Project 1: Topic Modelling Analysis of News Articles

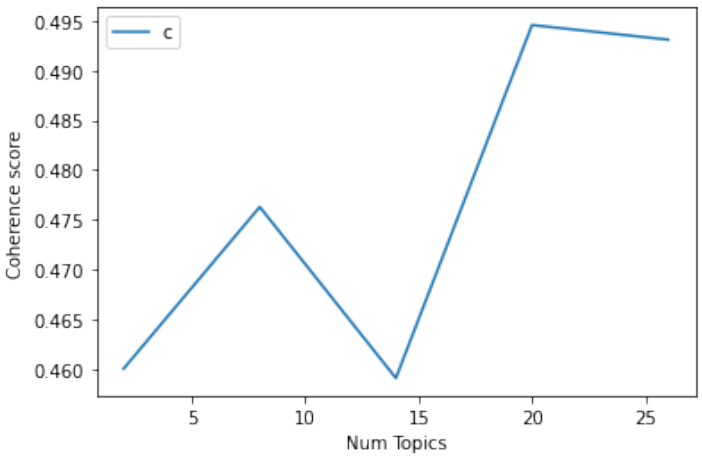
Alejandrina Jimenez Guzman
aj7354@rit.edu

Santurkar, Vaibhav
vs4503@rit.edu

October 18, 2021

1 Performance of the LDA model using coherence measure

For this project we used 107 articles from the Washington post and the NFL news website. Originally we used the Gensim's inbuilt version of the LDA algorithm but the coherence metric indicated that the topic modeling was not very good, and it was hard to find a proper name for some topics, therefore we generated a LDA model using mallet to obtain better quality of topics. We made use of the Gensim wrapper to implement Mallet's LDA from Gensim. And we obtained the following graph:



The results we obtained for different topics was:

```
Num Topics = 2  has Coherence Value of 0.4601
Num Topics = 8  has Coherence Value of 0.4763
Num Topics = 14  has Coherence Value of 0.4591
Num Topics = 20  has Coherence Value of 0.4946
Num Topics = 26  has Coherence Value of 0.4931
```

From these results we decided to use the model with 20 topics since it gave the highest Coherence Value before dropping and flattening out.

For this part of the project we used as guidelines the pipelines presented in [3, 2] and the concepts of Latent Dirichlet Allocation[1].

2 Tables of the 20 most likely words per Topic

For this section we decided to keep the weights of the words to indicate how they influenced in the decision for naming every topic. The weights reflect how important a keyword is to that topic.

0:Public Health and Safety	1:Upcoming Fixtures	2:Football Statistics	3:Politics and Activisim
0.025*"year"	0.028*"play"	0.038*"game"	0.020*"govern"
0.015*"pandem"	0.027*"washington"	0.036*"rush"	0.016*"elect"
0.015*"import"	0.020*"field"	0.030*"run"	0.016*"presid"
0.013*"month"	0.019*"run"	0.027*"back"	0.016*"biden"
0.013*"offici"	0.018*"lead"	0.025*"win"	0.015*"hous"
0.012*"high"	0.017*"quarter"	0.022*"td"	0.014*"million"
0.012*"percent"	0.017*"previou"	0.017*"yard"	0.014*"vote"
0.010*"sign"	0.014*"pa"	0.016*"rec"	0.014*"democrat"
0.009*"problem"	0.014*"goal"	0.014*"season"	0.012*"gener"
0.009*"countri"	0.013*"chief"	0.012*"year"	0.010*"mani"
0.009*"low"	0.013*"yard"	0.011*"thi"	0.010*"american"
0.009*"price"	0.013*"citi"	0.010*"career"	0.009*"report"
0.009*"larg"	0.013*"game"	0.010*"yd"	0.009*"republican"
0.008*"major"	0.012*"kansa"	0.010*"stat"	0.009*"accord"
0.008*"deliv"	0.011*"end"	0.010*"chanc"	0.008*"spend"
0.008*"polici"	0.011*"half"	0.010*"perform"	0.008*"parti"
0.008*"peopl"	0.010*"tackl"	0.009*"continu"	0.008*"trump"
0.008*"earli"	0.010*"line"	0.008*"mauric"	0.007*"white"
0.008*"issu"	0.009*"leav"	0.008*"base"	0.007*"organ"
0.007*"lose"	0.009*"mahom"	0.008*"top"	0.007*"democraci"

