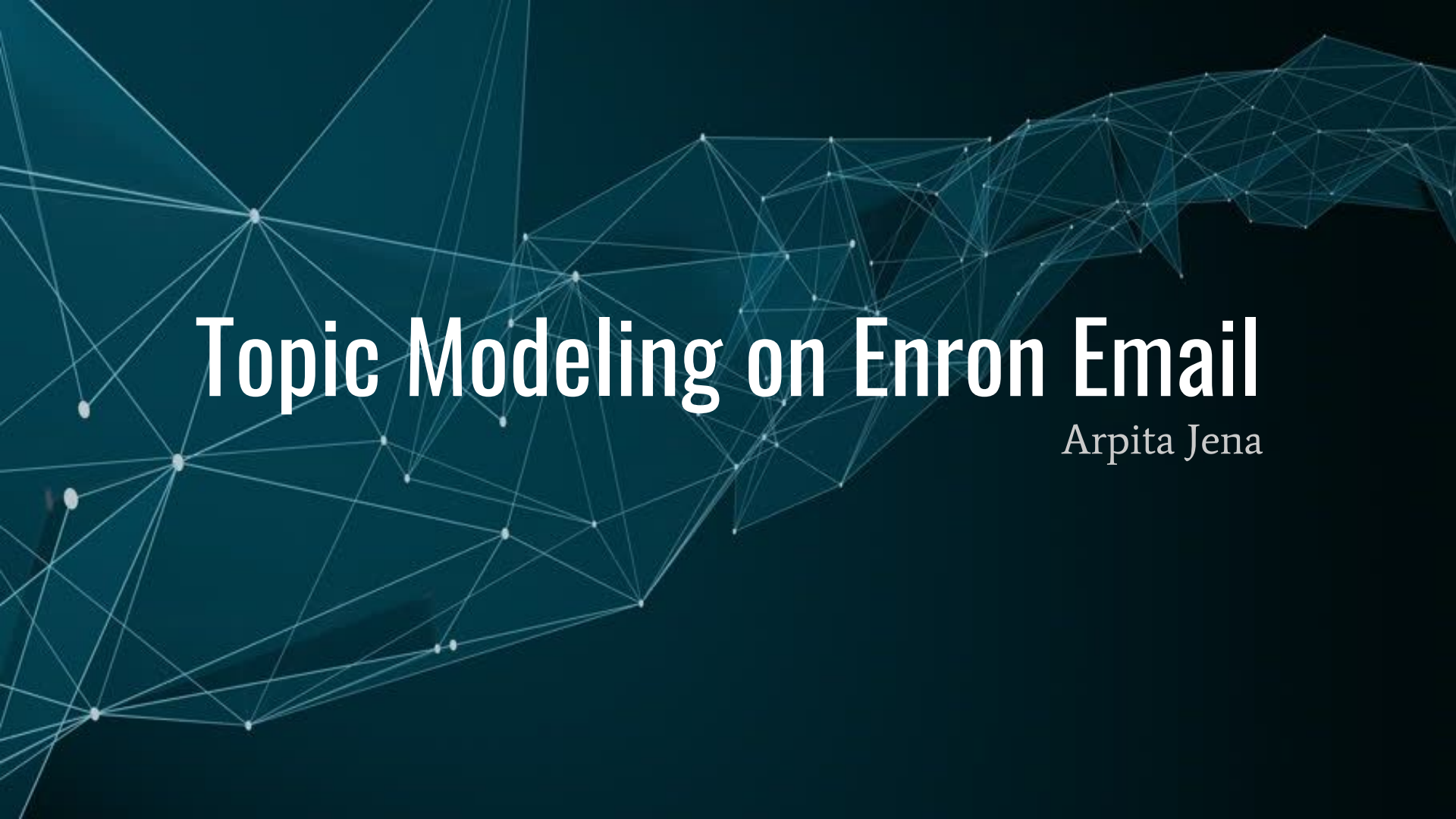


Topic Modeling on Enron Email

Arpita Jena

Data Overview

I have utilized the data from https://www.kaggle.com/wcukierski/enron-email-dataset

Total emails: 517401

Columns: Date, From, To, Subject, Mime-Version, Content-Type, X-From, X-To, X-cc, X-bcc, X-Folder, X-Origin, X-FileName, content, user

