



Final project

New visualization research or data analysis project

- **Research**: Pose problem, Implement creative solution
- **Data analysis**: Analyze dataset in depth & make a visual explainer

Deliverables

- **Research**: Implementation of solution
- Data analysis/explainer: Article with multiple interactive visualizations
- 6-8 page paper

Schedule

- Project proposal: Wed 2/19
- Design review and feedback: 3/9 and 3/11
- Final presentation: 3/16 (7-9pm) Location: TBD
- Final code and writeup: 3/18 11:59pm

Grading

- Groups of up to 3 people, graded individually
- Clearly report responsibilities of each member

3

Design Feedback (Week 10)

Signup for a 10 min slot

tps://docs.google.com/spreadsheets/d/1BtXmbQHrC3-chPT6kKS51Q-2p9XhbiM3Qct0N847yPM/edit?usp=sharing

- M 3/9 4-6pm
- T 3/10 7-8pm (SCPD only)
- W 3/11 4-6pm

Plan to give a 5 min presentation (mostly demo) of work so far. We will give oral feedback.

Final Presentation

M Mar 16 7-10pm, Location TBD

- Short presentation (5 min, mostly demo)
- Make sure there is time for questions



Text as data

Documents

Articles, books and novels Computer programs E-mails, web pages, blogs Tags, comments

Collection of documents

Messages (e-mail, blogs, tags, comments Social networks (personal profiles) Academic collaborations (publications)