4:Predictive Analysis	5:Medical Treatment	6:Fantasy Football Statistics	7:Business and Travel
0.051*"start"	0.027*"eihab"	0.061*"week"	0.036*"tip"
0.047*"em"	0.020*"famili"	0.025*"target"	0.018*"tri"
0.044*"question"	0.015*"falah"	0.024*"time"	0.017*"servic"
0.042*"sit"	0.014*"medic"	0.022*"yard"	0.016*"travel"
0.040*"season"	0.012*"franklin"	0.021*"play"	0.014*"time"
0.033*"back"	0.012*"doctor"	0.019*"catch"	0.013*"ye"
0.033*"nfl"	0.012*"eva"	0.015*"fantasi"	0.012*"hotel"
0.031*"answer"	0.012*"commun"	0.014*"high"	0.011*"app"
0.030*"point"	0.011*"pittsburgh"	0.012*"point"	0.010*"airport"
0.030*"lineup"	0.008*"caus"	0.012*"rb"	0.010*"clean"
0.029*"hope"	0.008*"begin"	0.011*"end"	0.010*"add"
0.024*"week"	0.008*"week"	0.011*"receiv"	0.010*"correct"
0.020*"good"	0.008*"hospit"	0.011*"big"	0.009*"cash"
0.019*"game"	0.007*"jewish"	0.011*"percent"	0.009*"charg"
0.015*"im"	0.007*"hill"	0.011*"option"	0.008*"gener"
0.013*"fantasi"	0.007*"treatment"	0.011*"william"	0.008*"rule"
0.011*"trademark"	0.007*"diseas"	0.010*"break"	0.008*"person"
0.010*"nt"	0.007*"life"	0.010*"carri"	0.008*"line"
0.009*"footbal"	0.006*"specialist"	0.010*"touchdown"	0.007*"work"
0.008*"realli"	0.006*"peopl"	0.010*"game"	0.007*"recommend"

8:NFL Match	9:Journalism and Literature	10:Public Health	11:Lifestyle
0.050*"nfl"	0.017*"oct"	0.048*"test"	0.024*"day"
0.033*"season"	0.015*"washington"	0.036*"vaccin"	0.021*"leav"
0.027*"week"	0.014*"state"	0.023*"coronaviru"	0.018*"mike"
0.026*"rank"	0.013*"unit"	0.020*"state"	0.016*"call"
0.023*"team"	0.012*"stori"	0.018*"case"	0.013*"live"
0.023*"game"	0.011*"post"	0.017*"report"	0.012*"becom"
0.017*"defens"	0.010*"saturday"	0.017*"peopl"	0.012*"place"
0.015*"sunday"	0.009*"space"	0.016*"antibodi"	0.012*"make"
0.014*"offens"	0.008*"news"	0.015*"health"	0.011*"hous"
0.011*"quarterback"	0.008*"day"	0.014*"data"	0.011*"turn"
0.011*"pass"	0.008*"editor"	0.013*"death"	0.011*"move"
0.011*"brown"	0.008*"photo"	0.013*"level"	0.011*"home"
0.010*"top"	0.008*"pm"	0.012*"number"	0.010*"close"
0.009*"make"	0.008*"minut"	0.011*"day"	0.010*"stay"
0.009*"touchdown"	0.008*"box"	0.010*"system"	0.009*"put"
0.008*"loss"	0.008*"imag"	0.009*"rate"	0.009*"crystal"
0.008*"yard"	0.008*"missil"	0.008*"provid"	0.008*"fire"
0.008*"footbal"	0.008*"world"	0.008*"show"	0.008*"drive"
0.007*"injuri"	0.008*"screen"	0.008*"hospit"	0.008*"anoth"
0.007*"rooki"	0.008*"set"	0.008*"includ"	0.008*"give"

12:Jobs and Employment	13:Law Enforcement and Violence	14:Player Trading Market	15:Food Prep
0.037*"work"	0.029*"word"	0.035*"team"	0.030*"home"
0.023*"compani"	0.015*"polic"	0.028*"player"	0.016*"make"
0.020*"job"	0.014*"post"	0.019*"trade"	0.016*"food"
0.018*"pay"	0.013*"term"	0.018*"nt"	0.012*"recip"
0.018*"employe"	0.012*"woman"	0.015*"time"	0.011*"detecto
0.017*"worker"	0.012*"peopl"	0.014*"big"	0.011*"find"
0.013*"scarlett"	0.009*"editor"	0.014*"year"	0.010*"agre"
0.013*"power"	0.008*"violenc"	0.012*"coach"	0.009*"appl"
0.013*"hour"	0.008*"compani"	0.011*"gruden"	0.009*"kind"
0.011*"appl"	0.008*"agre"	0.011*"make"	0.009*"problem
0.011*"post"	0.007*"accord"	0.010*"leagu"	0.009*"questio
0.011*"write"	0.007*"deliv"	0.010*"everi"	0.008*"cook"
0.009*"stori"	0.007*"morn"	0.010*"footbal"	0.008*"bake"
0.009*"begin"	0.007*"repres"	0.010*"email"	0.008*"replac"
0.009*"time"	0.007*"abus"	0.010*"anyon"	0.008*"air"
0.009*"hire"	0.007*"law"	0.009*"reid"	0.007*"design"
0.008*"posit"	0.007*"kill"	0.009*"start"	0.007*"sign"
0.008*"offer"	0.007*"puzzl"	0.008*"hit"	0.007*"indoor"
0.008*"day"	0.006*"slur"	0.008*"discuss"	0.007*"mold"
0.007*"amazon"	0.006*"durst"	0.008*"deal"	0.007*"list"

