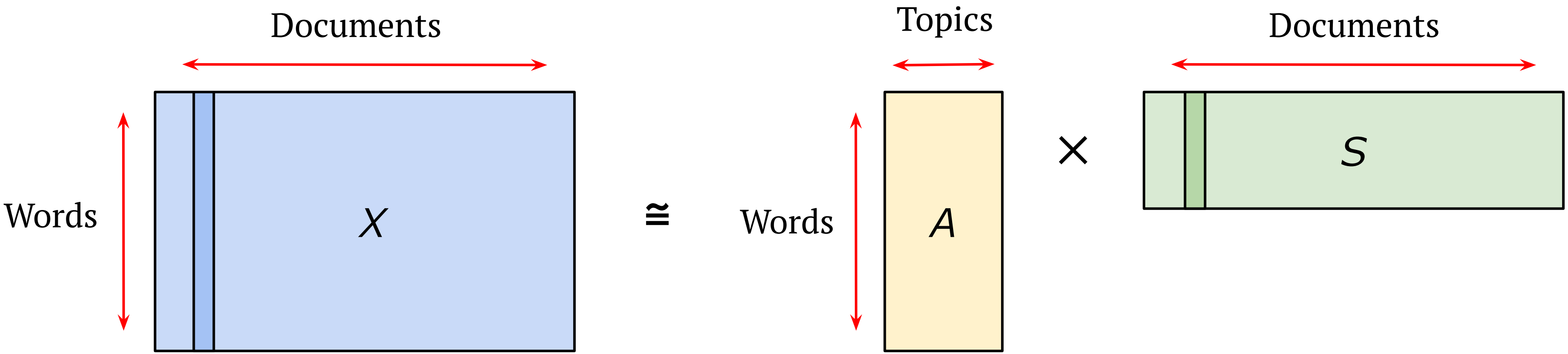
On a Guided Nonnegative Matrix Factorization

Joshua Vendrow, Jamie Haddock, Elizaveta Rebrova, Deanna Needell
UCLA Mathematics

Nonnegative Matrix Factorization (NMF)

Given a nonnegative matrix X (X ≥ 0), compute nonnegative A and S such that

$$X \approx AS$$

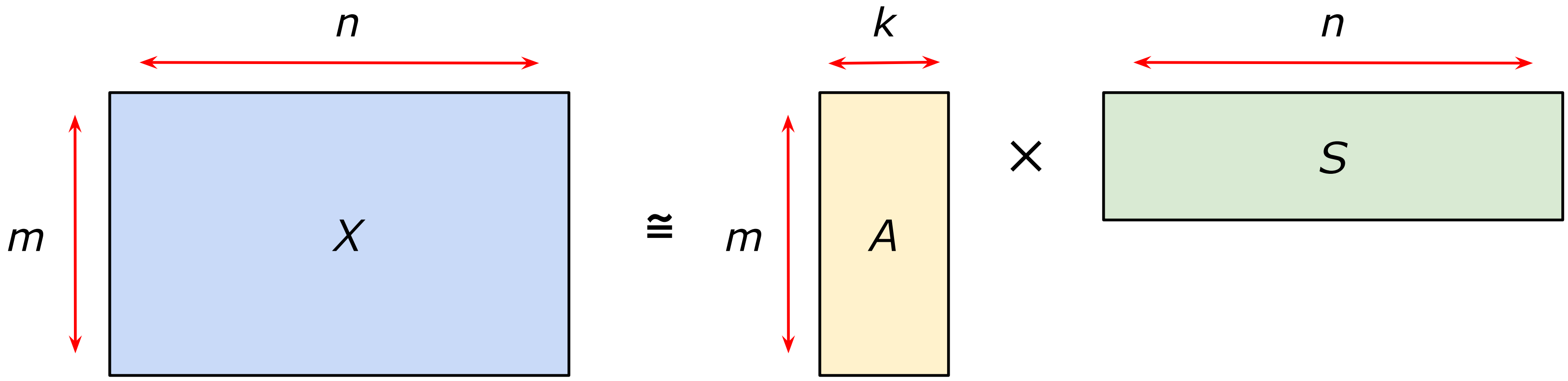


D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788, 1999.

