Topic analysis using Mallet and network graphs

Rob McDaniel
Senior Machine Learning Engineer
PayScale

About me

robm@payscale.com

github.com/robmcdan

https://www.linkedin.com/in/robmcdan

previous projects:

- Microsoft
- Lingistic.com

About PayScale

- using topics to categorize documents
- entity and semantic extraction from job descriptions
- topics (among other things) tell us if “accounting” is being used in the context of finance, or if it’s common use
- PayScale’s very different from Glassdoor and similar companies

Overview

what I’m going to cover:

semantics

topic models and how they work

how to use mallet

measuring topic interaction

high-level, code available for off-line analysis

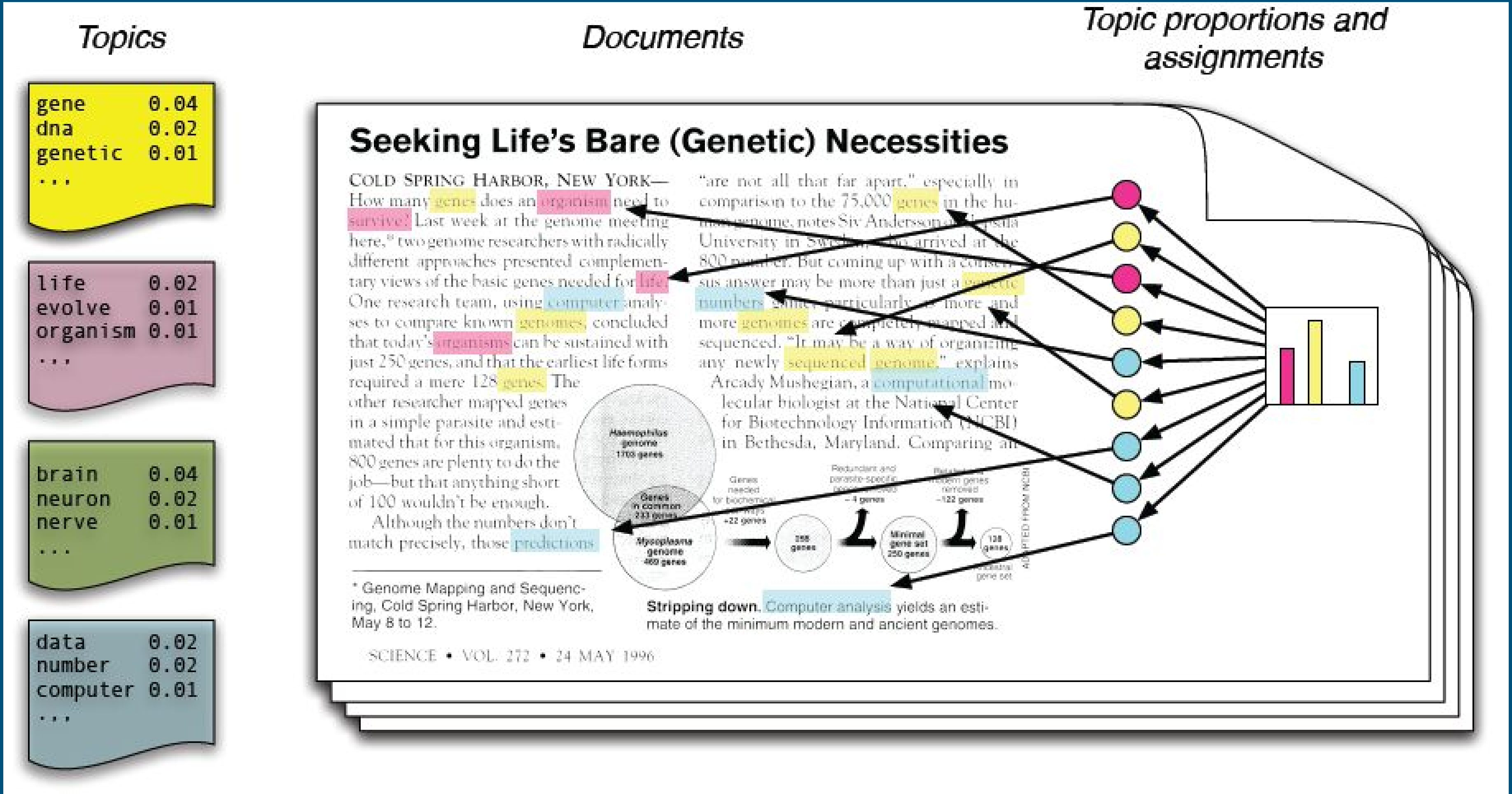
what are semantics

semantics: deals with meaning

the relationships of words together form the semantics

what is a topic model

- unsupervised; discovers themes in unstructured text
- bag of words model
- generative model
- can be thought of as a clustering algorithm
- topics: distributions over words
- document: distribution of topics