16:Literature and Publishing	17:Public Education	18:Lifestyle	19:Environment
0.021*"write"	0.035*"school"	0.013*"famili"	0.020*"chang"
0.013*"earhart"	0.020*"citi"	0.012*"make"	0.016*"climat"
0.013*"book"	0.019*"street"	0.012*"life"	0.015*"road"
0.011*"return"	0.015*"student"	0.011*"person"	0.015*"wine"
0.011*"piec"	0.014*"teacher"	0.010*"stori"	0.011*"farm"
0.010*"part"	0.012*"year"	0.010*"differ"	0.011*"restaur"
0.009*"museum"	0.011*"system"	0.009*"someth"	0.011*"state"
0.009*"vampir"	0.011*"counti"	0.009*"everyon"	0.010*"water"
0.009*"world"	0.010*"requir"	0.008*"light"	0.009*"grow"
0.009*"town"	0.010*"part"	0.008*"mani"	0.008*"park"
0.009*"news"	0.010*"french"	0.008*"follow"	0.008*"bottl"
0.008*"publish"	0.010*"local"	0.008*"alway"	0.008*"crop"
0.008*"poem"	0.009*"forc"	0.008*"peopl"	0.007*"rain"
0.008*"turn"	0.009*"staff"	0.007*"kid"	0.007*"local"
0.008*"author"	0.008*"war"	0.007*"realli"	0.007*"fish"
0.008*"lawmak"	0.007*"fight"	0.007*"import"	0.007*"weather"
0.008*"roll"	0.007*"free"	0.007*"talk"	0.006*"carbon"
0.008*"stori"	0.007*"kill"	0.007*"day"	0.006*"foot"
0.008*"everi"	0.006*"germain"	0.007*"son"	0.006*"river"
0.007*"public"	0.006*"news"	0.006*"agre"	0.006*"warm"

3 Projection of articles into the topic model

Topic	Title	Date
0.0	5 charts that explain inflation, wages, supply ...	9:56 a.m. EDT
11.0	Survivors of a California wildfire navigate lif...	9:00 a.m. EDT
16.0	Denver Art Museum plans to return four Cambodia...	10:00 p.m. EDT
5.0	Former president Bill Clinton discharged from h...	11:23 a.m. EDT
10.0	Russia's official coronavirus count faces quest...	6:00 a.m. EDT
17.0	Confederate streets in Alexandria targeted for ...	6:00 a.m. EDT
16.0	A woman won a million-euro Spanish literary pri...	9:55 a.m. EDT
7.0	Do you know how to tip a bellhop or housekeepin...	None
3.0	Advocates worry democracy is eroding on Biden's...	8:00 a.m. EDT
16.0	David Amess stabbing: Britain considers police ...	11:08 a.m. EDT
18.0	Adele's 'Easy on Me' new single breaks Spotify,...	5:59 a.m. EDT
8.0	NFL Week 6 scores and live updates - The Washin...	1:14 p.m. EDT
18.0	The NBA's Kyrie problem - The Washington Post	None
16.0	The new Rolling Stone: 'More immediate, more vi...	7:00 a.m. EDT