Size of data set: 1.43 GB

Data Cleaning

- Dropping redundant columns
- Extracting fields like from, to and email content
- Regular expression to get rid of html content, generic footer and forwards
- Word tokenization using SpaCy
- Selecting adjectives, nouns, pronouns for topic modeling
- Removing stop words and least frequent words from the data set

Word	Count	Word	Count
enron	321308	gas	102166
power	141105	information	97360
energy	130883	company	96776
time	122499	market	87426
mail	105698	california	78423

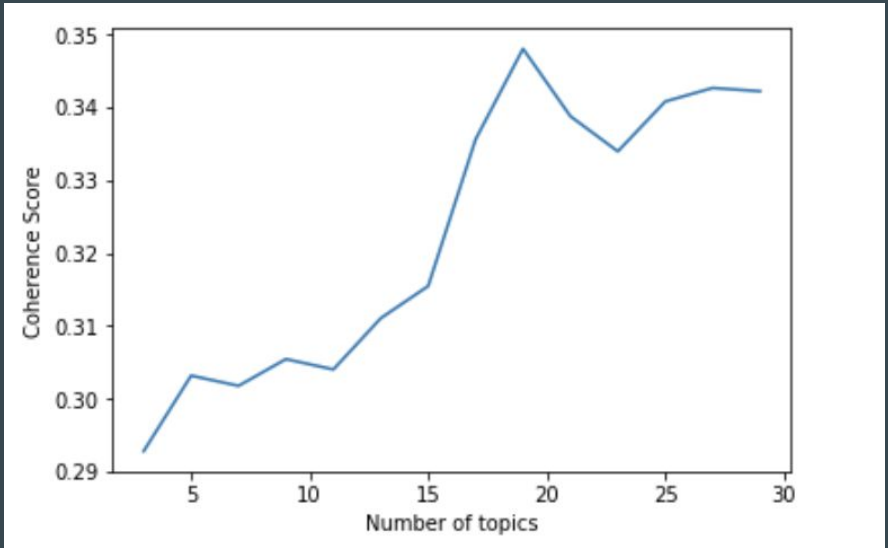
The final data set size reduced to 396 MB after cleaning and all the email contents size was 184 MB.

Topic modeling

Steps involved:

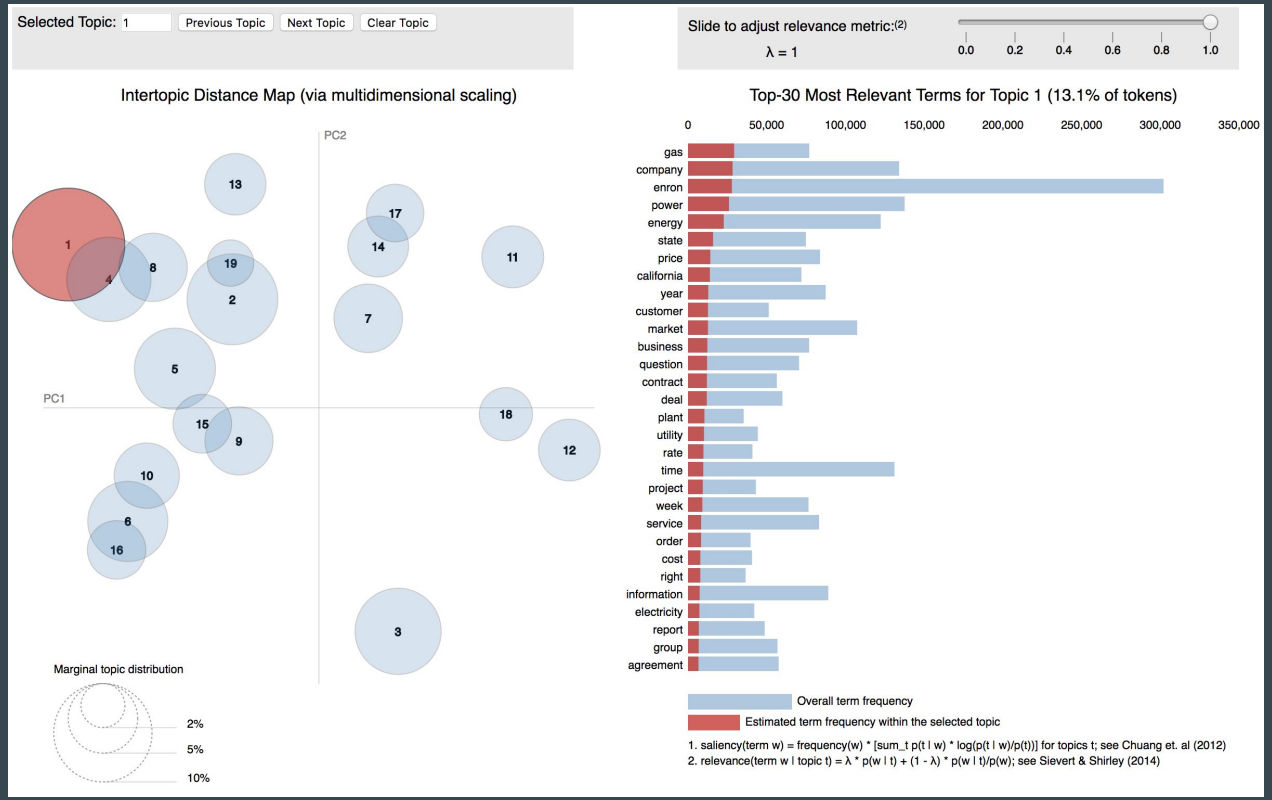
- Creating bigrams
- Document-term matrix
- Perform LDA for different number of topics
- Choosing the optimal number of topics based on coherence score
- Perform tfidf transformation
- Obtain NMF model and fit to the data
- Analysis of topics