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

history of LDA

originally used for finding patterns in genetic data

highly useful in today’s world of big data

many implementations available

Steps of topic modelling

vectorize training documents

train

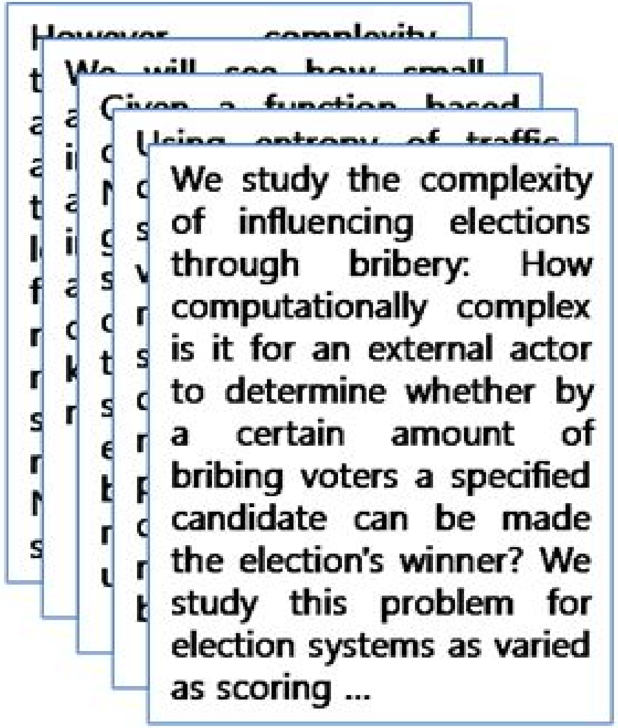
vectorize unseen documents

infer topics

vectorizing

- every document is represented as a numerical vector
- large vocabulary = sparse matrix
- multi-dimensional vector space model