13.0	Merriam-Webster promoted Typeshift, a word game...	October 15, 2021 at 5:00 p.m. EDT
5.0	Mourners of the Pittsburgh synagogue shooting. ...	None
18.0	Date Lab:These two tried to keep a secret from...	October 14, 2021 at 6:00 a.m. EDT
5.0	Second Glance: Doll shop, Oct. 10, 2021 - The W...	None
18.0	A museum dedicated to the family of John Wilkes...	October 11, 2021 at 9:00 a.m. EDT
7.0	How to clean your bathroom in 10 minutes, 30 mi...	October 14, 2021 at 7:00 a.m. EDT
15.0	How to prepare your home for winter - The Washi...	October 15, 2021 at 9:00 a.m. EDT
15.0	How to replace wired smoke detectors and when y...	October 15, 2021 at 7:00 a.m. EDT
15.0	How to check indoor air quality and when to tes...	7:00 a.m. EDT
10.0	Antibody tests can't give answers you want abou...	9:00 a.m. EDT
10.0	Am I eligible for a coronavirus booster shot? -...	None
10.0	Tracking the covid vaccine: Doses, people vacci...	None
19.0	At Alaska's most popular national park, climate...	October 15, 2021 at 5:10 p.m. EDT
7.0	How to responsibly dispose of your old CDs, DVD...	October 15, 2021 at 8:00 a.m. EDT
19.0	Earthshot Prize: These innovations could win 1 ...	6:00 a.m. EDT
19.0	Extreme weather and climate this summer challen...	8:00 a.m. EDT
3.0	Are Americans growing warier of more government...	12:13 p.m. EDT
3.0	The middle falls out - The Washington Post	October 15, 2021 at 1:16 p.m. EDT
3.0	Caroline Wren helped Publix heiress who funded ...	9:34 a.m. EDT
19.0	In India's Assam, death of Muslim man during ev...	9:47 a.m. EDT
13.0	China issues death sentence for man who set ex-...	1:38 p.m. EDT
13.0	Maduro ally Alex Saab extradited to U.S. as Ven...	12:36 a.m. EDT
13.0	SEPTA riders watched as woman was raped, police...	9:23 p.m. EDT
13.0	House of Representatives staffer arrested on ch...	8:53 p.m. EDT
17.0	Concord Review has published long student essay...	6:00 a.m. EDT
13.0	Robert Durst tests positive for covid-19 days a...	5:44 p.m. EDT
0.0	Prices rise from groceries to car rentals due t...	October 14, 2021 at 5:00 p.m. EDT
12.0	How people who quit their jobs are getting by f...	October 14, 2021 at 7:00 a.m. EDT
12.0	Apple employee Cher Scarlett is leading a worke...	October 14, 2021
12.0	Warehouses are looking for seasonal workers thi...	October 11, 2021 at 12:18 p.m. EDT
17.0	Deadlines arrive for school staff to be vaccina...	6:00 a.m. EDT
17.0	District to hire more pandemic staff to help sc...	7:00 p.m. EDT
13.0	Maxwell Bero, former Montgomery County teacher,...	October 15, 2021 at 7:00 p.m. EDT
3.0	McAuliffe outraises Youngkin in Virginia Govern...	4:09 p.m. EDT
9.0	Nike missiles around Washington - The Washingto...	4:12 p.m. EDT
9.0	The Soul Box Project puts an origami box on The...	5:17 p.m. EDT
18.0	D.C.-area forecast: Much cooler air rides in on...	6:00 a.m. EDT
10.0	Tracking coronavirus deaths, cases and vaccinat...	None
19.0	A sustainable-seafood claim by this waiter was ...	October 15, 2021 at 8:00 a.m. EDT
9.0	Where to find screenings of 'Rocky Horror Pictu...	October 15, 2021 at 6:00 a.m. EDT
5.0	Her unexplained jitteriness and weight loss wer...	10:00 a.m. EDT

