CSE 472
Assignment 3: Topic Modeling

Kazi Ashik Islam
1205007

April 24, 2017

Q1: When does the sampler converge if you don't consider burn-in and lag? (The 5 most frequent words of each topic remain same after certain iterations).

A1: To check if the output has converged, we checked when 100 / 200 / 300 consecutive outputs become same. The observation was as follows:

Consecutive Similar Output	Iteration NO.	Maximum Iteration
100	1641	10K
200	2869	10K
300	4361	10K

So from this observation, we can say that, the program actually didn't converge fully when burn-in and lag are not considered. The sampler may converge with increased number of iterations.

Q2: Is there any effect of burn-in? (The 5 most frequent words of each topic make more sense if we discard burn-in iterations).

A2: The effect of burn-in to the output is shown in the following table: -

Table 1: BURN-IN 100

Topic0	car	book	cost	want	auto
Topic1	henry	toronto	spencer	high	zoo
Topic2	oil	service	come	change	lights
Topic3	edu	article	writes	apr	engines
Topic4	insurance	geico	state	companies	good
Topic5	engine	power	turbo	steering	audi
Topic6	car	ford	probe	good	speed
Topic7	edu	gif	uci	ics	incoming
Topic8	space	bill	moon	sci	back
Topic9	science	part	mars	nasa	internet
Topic10	george	diesels	people	cars	heard
Topic11	edu	writes	article	mustang	camaro

Topic12	cars	manual	small	toyota	seat
Topic13	don	make	time	even	two
Topic14	mission	hst	shuttle	solar	pat
Topic15	car	clutch	shifter	sho	shift
Topic16	system	oort	writes	ray	cloud
Topic17	station	launch	option	shuttle	redesign
Topic18	etc	earth	day	large	planets
Topic19	sky	light	rights	night	uiuc

Table 2: BURN-IN 1K

Topic0	cars	manual	small	seat	toyota
Topic1	engine	power	turbo	steering	audi
Topic2	oil	service	come	lights	change
Topic3	edu	writes	article	mustang	blah
Topic4	car	ford	probe	speed	miles
Topic5	edu	gif	uci	ics	incoming
Topic6	sky	light	people	rights	night
Topic7	henry	toronto	spencer	zoo	svr
Topic8	mission	hst	shuttle	solar	pat
Topic9	writes	edu	article	apr	engines
Topic10	book	cost	want	saturn	years
Topic11	insurance	geico	good	state	radar
Topic12	space	science	internet	information	world
Topic13	cars	diesels	heard	matter	emissions
Topic14	etc	earth	day	large	planets
Topic15	bill	moon	george	back	howell
Topic16	station	launch	option	shuttle	redesign
Topic17	don	make	even	time	find
Topic18	system	mars	oort	spacecraft	part
Topic19	clutch	shifter	sho	car	shift

Table 3: BURN-IN 5K

Topic0	edu	writes	article	apr	eliot
Topic1	oil	service	come	lights	change
Topic2	cars	diesels	heard	matter	emissions
Topic3	science	internet	part	mars	information
Topic4	system	oort	writes	ray	cloud
Topic5	henry	toronto	spencer	zoo	svr
Topic6	space	nasa	sci	long	world
Topic7	car	book	cost	auto	want

Topic8	engine	turbo	small	toyota	power
Topic9	bill	moon	george	back	howell
Topic10	station	launch	option	shuttle	redesign
Topic11	insurance	geico	good	state	radar
Topic12	writes	article	edu	apr	mustang
Topic13	sky	light	people	rights	night
Topic14	car	ford	probe	speed	miles
Topic15	etc	earth	day	large	planets
Topic16	clutch	shifter	sho	shift	gear
Topic17	edu	gif	uci	ics	incoming
Topic18	don	make	even	time	find
Topic19	mission	hst	shuttle	solar	pat

(10K iterations and stride of size 50.)

From the above tables we can see that there is small difference in output for different values of burn-in.

The motivation behind burn-in is that, at first, topics are assigned to words randomly. So these initial assignments shouldn't be of any value to us. We should consider assignments when the sampler is near convergence. That is why we discard some of the initial iterations.

Q3: Is there any effect of stride/lag?