LDA: Coherence Score Plot



The optimal number of topics is 19

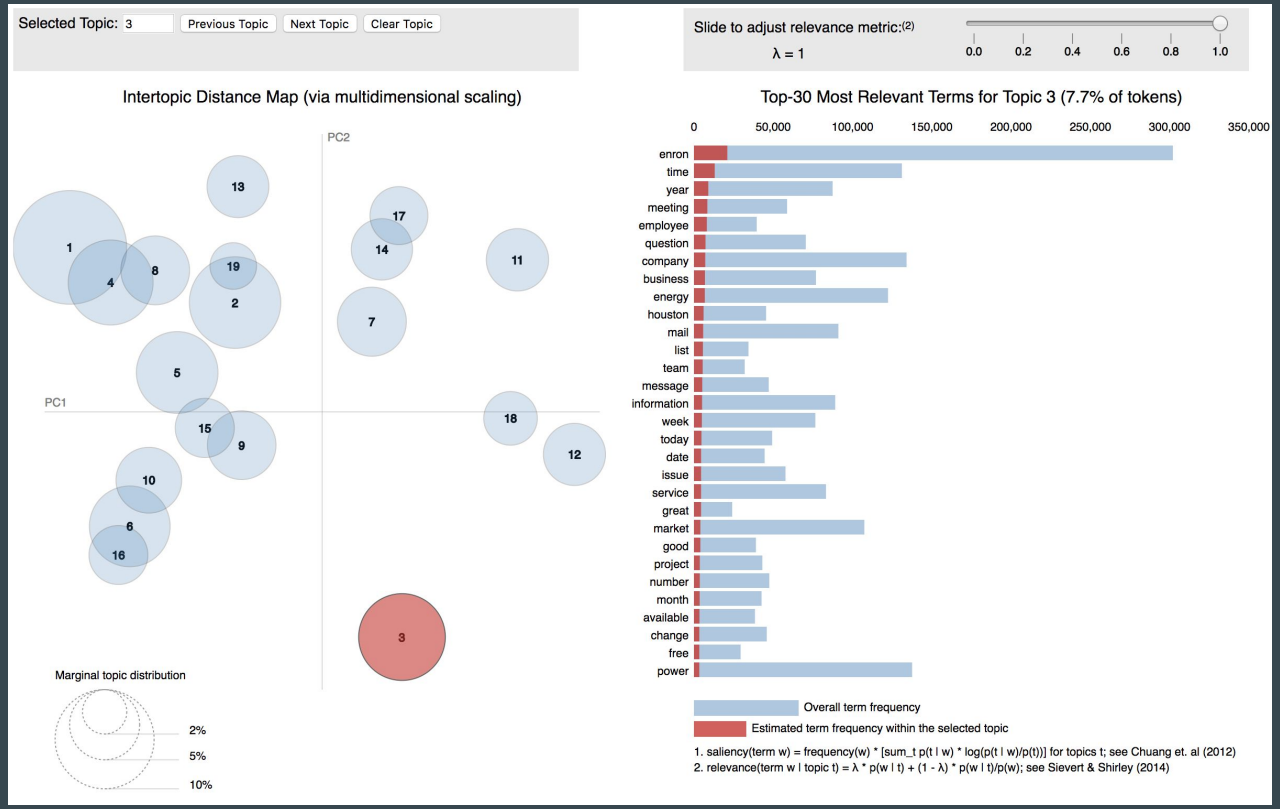
LDA: Topics Visualization



The most important topic seems to be related with California energy crisis.