Nonnegative Matrix Factorization (NMF)

Formulated as the optimization task:

$$\underset{\boldsymbol{A}\geq 0, \boldsymbol{S}\geq 0}{\arg\min} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|_F^2$$

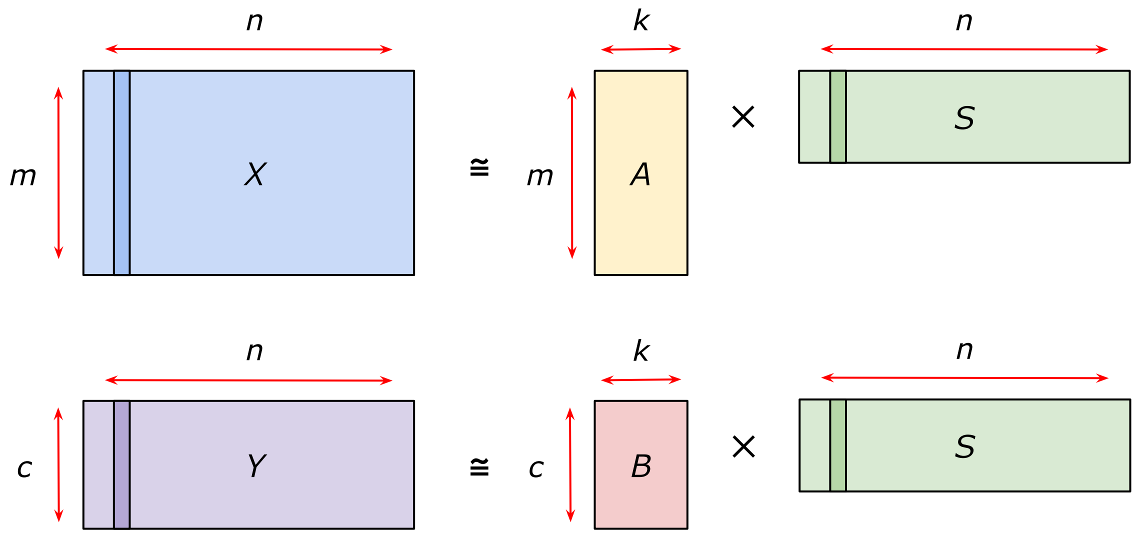


H. Lee, J. Yoo, and S. Choi, “Semi-supervised nonnegative matrix factorization,” IEEE Signal Proc. Let., vol. 17, no. 1, pp. 4–7, 2009.

Semi-Supervised Nonnegative Matrix Factorization (SSNMF)

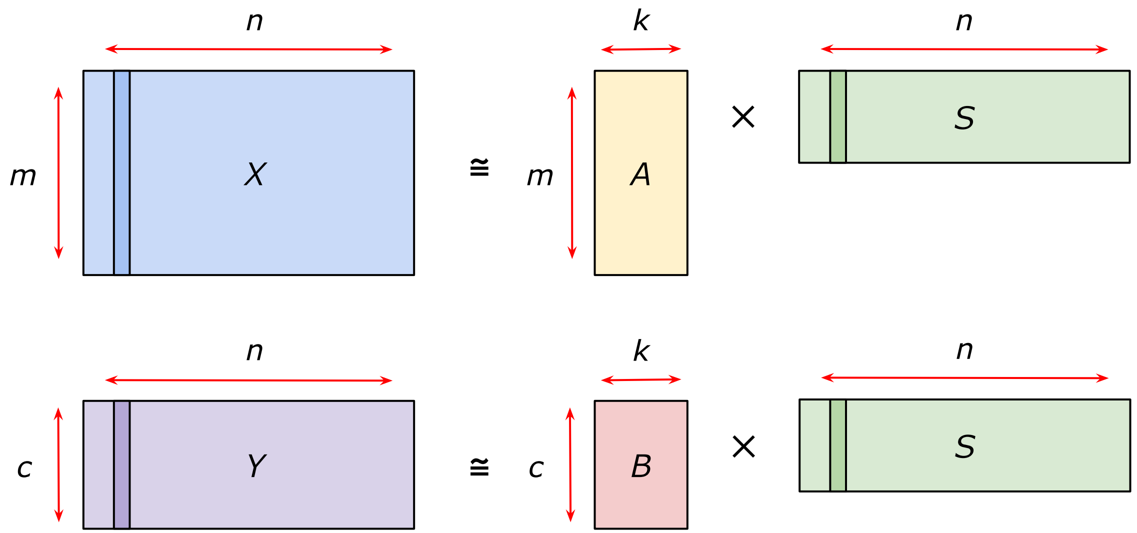
Given nonnegative matrix X and label matrix Y, compute nonnegative A, B, and S such that