_		1125	6	5/2 Jan
	R		ro Ply	Poter
AMI Reserve Gospion date		rezage (HTML) - M		A.
Provent Dents Hone to Code: Dents - Dents - Dents - Dents Hone to Code: Dents - Dents	Crispota Net a Image	des - 10 - Sents Conder	H	
deroethridioto on Auser Trote de		Set 3x4820842	a 23	
noerder Briter am Alf Rasers Boostor date 45 200 4500 PM				THE A
1999 - House AI J Jan J Johnson Markowski, Sandar J Markawski, Sandar Markawski, Sa	T CLR 1.1.4(22) MET G anh 2, MS-RTC LM () a Number (relev, Number (es aprices, Number (rel)	LR 25.5027, MIT CLR 15.84506.50, MIT CLR Belle Ballet, Def walter, Natherf source fairt, Boolean for farmer, Cancer Tabel View reder or		
aar at System Number Sning ToNamber/Sning At, NamberSy inall at System Number Paraelast11/Sning 1, NamberStyles at MuCalculat_Click(Object sender, EventAcgs e) in E-DATA1	DBExceptionEnabler Sa	npleWith ExceptionSample ups. vb line 8		
eer et Soam Nandes Saing FilonalerSteing en SanderSt and et Soem Nandes Saing FilonalerSteing et SanderStein er beschlinder, Gläd/Hjert ender, FormArp ej is E DATAT	DBExceptionEnadlerSe	nghê Nêvê Exerquê xwî dine 8		125
en et forsta hande Stepf Challer/Step et , Nanderfen in et forsta Nander Hondersberg viel i Education in Calader, Challoffers welte, Bennetyn et i Education Calader, Challoffers welte, Bennetyn et i Education Education I Frenchung Machine Die Gelf Bennethause Machine Di	che pet ges Google (* recente beber Jason Chaang	ng def kan Exception Sample ange verslage 3 - N. S. C. (m. R. (m	p laged 2	8.5% 0.0 5 0 0
en et forse have have been for a for a second of the secon	den Die Des Google (2) Frende Johns Jacon Chuang Walt John (1)	ngelifik Europia Sangta aga Afrika I - • • • • • • • • • • • • • • • • • • •	p Lope 20	
 S) 	ne på gå Coogle () rende blev Jacon Chaang Mail unde (R Best bilevation	ngdrife Ecoperationgel age video 1 • • • ≠ € (a to (g towe) best there in the state interes •	p lager 22	alas Oro o R A
er dens hand hand bland bland bland skiller. Na bland frei far den sen skale hand bland skiller i stander in dense skale bland skiller i stander in dense skiller i stander i far den skiller i stander i	des pels gels Coogle (2) Fronds Johns Jacon Chaung Wall John (2) Bels Information - Strindiz	ng diff & Engelinkangk age Alard - % # (Leo (Leon) Annohang Mar Annohang Mar Annoh	p kopt fo	Alartan Adartan Liaking br a tarb
ra func handhalan shadhalan shadhala	ne pa ge Coople () rende balan Josen Charles () rende balan Josen Charles Martin Labo (R) Beis belanden Martin - Martin	ngentin Europein kangel opp of daret "	p Lupet 20	All 5
s den beder half de der helen i se de	den peter Congle (*) formen beer State Congle (*) formen beer State Change Wall tabe (*) Beek bilanstilen - State State State State State State St	elefti Experience aprovider l'	p topa to	All S O - (All S All S Al
ra frank holden hand frank holden hand holden holde	An an ge Coople () An ge Coople () And the Coopl	The second secon	p kept in	ADIA O - O Adrese Liables track Adrese Dictor
• Comparison • Comparison<	na para Congle (2) read para Congle (2) read para Congle (2) read balanti Succession (2) Succession (2) Success	And the Comparison of the set of	p 1007 P	AD 5 0 -
The second secon	tan an ya Courde (Torus block Standard (Standard) Standard (Standard) Standard) Standard (Standard) Standard) Standard (Standard) Standard) Standard (Standard) Standard) Standard) Standard (Standard) Sta		to the lateral second s	ADD 8 Or 0 Add 8 Add
и сви во молосни различи и сви р На на	Carling Courses and an angle C		bitratio	All A
standarden skalanden skalande Skalanden skalanden skal Skalanden skalanden skala	ne ge Couge () rink ge Couge () rink ge Couge () Storm the Storm Storm the Storm Heat Heat () Des Heat () Heat Heat () Des Heat () Heat () He		r 1997 Z	ALC A ALC ALC A ALC ALC ALC A ALC ALC ALC ALC ALC ALC ALC ALC ALC ALC
и сви во молосорани страновани с молосорани с трановани с молосорани с м С молосорани с молос	An		p tost 2	ALL & A
status balancia da la desarra da la desar la desarra da la desarra d desarra da la desarra da la desa Regiona da la desarra da	Antipological de la factoria de la consectiona de la consectión de la cons		p tost 2	LAR S
s de la desarrada de la desarrada Estarrada de la desarrada d	Construction of the second secon		2 loss 2	LT I I I I I I I I I I I I I I I I I I I
standard de la construcción de l	An and a second		t loss	ACION CONTRACTOR
status balancia da la compositiva da la compos	Control Marcola Control Contro Control Control Control		2 tost 2	ALL S
	An		2 these of the second s	ALL IS ALL ALL ALL ALL ALL ALL ALL ALL ALL AL
	Construction of the second secon		hieron	ALL O THE ALL AND ALL



Why Visualize Text?

Understanding: get the "gist" of a document

Grouping: cluster for overview or classification

Compare: compare document collections, or inspect evolution of collection over time

Correlate: compare patterns in text to those in other data, e.g., correlate with social network

9

Example: Health Care Reform

Background

Initiatives by President Clinton Overhaul by President Obama

Text data

News articles Speech transcriptions Legal documents

What questions might you want to answer? What visualizations might help?

A Concrete Example

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of yo







WordTree: Word Sequences





Gulf of Evaluation

Many (most?) text visualizations do not represent text directly. They represent the output of a **language model** (word counts, word sequences, etc.)

Can you interpret the visualization?

How well does it convey the properties of the model?

Do you trust the model?

How does the model enable us to reason about the text?

Text Visualization Challenges

High Dimensionality

Where possible use text to represent text... ... which terms are the most descriptive?

Context & Semantics

Provide relevant context to aid understanding Show (or provide access to) the source text

Modeling Abstraction

Determine your analysis task Understand abstraction of your language models Match analysis task with appropriate tools and models

17

Topics

Text as Data Visualizing Document Content Visualizing Conversation Document Collections



Words as nominal data?