A3: The effect of stride is shown in the following tables: -

Table 4 STRIDE SIZE 5

Topic0	clutch	car	shifter	sho	shift
Topic1	station	launch	option	shuttle	redesign
Topic2	sky	light	rights	people	point
Topic3	edu	gif	uci	ics	incoming
Topic4	oil	service	come	lights	change
Topic5	henry	toronto	spencer	zoo	svr
Topic6	space	science	internet	information	world
Topic7	edu	writes	article	mustang	camaro
Topic8	insurance	geico	good	radar	two
Topic9	engine	power	turbo	steering	audi
Topic10	don	make	even	people	two
Topic11	bill	moon	back	george	support
Topic12	book	want	cost	auto	saturn
Topic13	etc	earth	day	large	planets
Topic14	mission	hst	shuttle	solar	pat
Topic15	writes	edu	article	apr	engines
Topic16	cars	diesels	heard	matter	emissions

Topic17	cars	manual	small	toyota	seat
Topic18	car	ford	probe	speed	pretty
Topic19	system	part	mars	oort	spacecraft

Table 5 STRIDE SIZE 10

Topic0	nasa	part	mars	spacecraft	gov
Topic1	clutch	shifter	car	sho	shift
Topic2	etc	earth	day	large	planets
Topic3	writes	edu	article	people	don
Topic4	edu	article	writes	apr	engines
Topic5	car	ford	probe	speed	miles
Topic6	mission	hst	shuttle	solar	pat
Topic7	mustang	diesels	george	cars	emissions
Topic8	insurance	geico	good	companies	state
Topic9	writes	edu	system	oort	ray
Topic10	henry	toronto	spencer	zoo	svr
Topic11	book	cost	want	car	saturn
Topic12	sky	light	night	bright	uiuc
Topic13	engine	power	turbo	steering	audi
Topic14	don	make	edu	writes	even
Topic15	space	bill	moon	sci	internet
Topic16	oil	service	come	change	lights
Topic17	station	launch	option	shuttle	space
Topic18	cars	manual	small	seat	toyota
Topic19	edu	gif	uci	ics	incoming

Table 6 STRIDE SIZE 20

Topic0	car	book	cost	want	auto
Topic1	bill	moon	back	support	act
Topic2	cars	diesels	heard	matter	emissions
Topic3	henry	toronto	spencer	zoo	svr
Topic4	mission	hst	shuttle	solar	pat
Topic5	space	internet	science	nasa	world
Topic6	part	mars	spacecraft	system	propulsion
Topic7	etc	earth	day	large	planets
Topic8	car	ford	probe	speed	feel
Topic9	engine	toyota	small	seat	cars
Topic10	insurance	geico	good	state	radar
Topic11	edu	article	writes	apr	engines

Topic12	don	make	even	find	time
Topic13	sky	light	people	rights	night
Topic14	edu	gif	uci	ics	incoming
Topic15	writes	edu	oort	system	ray
Topic16	oil	service	come	change	lights
Topic17	clutch	shifter	sho	car	shift
Topic18	george	mustang	blah	howell	edu
Topic19	station	launch	option	shuttle	redesign

Table 7 STRIDE SIZE 50

Topic0	space	science	internet	world	sci
Topic1	henry	toronto	spencer	zoo	svr
Topic2	etc	earth	day	large	planets
Topic3	edu	gif	uci	ics	incoming
Topic4	oil	service	come	change	lights
Topic5	station	launch	option	shuttle	redesign
Topic6	system	mars	oort	spacecraft	gov
Topic7	mission	hst	shuttle	solar	pat
Topic8	don	make	writes	want	cost
Topic9	insurance	geico	good	radar	state
Topic10	edu	writes	article	apr	eliot
Topic11	clutch	shifter	sho	car	shift
Topic12	sky	light	people	rights	night
Topic13	engine	power	turbo	steering	audi
Topic14	edu	writes	george	mustang	howell
Topic15	bill	moon	back	time	support
Topic16	cars	manual	find	seat	small
Topic17	book	make	don	auto	want
Topic18	cars	diesels	writes	article	heard
Topic19	car	ford	probe	speed	miles

(10K iterations, 5K burn-in)

From the above tables, we can see that the output differs as the stride size is changed.

The motivation behind using strides, was to make the successive samples (generated by the gibbs sampler) un-correlated. As the output changes with stride size, we can conclude that, the samples are correlated when the stride size is small. So the use of strides is justified.