$$X \approx AS,\, Y \approx BS$$

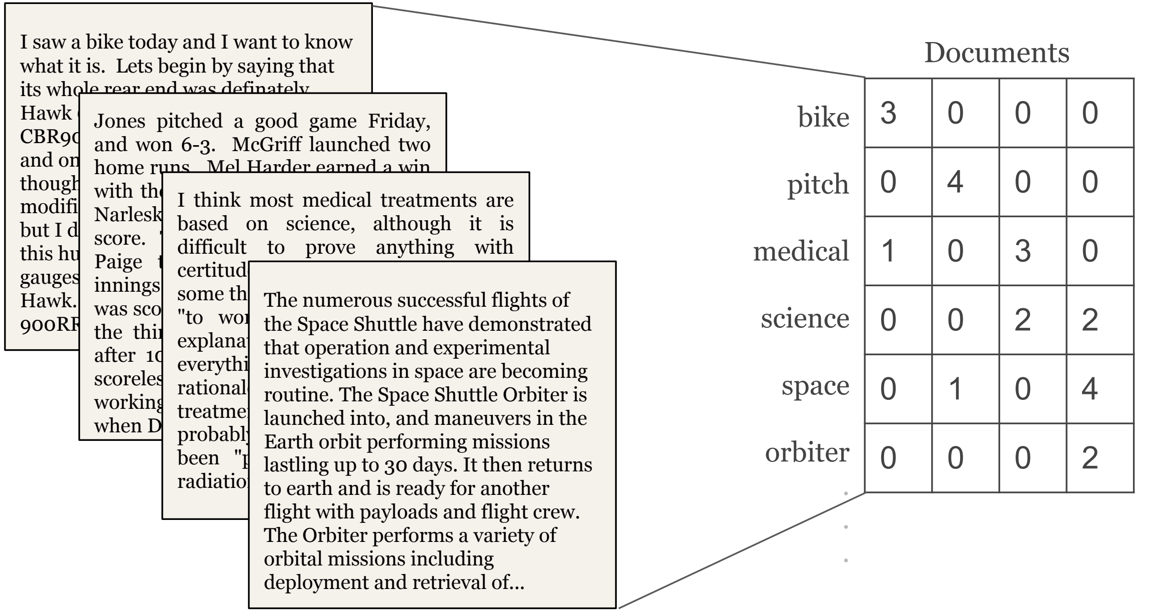


Semi-Supervised Nonnegative Matrix Factorization (SSNMF)

Formulated as the optimization task: $$\underset{\boldsymbol{A},\boldsymbol{S},\boldsymbol{B}\geq 0}{\arg\min}\ \underbrace{\|\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S}\|_F^2}_{\text{Reconstruction Error}} + \lambda \underbrace{\|\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{S}\|_F^2}_{\text{Classification Error}}$$