High dimensional (10,000+)

More than equality tests

- Correlations: Hong Kong, San Francisco, Bay Area
- Order: April, February, January, June, March, May
- Membership: Tennis, Running, Swimming, Hiking, Piano
- Hierarchy, antonyms & synonyms, entities, ...

Words have meanings and relations

Text Processing Pipeline

Tokenization

Segment text into terms. Remove stop words? *a*, *an*, *the*, *of*, *to*, *be* Numbers and symbols? #cardinal, @Staanford, OMG!!!!!!!! Entities? Palo Alto, O'Connor, U.S.A.

22

Text Processing Pipeline

Tokenization

Segment text into terms. Remove stop words? a, an, the, of, to, be Numbers and symbols? #cardinal, @Stanford, OMG!!!!!!!! Entities? Palo Alto, O'Connor, U.S.A.

Stemming

Group together different forms of a word. Porter stemmer? visualization(s), visualize(s), visually -> visual Lemmatization? goes, went, gone -> go

Text Processing Pipeline

Tokenization

Segment text into terms. Remove stop words? a, an, the, of, to, be Numbers and symbols? #cardinal, @Stanford, OMG!!!!!!!! Entities? Palo Alto, O'Connor, U.S.A.

Stemming

Group together different forms of a word. Porter stemmer? visualization(s), visualize(s), visually -> visual Lemmatization? goes, went, gone -> go

Ordered list of terms

24

The Bag of Words Model

Ignore ordering relationships within the text

A document \approx vector of term weights

Each term corresponds to a dimension (10,000+) Each value represents the relevance

For example, simple term counts

Aggregate into a document x term matrix

Document vector space model

Document x Term matrix

Each document is a vector of term weights Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0









Tag Clouds

Strengths

Can help with gisting and initial query formation

Weaknesses

Sub-optimal visual encoding (size vs. position) Inaccurate size encoding (long words are bigger) May not facilitate comparison (unstable layout) Term frequency may not be meaningful Does not show the structure of the text



Keyword Weighting

Term Frequency

tf_{td} = count(t) in d Can take log frequency: log(1 + tf_{td}) Can normalize to show proportion: tf_{td} / Σ_t tf_{td}



Keyword Weighting

Term Frequency tf_{td} = count(t) in d

TF.IDF: Term Freq by Inverse Document Freq $ff.idf_{td} = log(1 + ff_{td}) \times log(N/df_t)$

 $df_t = # docs containing t; N = # of docs$



How do people describe text?

Asked 69 graduate students to read and describe dissertation abstracts

Each given 3 documents in sequence; summarized each using keypharases, then summarized the 3 together as a whole using keyphrases

Were matched to both familiar and unfamiliar topics; topical diversity within a collection was varied systematically





















Yelp: Review Spotlight [Yatani 2011] 109 amazing around baked bar bass best chef delicious eat b) best sf sure in striped bass best sushi baked sea bass other person slow service fresh fish sushi bar more hour baked mussel exp sushi chef only thing sake long time good food sushi restaurant long wait tabl hawaiian roll long line reasonable price baked mango small place delicious everything Mentioned 63 times ^ possess sage of the halos wisdom , and know in advance sushi zone only accepts cash and the waits will be long and arduous . yes , its a long wait , learn the master of zen if you want to eat here .

Tips: Descriptive KeyphrasesUnderstand the limitations of your language modelBag of words:Bay to computeSingle wordsLoss of word orderingSelect appropriate model and visualizationGenerate longer, more meaningful phrasesAdjective-noun word pairs for reviewsShow keyphrases within source text



Information Retrieval

Search for documents Match query string with documents Visualization to **contextualize results**