Documents

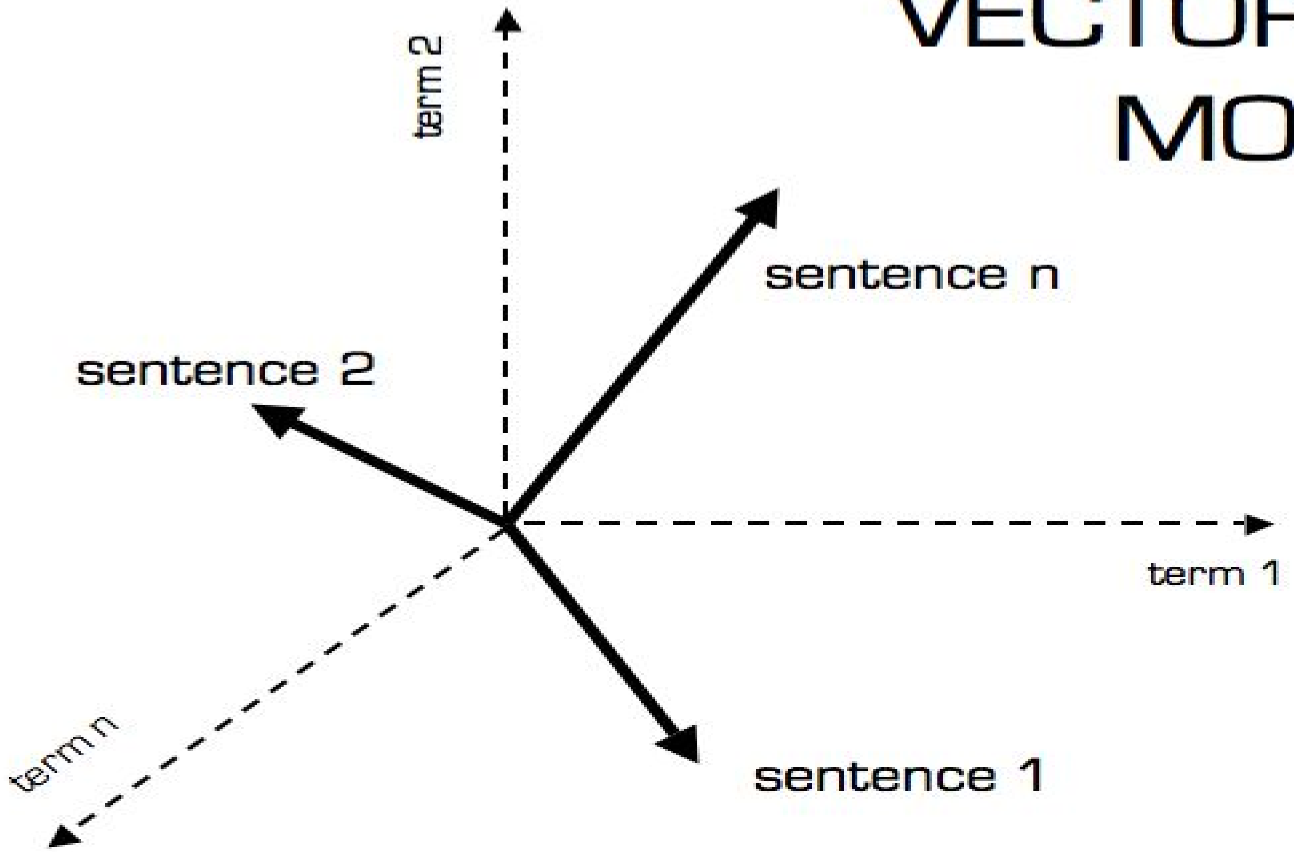


Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

VECTOR SPACE MODEL



training

- training on word sequences
- non-deterministic; set random seed for consistent re-modelling
- must predetermine number of topics

inference

- infer topics from unseen documents
- deterministic
- must use original vocabulary

The corpus

GOP and Democratic debates for 2016 election cycle

http://www.presidency.ucsb.edu/debates.php

Why presidential debates?

- nicely chunked into little context chunklets
- wide variety of speakers and context
- limited vocabulary
- no legal issues around using customer’s data

Prepare the corpus

- garbage in, garbage out
- stop word removal
- isolating key phrases
- parsing relevant items

isolating key phrases

what are ngrams

finding likely ngrams

nltk

- NLTK (natural language tool kit)
- built in collocation measures
- likelihood measure finds ngrams which go together often, based on prior occurance

ngram likelihood

```
#tokenize sentence into words

bigram_measures = nltk.collocations.BigramAssocMeasures()

bigram_finder = BigramCollocationFinder.from_words(words)

bigram_finder.score_ngrams(bigram_measures.likelihood_ratio)
```

Example ngrams by likelihood measure

Secretary_Clinton	1928.6349782078692
United_States	1582.65386804903
Senator_Sanders	1430.490764995243
Senator_Rubio	1109.1112295674834
Wall_Street	1050.378694855774
Senator_Cruz	1000.1855691365586
New_Hampshire	908.0197777658559
President_Obama	788.9489874026619
Governor_Christie	732.7257542072496
Governor_Bush	728.8350811850478
North_Korea	597.3617719755027
Governor_Kasich	589.1031551108515
commercial_break	556.4690710489392
Senator_Paul	547.8797358958528
Hillary_Clinton	486.48080519364464
health_care	460.21407947612664
Donald_Trump	447.3320619557984
bell_rings	439.30608437343244
climate_change	436.45248108215134
Barack_Obama	409.1356822232396
foreign_policy	399.3244418178018
White_House	386.00309121441626
Des_Moines	380.3797449598837
Dana_Bash	349.476587224584
Ronald_Reagan	345.09198162490304
Middle_East	325.27298204530655

why not just nltk?

- ngrams by collocation are neat, but don’t capture semantics
- when words are slightly different, they appear unrelated
- still useful for seeding LDA

replacements and deletions

mallet allows for replacement and deletion of words

example:

New Hampsire -> New_Hampshire

I’m -> <deleted>

why replace?

- allows the inclusion of ngrams into mallet
- topic: “elections, new_hampshire” instead of “elections, new, hampshire”
- remove words we don’t care about

generating topic models using mallet

installing mallet (fork available on my github account which ignores case sensitivity)

Java

./bin/mallet is a wrapper for accessing mallet features

process

Process ngrams (see code sample on github)

build replacement file (see code sample on github)

parse debates (see code sample on github)

vectorize and train lda

predict topics

build graph

import data step

can import either a single file, one example per line (mallet import-file)

or can import a directory of files (mallet import-dir)

train model step

./bin/mallet train-topics

pass sequences file

specify outputs

interpreting the Mallet output files

doc_topics -- the proportion of topics (columns) in each document (rows)

topic_keys -- N words for each topic, to “describe” it

topic_counts -- a count of each topic word and how many times it occurs in each topic

import unseen document

import-file as before, but must use --use-pipe-from flag

vectorizes according to exisiting model vocab

infer topics

infer-topics -- specify input doc, and inferencer file

Putting topics together: interactions

measuring interactions using KL Divergence

- measures the differences in P(W | T) across documents
- captures how often topics occur with other topics
- topics that occur with others must be related
- threshold is important