Bag of Words Representation



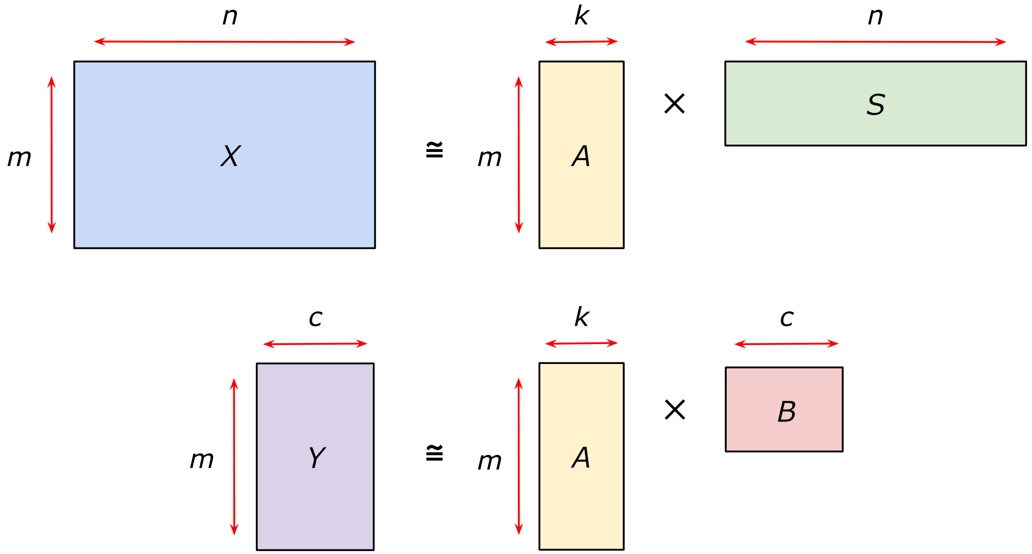
Documents

Term-Document Matrix

Guided Nonnegative Matrix Factorization (Guided NMF)

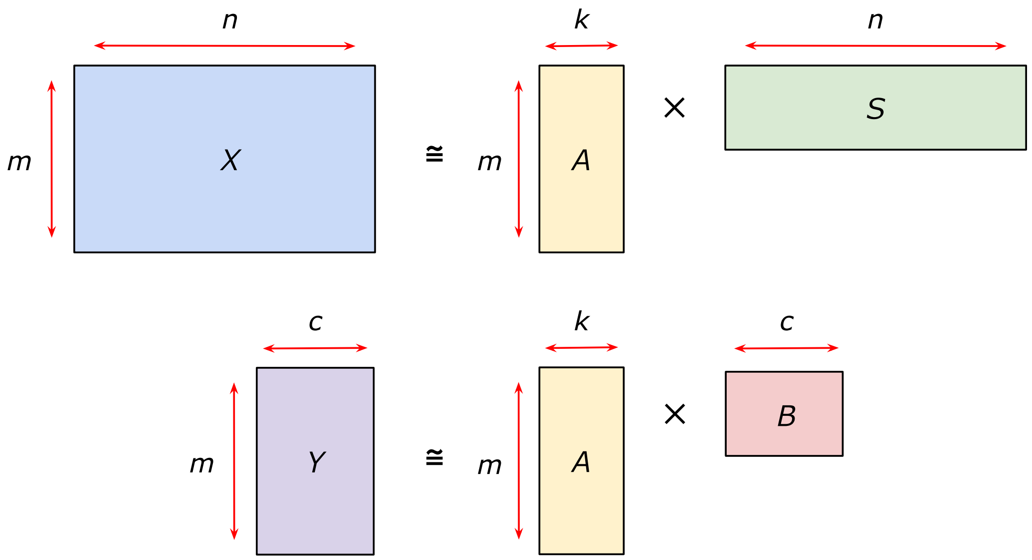
Given nonnegative matrix X and label matrix Y, compute nonnegative A, B, and S such that