10.0	Rabbit test detected pregnancy by injecting uri...	7:00 a.m. EDT
19.0	Alligator gar: Kansas fisherman caught 40-pound...	October 14, 2021
9.0	China sends three astronauts to its first perma...	October 15, 2021 at 10:10 a.m. EDT
1.0	Washington Football Team vs. Chiefs: Live updat...	2:23 p.m. EDT
1.0	Andy Reid faces another former assistant in Ron...	5:00 a.m. EDT
2.0	Dodgers-Braves NLCS Game 1: Austin Riley, Ozzie...	12:02 a.m. EDT
8.0	NFL Week 6: Sunday TV schedule features Ravens ...	3:00 a.m. EDT
9.0	As the push for a new owner continues, the Wash...	1:22 a.m. EDT
9.0	Points-hungry D.C. United settles for draw with...	11:20 p.m. EDT
9.0	Nick Rolovich's future remains unclear after wi...	11:56 a.m. EDT
1.0	Kahleah Copper has helped push Chicago Sky to v...	6:00 a.m. EDT
15.0	Selling home-cooked food is getting easier, tha...	October 13, 2021 at 8:00 a.m. EDT
17.0	Hubert Germain, last French WWII 'Companion of ...	October 15, 2021 at 6:39 p.m. EDT
9.0	Pictures of what happened this week: A painting...	None
18.0	Carolyn Hax: Teen wants space, gets upset when ...	12:00 a.m. EDT
18.0	Ask Amy: Couple operate a business, but only on...	12:00 a.m. EDT
18.0	Miss Manners: Illness takes patient's hair, lea...	12:00 a.m. EDT
12.0	IATSE strike averted as union reaches deal with...	11:10 p.m. EDT
16.0	Amelia Earhart's poems reveal her inner thought...	6:00 a.m. EDT
18.0	Suzanne Valadon exhibition at the Barnes is a s...	October 15, 2021 at 7:00 a.m. EDT
9.0	New movies to stream from home this week. - The...	October 14, 2021 at 10:00 a.m. EDT
16.0	'Dracula' brought vampires into the limelight. ...	8:00 a.m. EDT
18.0	'My Monticello, by Jocelyn Nicole Johnson book ...	October 15, 2021 at 7:00 a.m. EDT
15.0	This white chili with butternut squash recipe b...	10:00 a.m. EDT
15.0	7 apple recipes for baking pies, cakes, tarts a...	10:00 a.m. EDT
15.0	5 dieting truths everyone can – or should – a...	October 15, 2021 at 12:00 p.m. EDT
19.0	Carbon farming, one of the oldest agricultural ...	October 15, 2021 at 12:00 p.m. EDT
7.0	Why airport power outlets don't work - The Wash...	8:00 a.m. EDT
7.0	Do you know how to tip a bellhop or housekeepin...	None
7.0	How to tip without cash while traveling - The W...	October 15, 2021 at 3:38 p.m. EDT
16.0	Istria, Croatia: Exploring food, architecture a...	October 15, 2021 at 10:00 a.m. EDT
8.0	2021 NFL season, Week 6: What we learned from B...	Oct 14, 2021 at 11:23 PM
8.0	NFL Week 6 bold predictions: Cardinals suffer f...	Oct 15, 2021 at 10:50 AM
2.0	RB Index, Week 6: Browns, Cowboys lead NFL's to...	Oct 14, 2021 at 11:22 AM
8.0	Top 10 surprising performances so far: Cordarre...	Oct 14, 2021 at 01:38 PM
1.0	2022 NFL Draft: Top 25 Senior Bowl prospects at...	Oct 14, 2021 at 03:57 PM
14.0	Las Vegas Raiders left with 'a lot to process' ...	Oct 13, 2021 at 11:10 PM
8.0	Week 6 NFL underdogs: Will the real Washington ...	Oct 13, 2021 at 08:23 AM
8.0	NFL Power Rankings, Week 6: Bills dethrone Card...	Oct 12, 2021 at 08:08 AM
14.0	NFL trade deadline: An executive's guide to pla...	Oct 13, 2021 at 11:47 AM
8.0	Offensive Player Rankings, Week 6: Five QBs und...	Oct 12, 2021 at 01:42 PM

14.0	Jon Gruden's words cut against NFL's efforts to...	Oct 12, 2021 at 12:21 AM
8.0	Seven NFL teams at risk of falling apart due to...	Oct 10, 2021 at 10:11 PM
14.0	Devin White: The impact of my late brother's me...	Sep 24, 2021 at 10:31 AM
4.0	2021 NFL Fantasy Football Start 'Em, Sit 'Em We...	Oct 13, 2021 at 10:00 AM
4.0	2021 NFL Fantasy Football Start 'Em, Sit 'Em We...	Oct 13, 2021 at 10:00 AM
4.0	2021 NFL Fantasy Football Start 'Em, Sit 'Em We...	Oct 13, 2021 at 10:00 AM
4.0	2021 NFL Fantasy Football Start 'Em, Sit 'Em We...	Oct 13, 2021 at 10:00 AM
4.0	2021 NFL Fantasy Football Start 'Em, Sit 'Em We...	Oct 13, 2021 at 10:00 AM
4.0	2021 NFL Fantasy Football Start 'Em, Sit 'Em We...	Oct 13, 2021 at 10:00 AM
6.0	Marcas Grant's 2021 NFL Fantasy Football Sleepe...	Oct 12, 2021 at 07:56 PM
6.0	Fantasy waiver wire targets for Week 6 of 2021 ...	Oct 11, 2021 at 03:40 PM

References

[1] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

[2] Dwivedi Priya. Nlp: Extracting the main topics from your dataset using lda in minutes. 2018.

[3] Dua Sejal. Nlp preprocessing and latent dirichlet allocation (lda) topic modeling with gensim. 2021.