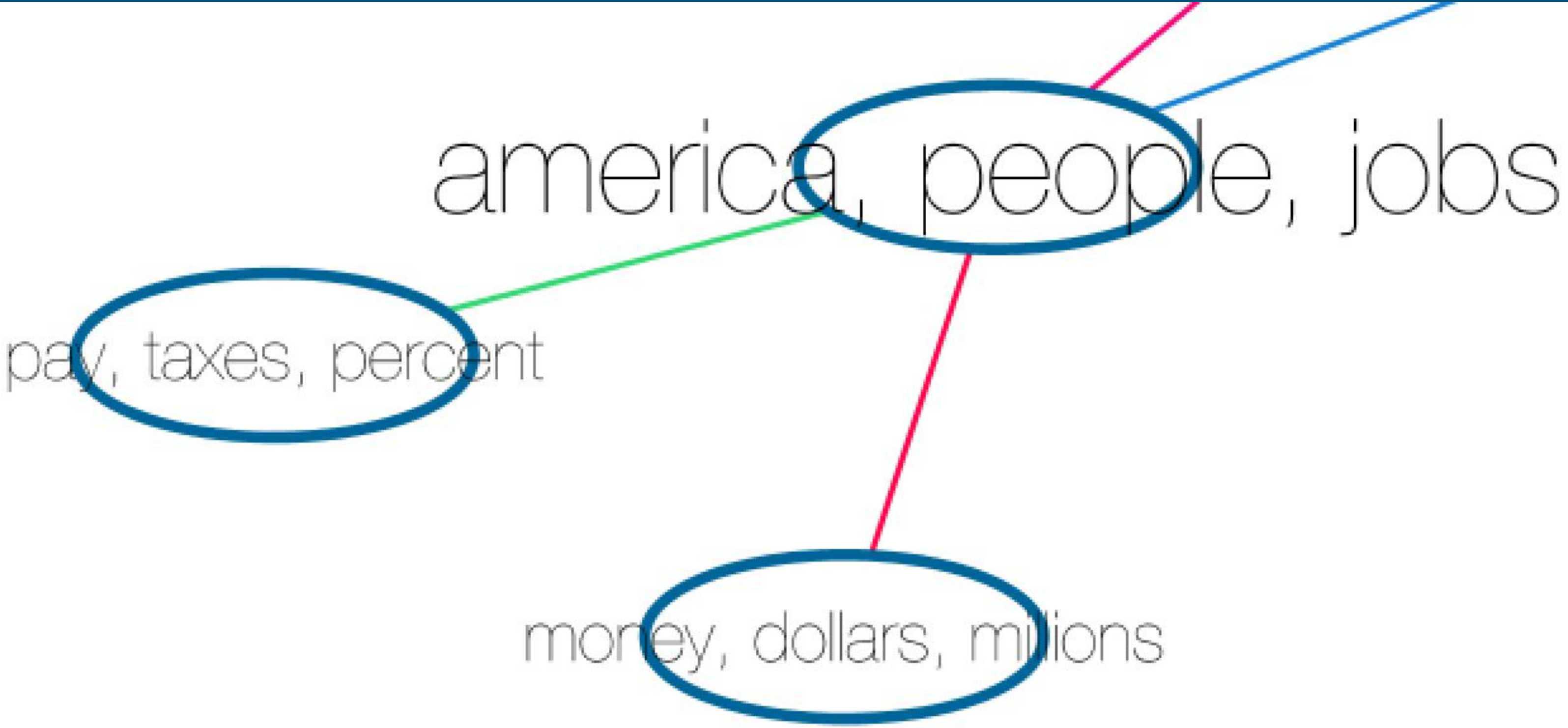
Interactions as networks

Topic -> Node

Divergence -> weight

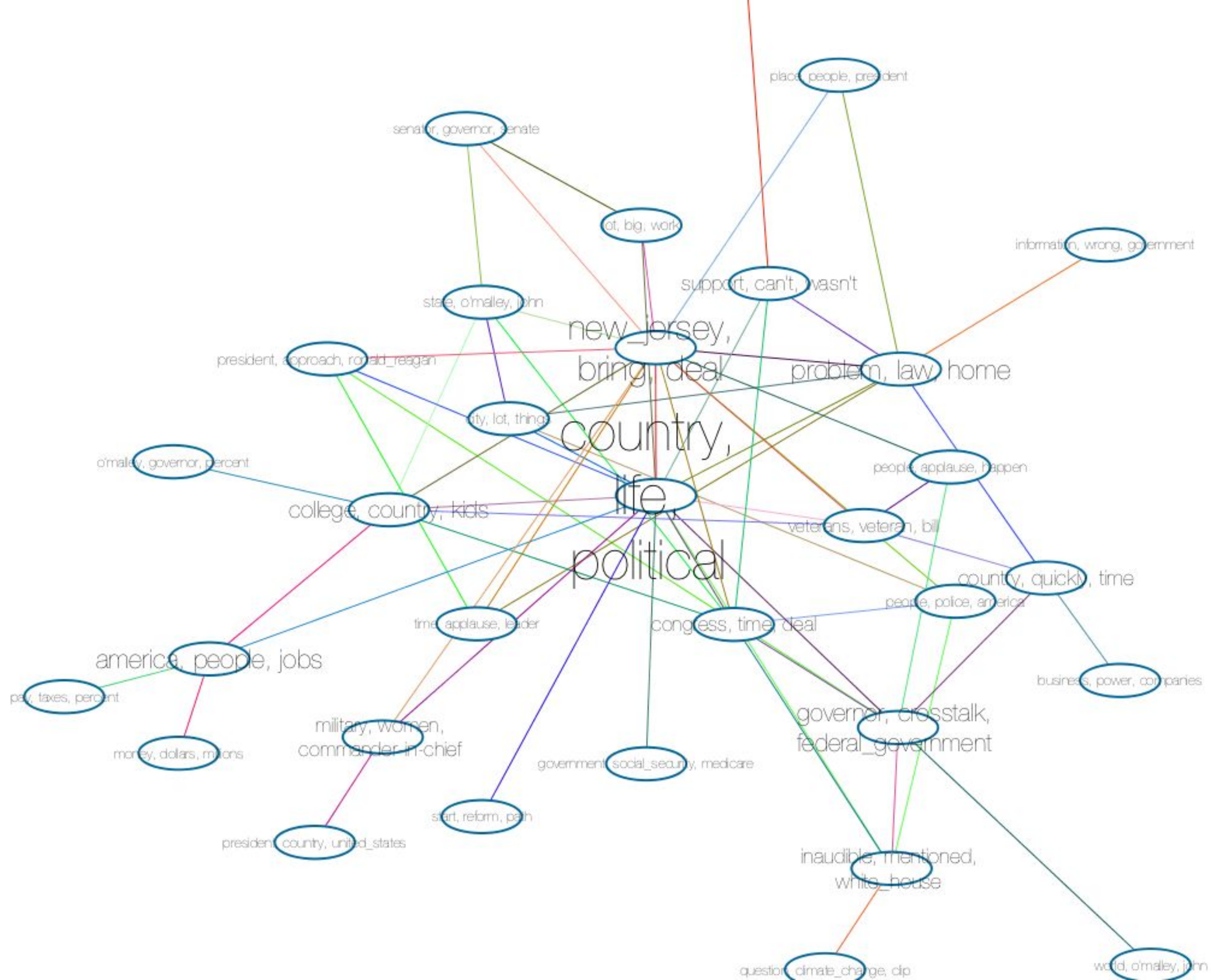
generates undirected network

networkX python package will output to graphml format


america, people, jobs
pay, taxes, percent
money, dollars, millions

Graphing in Cytoscape

- cytoscape -- open source
- popular in bioinformatics
- complex networks
- http://www.cytoscape.org/
- http://diging.github.io/tethne/api/tutorial.mallet.html
-



place, people, president
senator, governor, senate
lot, big, work
information, wrong, government
support, can't, wasn't
state, o'malley, john
new_jersey, bring, deal
president, approach, ronald_reagan
problem, law, home
city, lot, thing
country, life, political
o'malley, governor, percent
people, applause, happen
college, country, kids
veterans, veteran, bill
country, quickly, time
people, police, america
time, applause, leader
congress, time, deal
america, people, jobs
business, power, companies
pay, taxes, percent
governor, crosstalk, federal_government
military, women, commander-in-chief
money, dollars, millions
government, social_security, medicare
start, reform, path
president, country, united_states
inaudible, mentioned, white_house
question, climate_change, clip
world, o'malley, john

Possible improvements

remove noise

spelling correction

better data sampling

Resources

https://github.com/robmcdan/Mallet

https://networkx.github.io/

http://www.cytoscape.org/

example code:

https://github.com/robmcdan/datapalooza

references & resources

http://diging.github.io/tethne/api/tutorial.mallet.html

https://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

http://mimno.infosci.cornell.edu/papers/mimno-semantic-emnlp.pdf

http://mallet.cs.umass.edu/about.php

http://yosinski.com/mlss12/MLSS-2012-Blei-Probabilistic-Topic-Models/