$$X \approx AS,\ Y \approx AB$$



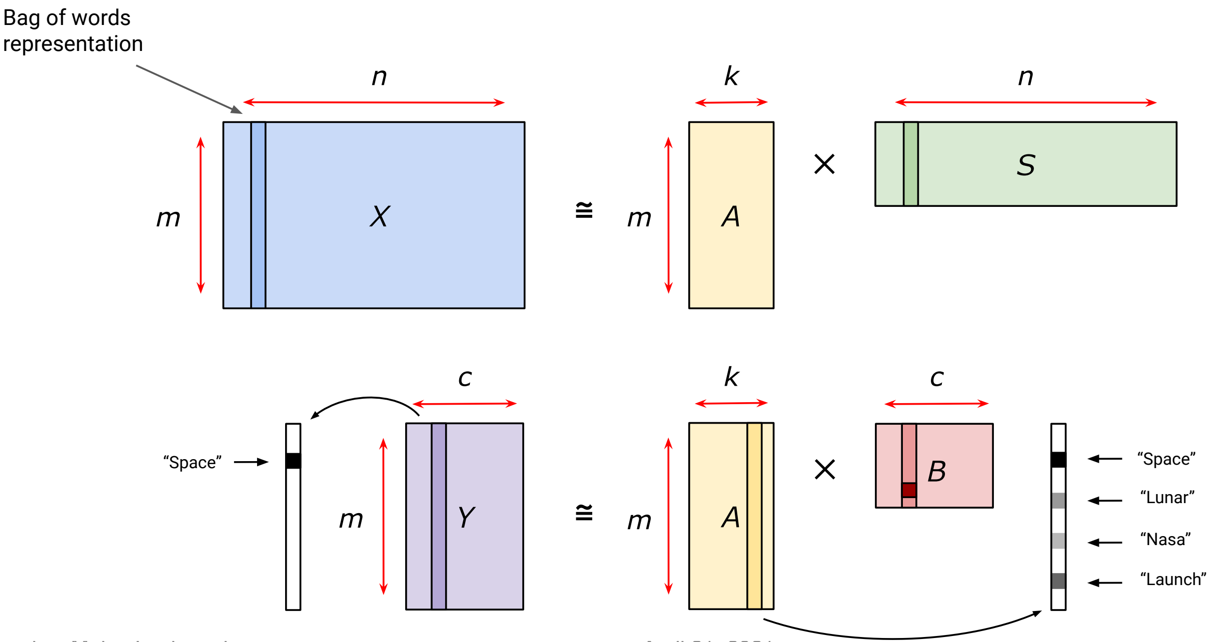
Guided Nonnegative Matrix Factorization (Guided NMF)

Formulated as the optimization task: $$\min_{\boldsymbol{A}\geq 0, \boldsymbol{S}\geq 0, \boldsymbol{B}\geq 0} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\|_F^2$$



Guided Nonnegative Matrix Factorization (Guided NMF)

Example:



20 Newsgroups Data Set

- 20,000 documents containing the text of messages from 20 newsgroups
- We use a subset of 10 newsgroups w/ 100 documents each:

  Graphics, hardware, forsale, motorcycles, baseball, medicine, space, guns, mideast, and religion

rec.motorcycle

```
Newsgroup: rec.motorcycles
document_id: 103140
From:
coburnn@spot.Colorado.EDU
Subject: Identify this
bike for me

OK,
    I saw a bike today and
I want to know what it is.
Lets begin by saying that
its whole rear end...
```

rec.sport.baseball

```
Newsgroup:
rec.sport.baseball
document_id: 102591
From: dxf12@po.CWRU.Edu
Subject: (ATAS) N.L. games
8/2-8/5 & standings of all

Philadelphia at Chicago:
Teams tied for 1st after
Sunday Dick Redding
battled Chet Brewer in the
first game of a...
```

sci.med

```
Newsgroup: sci.med
document_id: 58108
From: geb@cs.pitt.edu
Subject: Re: "CAN'T
BREATHE"

Where did you read this?
I don't think this is
true. I think most medical
treatments are based on
science, although it is
difficult to prove...
```

sci.space

```
Newsgroup: sci.space
document_id: 59850
From: leech@cs.unc.edu
Subject: Space FAQ 14/15 -
How to Become an Astronaut

HOW TO BECOME AN ASTRONAUT

First the short form,
authored by Henry Spencer,
then an official NASA
announcement...
```

K. Lang, “20 newsgroups,” Jan 2008.

NMF on 20 Newsgroups Data Set

Topic 1	Topic 2	Topic 3	Topic 4
drive	israel	space	people
card	jews	nasa	god
mb	israeli	year	think
mac	turkish	games	know
color	arab	team	jesus
video	armenian	game	gun
system	people	shuttle	say
monitor	arabs	launch	see
apple	armenians	data	believe
software	jewish	runs	time

Topic 1	Topic 2	Topic 3	Topic 4
games	geb	drive	people
team	pitt	system	god
game	dsl	card	know
runs	cadre	space	think
year	njxp	mb	israel
pitching	chastity	mac	say
win	skepticism	mail	jesus
last	shameful	software	gun
baseball	banks	new	time
players	intellect	use	see

Tables 1 and 2: Rank 4 NMF on 20 Newsgroups data set.