Fact: A demand supply gap was created by energy companies, mainly Enron, to create an artificial shortage.

LDA: Topics Visualization



Here the topic seems to be mostly related to meeting, questions, issues, business.

LDA: Most common words in topics

Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10	Topic # 11
enron	time	power	enron	enron	enron	enron	database_engine	time	enron	gas
company	power	service	information	energy	information	price	error_borland	market	power	company
power	message	enron	week	mail	power	market	energy	mail	market	enron
energy	enron	company	time	power	time	gas	enron	power	service	power
year	energy	energy	scheduled_outage	service	mail	time	price	week	contract	energy
market	information	agreement	message	information	energy	information	company	energy	trading	state
time	market	time	company	company	email	energy	email	price	business	price
business	mail	information	service	market	california	state	information	enron	year	california
information	price	insufficient_memory	comment	houston	state	year	question	california	agreement	year
corp	group	operation_alias	report	meeting	ferc	business	time	service	california	customer
number	question	market	email	list	market	power	service	company	company	market
agreement	service	contract	group	communication	service	report	deal	information	issue	business
kay_mann	change	price	energy	california	data	company	mail	state	transaction	question
project	meeting	deal	power	group	date	good	market	order	state	contract
electricity	cost	issue	draft	time	change	issue	order	year	energy	deal
issue	company	cost	issue	business	corp	deal	power	today	time	plant
transaction	deal	group	mail	state	report	today	number	rate	gas	utility
question	date	business	question	message	message	natural_gas	utility	gas	change	rate
contract	sure	conference	year	price	company	service	change	deal	price	time
trade	week	data	market	trading	agreement	utility	state	electricity	information	project

Topic 1: Energy issue

Topic 6: California specific issues

Topic 10: Trading and business related

NMF: Most common words in topics

Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10	Topic # 11
schedule	perlingiere	enron	kay	description	database	deal	sara	request	vince	delainey
final	debra	employee	mann	time	alias	kate	shackleton	resource	kaminski	david
variance	smith	america	corp	chairperson	unknown	symes	smith	approval	shirley	regard
hour	america	north	ben	calendar	closed	evelyn	street	srrs	crenshaw	guy
preferred	texas	company	jacoby	entry	error	prebon	houston	auth	stinson	john
hourahead	north	corp	suzanne	detailed	operation	kerri	texas	name	interview	janet
ancillary	street	sally	enron	migration	borland	broker	fax	type	presentation	dave
detail	fax	beck	fred	central	hourahead	peak	north	application	john	rob
log	houston	mark	kathleen	outlook	engine	mike	america	date	resume	milnthorp
file	phone	stock	sheila	appointment	insufficient	sharen	phone	approver	gibner	mark
table	legal	fund	lisa	standard	memory	cason	isda	day	research	lavorato
export	department	service	adam	date	file	volume	cheryl	data	sam	dietrich
import	enron	energy	john	team	manual	trade	enron	read	communication	kevin
date	cook	business	chris	mtg	intervention	amerex	sheila	directory	group	reviewer
message	gisb	communication	roseann	stacey	download	chris	susan	kobra	zimin	calger
load	mary	kean	booth	white	log	metoyer	glover	alternate	vasant	tim
epmi	cordially	steven	carnahan	oncall	final	trader	credit	eol	martin	rodney
ectrt	regard	year	engeldorf	room	hour	number	tanya	email	leppard	forster
tblloads	floor	dynegy	bill	conference	date	kimberly	kaye	backoffice	william	greg
energy	draft	consumer	mitro	invitation	message	counterparty	nelson	requester	christie	eric

Topic 5: Appointment related

Topic 7: Deal and broker

Topic 9: Application and approval

Topic 10: Interview + resume

Conclusion

- The topics generated from LDA and NMF seem significantly different.
- LDA has a high prevalence of energy and power related terms among all the topics but with varied importance.
- Topics generated from NMF have more variance in terms of words present in them.
- People's name across different topics might indicate their activity in those domains.
- The topics need to be thoroughly analyzed to make any conclusion on performance of both the methods.