Google scholar acronym resolution Search Advanced Scholar Search	My Cit	tations
Scholar Articles and patents 💽 anytime 💽 include citations 💽 🔀 Create email alert	Results 1 - 10 of about 63,200. (0.10	0 sec)
A supervised learning appropriate to action multichilication Johan D - Twang The Spherent Constants - 2005 - multico Chilicitat recence or ca Recently the fields of Genetics and Medicine have become specially interested in acromym resolution (Pustopolicy et al., 2001) - vel al. 2002, Pustopolicy et al. acromym resolution sechnicipe searches for definitions of acromyms within noun phrases	(PDF) from nrc-onc.gc.ca	
<u>Biomedical term mapping databases</u> 10 Wen, JI Chaeg, J Putsipnika,	HTMLIfom nih.cox Find it@Stanford	
Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution	Find it@Stanford	
Tockel 4. Globinu Climate Dynamics, 2003 - Springer The long simulations CC and CS are split into two 30-year datasets CC1 and CS1 for the period 1960-1969 and CC2 and CS2 for the period 2017-2099 Full name Acronym Resolution Period Coupled Control CC TS3 1950-2099 Yes Cited by 137 - Related and clines - ALE Swareson Ende by 137 - Related and clines - ALE Swareson		
Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. MS Tutle, NE Olson, KD Keck, VIG Cole Methods of information, 1999 - ukpric ac.uk Tiele not supplied (PMID: 1656483). Concept definition and manipulation are supported through as of tools with functionality such as acromym resolution, word completion, term completion, appling correction, display of the authoritative form of the term entered by the user fuel by 56. Feldeta and tuncials - Cached - Trini diffSathord - R.D.Mert - Ad Stravisions		
PPP A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901–2000) and 16 scenarios (2001–2100)	[PDF] from ipcc-data.org	







Concordance

What is the common local context of a term?









Recurrent themes in speech





Glimpses of structure

Concordances show local, repeated structure But what about other types of patterns?

For example

Lexical: <A> at Syntactic: <Noun> <Verb> <Object>

68

Phrase Nets [van Ham 2009]

Look for specific linking patterns in the text: 'A and B', 'A at B', 'A of B', etc Could be output of regexp or parser

Visualize extracted patterns in a node-link view

 $\mathsf{Occurrences} \rightarrow \mathsf{Node size}$

Pattern position → Edge direction

Darker color \rightarrow higher ratio of out-edges to in-edges















Visualizing Conversation

Many dimensions to consider:

Who (senders, receivers) What (the content of communication) When (temporal patterns)

Interesting cross-products:

What x When → Topic "Zeitgeist" Who x Who → Social network Who x Who x What x When → Information flow





















Named Entity Recognition

Identify and classify named entities in text:

John Smith \rightarrow PERSON Soviet Union \rightarrow COUNTRY 353 Serra St \rightarrow ADDRESS (555) 721-4312 \rightarrow PHONE NUMBER

Entity relations: how do the entities relate?

Simple approach: do they co-occur in small window of text?



Parallel Tag Clouds [Collins 09]