Guided NMF on 20 Newsgroups Data Set

Topic 1	Topic 2	Topic 3	Topic 4
pitch	*medical*	*space*	people
expected	tests	nasa	know
curveball	disease	shuttle	think
stiffness	diseases	launch	time
loosen	prejudices	sci	use
shoulder	services	lunar	new
shea	graduates	orbit	see
rotation	health	earth	say
game	patients	station	us
giants	available	mission	god

Table 3: Rank 4 Guided NMF on 20 Newsgroups data set with keywords *pitch*, *medical*, and *space*.

Topic 1	Topic 2	Topic 3	Topic 4
motorcycle	*sale*	*religion*	people
bike	offer	christian	know
dod	condition	judaism	think
wheelie	shipping	freedom	time
shaft	asking	christians	use
bikes	includes	islam	new
rider	mb	compulsion	space
riding	excellent	avi	see
scene	price	life	say
ski	best	gunpoint	us

Table 4: Rank 4 Guided NMF on 20 Newsgroups data set with keywords *motorcycle*, *sale*, and *religion*.

Guided NMF on 20 Newsgroups Data Set

Topic 1	Topic 2	Topic 3	Topic 4
pitch	*medical*	*space*	people
expected	tests	nasa	know
curveball	disease	shuttle	think
stiffness	diseases	launch	time
loosen	prejudices	sci	use
shoulder	services	lunar	new
shea	graduates	orbit	see
rotation	health	earth	say
game	patients	station	us
giants	available	mission	god

Table 3: Rank 4 Guided NMF on 20 Newsgroups data set with keywords *pitch*, *medical*, and *space*.

Topic 1	Topic 2	Topic 3	Topic 4
motorcycle	*sale*	*religion*	people
bike	offer	christian	know
dod	condition	judaism	think
wheelie	shipping	freedom	time
shaft	asking	christians	use
bikes	includes	islam	new
rider	mb	compulsion	space
riding	excellent	avi	see
scene	price	life	say
ski	best	gunpoint	us

Table 4: Rank 4 Guided NMF on 20 Newsgroups data set with keywords *motorcycle*, *sale*, and *religion*.

Guided NMF on 20 Newsgroups Data Set

Topic 1	Topic 2	Topic 3	Topic 4
pitch	*medical*	*space*	people
expected	tests	nasa	know
curveball	disease	shuttle	think
stiffness	diseases	launch	time
loosen	prejudices	sci	use
shoulder	services	lunar	new
shea	graduates	orbit	see
rotation	health	earth	say
game	patients	station	us
giants	available	mission	god

Table 3: Rank 4 Guided NMF on 20 Newsgroups data set with keywords *pitch*, *medical*, and *space*.

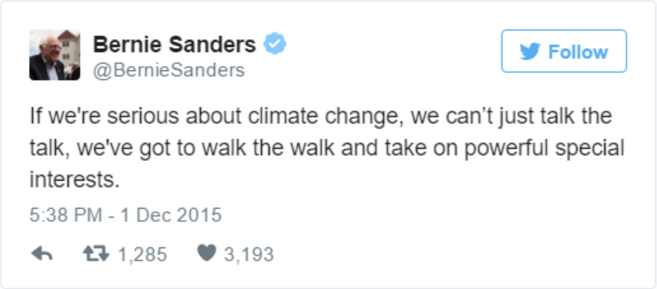
Topic 1	Topic 2	Topic 3	Topic 4
motorcycle	*sale*	*religion*	people
bike	offer	christian	know
dod	condition	judaism	think
wheelie	shipping	freedom	time
shaft	asking	christians	use
bikes	includes	islam	new
rider	mb	compulsion	space
riding	excellent	avi	see
scene	price	life	say
ski	best	gunpoint	us

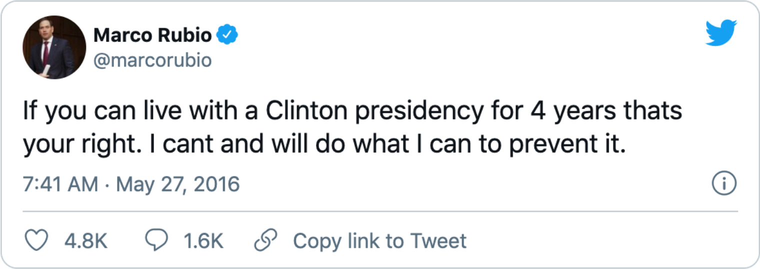
Table 4: Rank 4 Guided NMF on 20 Newsgroups data set with keywords *motorcycle*, *sale*, and *religion*.

Twitter Political Data Set

- A data set of tweets sent by political candidates during the 2016 election season
- We subset the tweets from eight politicians, four Republicans and four Democrats:

  Hillary Clinton, Tim Kaine, Martin O'Malley, Bernie Sanders, Ted Cruz, John Kasich, Marco Rubio, and Donald Trump





J. Littman, L. Wrubel, and D. Kerchner, “2016 United States Presidential Election Tweet Ids,” 2016.

NMF vs. Guided NMF on Twitter Political Data Set

Topic 1	Topic 2	Topic 3	Topic 4
thank	govpencein	gopdebate	tedcruz
trump2016	indiana	imwithhuck	cruz
maga[1]	indiana_edc	jeb	cruzcrew
great	state	tonight	ted
america	jobs	president	choosecruz
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
kasich	hillary	randpaul	fitn
john	trump	iowa	new
johnkasich	people	iacaucus	hampshire
ohio	donald	caucus	johnkasich
gov	president	tonight	nh

[1] Here “maga" abbreviates “makeamericagreatagain."

Table 5: Rank 8 NMF on Twitter political data set.

Topic 1	Topic 2	Topic 3	Topic 4
economy	*obamacare*	govpencein	gopdebate
jobs	fullrepeal	indiana	kasich
tax	repeal	indiana_edc	randpaul
plan	replace	state	john
create	fight	jobs	tonight
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
tedcruz	hillary	johnkasich	people
thank	trump	new	need
cruz	donald	fitn	must
cruzcrew	clinton	kasich	berniesanders
ted	president	hampshire	country

Table 6: Rank 8 Guided NMF on Twitter political data set with keywords *economy* and *obamacare*.

NMF vs. Guided NMF on Twitter Political Data Set

Topic 1	Topic 2	Topic 3	Topic 4
thank	govpencein	gopdebate	tedcruz
trump2016	indiana	imwithhuck	cruz
maga[1]	indiana_edc	jeb	cruzcrew
great	state	tonight	ted
america	jobs	president	choosecruz
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
kasich	hillary	randpaul	fitn
john	trump	iowa	new
johnkasich	people	iacaucus	hampshire
ohio	donald	caucus	johnkasich
gov	president	tonight	nh

[1] Here "maga" abbreviates "makeamericagreatagain."

Table 5: Rank 8 NMF on Twitter political data set.

Topic 1	Topic 2	Topic 3	Topic 4
economy	*obamacare*	govpencein	gopdebate
jobs	fullrepeal	indiana	kasich
tax	repeal	indiana_edc	randpaul
plan	replace	state	john
create	fight	jobs	tonight
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
tedcruz	hillary	johnkasich	people
thank	trump	new	need
cruz	donald	fitn	must
cruzcrew	clinton	kasich	berniesanders
ted	president	hampshire	country

Table 6: Rank 8 Guided NMF on Twitter political data set with keywords *economy* and *obamacare.*

NMF vs. Guided NMF on Twitter Political Data Set

Topic 1	Topic 2	Topic 3	Topic 4
thank	govpencein	gopdebate	tedcruz
trump2016	indiana	imwithhuck	cruz
maga[1]	indiana_edc	jeb	cruzcrew
great	state	tonight	ted
america	jobs	president	choosecruz
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
kasich	hillary	randpaul	fitn
john	trump	iowa	new
johnkasich	people	iacaucus	hampshire
ohio	donald	caucus	johnkasich
gov	president	tonight	nh

[1] Here "maga" abbreviates "makeamericagreatagain."