adverted	adjourned	allocatur	adequate	8	Midavit	about	abuse	abuse	appeal	ballot	accused	agency
ala	alia	analysis	affirmed	bankruptcy	bargarring	argand	advited	aliens	argument	banc	agency	agency
anent	allocution	antitrust	aid	barge	Case .	asked	amrmed	appropriate	attached	black	annuity	agency s
appellant	arbitration	app	aiu	cargo	cocaine	called	appellee	asylum	binding	boat	antidumping	authority
appellant's	asbestos	arbitration	appeal	charter		cocaine	argued	asylum	brief	candidate	application	bargaining
annellee	closure	assets	appelee	coverage	coun	cocame	beleve (c)	circuit	cited	case	art	brief
apponde	commenced	bankruptcy	argument	annen	defendant	100 Martinety	oocaine wool	cited	collateral	class	board	broadcast
asseveration	conveniens.	believe	before	dest	defendant's	defendant	denied	contended	copy	commerce	claim	capricious
boat	copyright	benefit	coal	deblor	atest	daciplinary	disability	courts courts	anter defender.	country	compounds	carrier
brief	date	bottlers	cocaine	drilling	denied	enough	distribution	don declared	determine	county	construction	competition
official states	defendant		conspiracy	and an and a second sec	district	fire	district	denial	disfavor	death	contract	costs
commonwealth	disenfranchised	class	contention	gas	and the	get	drug	deportation	doc	desagragation	combine data	disposition
defendant	foreign	context.	court's	habeas	employees	gun	evidence	discretion	doctrine	disenfranchised	decision	emissions
dei	forcigit	creditors	crack	homestead	filed	had	farm	disposition	estoppel	donos	description	employees
ensued	Tard	debtor	decisional	indemnity	firearm	harassing	grams	district	examination	alaction	device	explanatory
event	heroin	exercise	denied	instant	follows	have	had	errs	forthwith	electors	embodied	facilities
ferritin	injunction	fiduciary	disclosed	insurance	grievance	her	her	except	furnished	immunities	equivalent	gas
guidelines	inter	have	dispensed	mourance	hereby	Tier	hie	fear	further	in the second	1010 1011	interpretation
here	internal	here	dstribution	jurists	nis	nim	impair	fish	graces be	ivary	infringement	intervenor
incarcerative	keeplock	inasmuch	district	law.	job	his	inmates	habitat	judgment	jail	Inningement	labor
interiore	marks	insurance	drug emor	liability	judgment	job	jury	his	judicata	law	invalid	Icense
jury		interest	fact	loan	magistrate	judge	medical	immigration	material	migrant	invention	memoranda
limned	millions	Jurisdiction	from	marihuana	marijuana	kilogramusa	othomaphotom	jurisdiction	nevertheless	mitigation	inventor	petitioner
Ist	narcotics	legislation	interlocutory	maritime	redical	launor	emaniprietan	IIIIe land	opinion	norelab.dory	layer	ninelines
might	ingines	liability	joined	mitigation	motion	might	noffice	methamphetamine	oral	ordinance	inetest iterat	preceding
mortgage	plaintiff	majority	legal	negligence	office	more	opinion	novo	ordor	payday	means	promulgated
plausible	plaintiff's	market	magistrate	nre	panel	one	nain	TIOVO.	order	phase	merchandise	proposed
point	principal	notes	material	oil	paupers	estrich	postconviction	panel	persuasive	preceding	noninfringement	rate
print rescript.	quotation	our	merits	parish	plaintiff	para	pravia processor and star	persecution	plaintiff's	qualified race	obion	Tate
601/	racketeering	Jarien	miner's	nlatform	plaintiff's	police	reversed	petition	precedential	racial	patent	regulations
Say	revisurance	plaintiff	mining	plation	pneumoconiosis	prisoner	search	patient	record	section	patentee	rehearing
see	respect	plainun	opinion	recovery	police	cho	sentence	provided	res	sentence	product	reprinted
some	600	pian	oral	ref'd	pulmonary	SIIC	sexual	public	submitted	sheriff	protein	rulemaking
supra	sharsholders	plenary	order	removed	pursuant	suit	she	pursuant	suspended	etudente	reissue	section
think	shares	policy	pneumoconiosis	rigging	recommendation	supra	subd	sector	ten Dereture 59	tion to	retirement	see
tit	sterile	provision	present	seaman	search	tentative	Supervise	cuitable	unanimaun	turtie Boot log	said	service
town	subway	recognized	process	stated	sentence	thought	lestined	Suitable	Containing the last	vessel	signal	snipper
trialworthy	summation	section	published	suit	sitting	told	taid	tribal	untavorable	vote	specific	tariff
vessel	trade	settlement	pulmonary	usury	unanimous	tije	trial	tribe	unpublished	voters	structure	transmission
viz	vacated	sytup	sentence	vessel	union	want	unpublished	unanimous	value		surface	union
whom	waybill	which	which we are a set of the set of	writ	upon	when who	vendict	water	vehicle	white	vaccination	united
whom	where	would	wrote	whit	warrant	would	work	without	vol	zone	veterans	waste
First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth	Eleventh	Federal	DC

107

Topic modeling

Topic modeling approaches

Assume documents are a mixture of topics Topics are (roughly) a set of co-occurring terms Latent Semantic Analysis (LSA): reduce term matrix Latent Dirichlet Allocation (LDA): statistical model













Summary

High Dimensionality

Where possible use text to represent text... ... which terms are the most descriptive?

Context & Semantics

Provide relevant context to aid understanding. Show (or provide access to) the source text.

Modeling Abstraction

Understand abstraction of your language models. Match analysis task with appropriate tools & models.

Currently: from bag-of-words to *vector space embeddings*