Table 5: Rank 8 NMF on Twitter political data set.

Topic 1	Topic 2	Topic 3	Topic 4
economy	*obamacare*	govpencein	gopdebate
jobs	fullrepeal	indiana	kasich
tax	repeal	indiana_edc	randpaul
plan	replace	state	john
create	fight	jobs	tonight
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
tedcruz	hillary	johnkasich	people
thank	trump	new	need
cruz	donald	fitn	must
cruzcrew	clinton	kasich	berniesanders
ted	president	hampshire	country

Table 6: Rank 8 Guided NMF on Twitter political data set with keywords *economy* and *obamacare*.

NMF vs. Guided NMF on Twitter Political Data Set

Topic 1	Topic 2	Topic 3	Topic 4
thank	govpencein	gopdebate	tedcruz
trump2016	indiana	imwithhuck	cruz
maga[1]	indiana_edc	jeb	cruzcrew
great	state	tonight	ted
america	jobs	president	choosecruz
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
kasich	hillary	randpaul	fitn
john	trump	iowa	new
johnkasich	people	iacaucus	hampshire
ohio	donald	caucus	johnkasich
gov	president	tonight	nh

[1] Here "maga" abbreviates "makeamericagreatagain."

Table 5: Rank 8 NMF on Twitter political data set.

Topic 1	Topic 2	Topic 3	Topic 4
economy	*obamacare*	govpencein	gopdebate
jobs	fullrepeal	indiana	kasich
tax	repeal	indiana_edc	randpaul
plan	replace	state	john
create	fight	jobs	tonight
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
tedcruz	hillary	johnkasich	people
thank	trump	new	need
cruz	donald	fitn	must
cruzcrew	clinton	kasich	berniesanders
ted	president	hampshire	country

Table 6: Rank 8 Guided NMF on Twitter political data set with keywords *economy* and *obamacare*.

Guided NMF vs. Seeded LDA on 20 Newsgroups

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.83**	**0.88**	**0.88**	**0.87**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.86**	**0.87**	0.88	0.87
	Seeded LDA	0.37	0.5	**0.91**	**0.89**
10	Guided NMF	**0.88**	0.89	0.89	0.89
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 7: AUC scores for 20 Newsgroups data set on documents in the space class

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.89**	**0.9**	**0.9**	**0.9**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.9**	**0.9**	0.9	**0.9**
	Seeded LDA	0.37	0.5	**0.91**	0.89
10	Guided NMF	**0.87**	0.9	0.9	0.9
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 8: AUC scores for 20 Newsgroups data set on documents in the baseball class

Guided NMF vs. Seeded LDA on 20 Newsgroups

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.83**	**0.88**	**0.88**	**0.87**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.86**	**0.87**	0.88	0.87
	Seeded LDA	0.37	0.5	**0.91**	**0.89**
10	Guided NMF	**0.88**	0.89	0.89	0.89
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 7: AUC scores for 20 Newsgroups data set on documents in the space class

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.89**	**0.9**	**0.9**	**0.9**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.9**	**0.9**	0.9	**0.9**
	Seeded LDA	0.37	0.5	**0.91**	0.89
10	Guided NMF	**0.87**	0.9	0.9	0.9
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 8: AUC scores for 20 Newsgroups data set on documents in the baseball class

Guided NMF vs. Seeded LDA on 20 Newsgroups

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.83**	**0.88**	**0.88**	**0.87**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.86**	**0.87**	0.88	0.87
	Seeded LDA	0.37	0.5	**0.91**	**0.89**
10	Guided NMF	**0.88**	0.89	0.89	0.89
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 7: AUC scores for 20 Newsgroups data set on documents in the space class

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.89**	**0.9**	**0.9**	**0.9**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.9**	**0.9**	0.9	**0.9**
	Seeded LDA	0.37	0.5	**0.91**	0.89
10	Guided NMF	**0.87**	0.9	0.9	0.9
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 8: AUC scores for 20 Newsgroups data set on documents in the baseball class

Thank You

J. Vendrow, J. Haddock, E. Rebrova, D. Needell. “On a Guided Nonnegative Matrix Factorization.” Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2021.