ANNUAL MEETING

November 10-13, 2019
Honolulu, HI

Hilton Hawaiian Village
Waikiki Beach Resort

Text Modeling Even Your Boss Will Understand

Using LDA to get data from free-form text

Bret Shroyer, FCAS
Gross Consulting

Andy Doll, FCAS
Capital Insurance Group

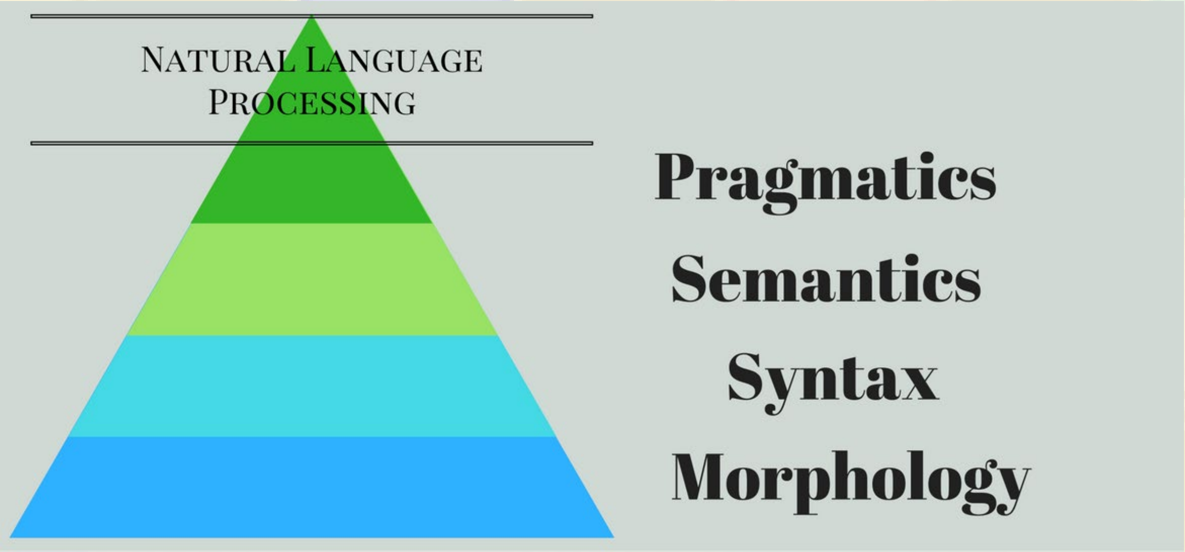

Alexa, tell me a joke...

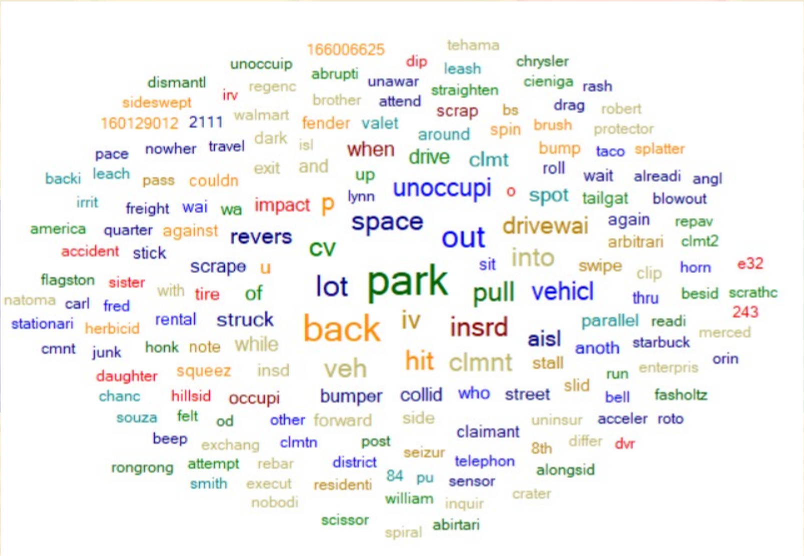
Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Text Analysis Comes in Flavors

- Natural Language Processing (NLP) is a general term describing Text Mining and Text Analysis
- Two broad categories of NLP:
  - Semantic analysis (Alexa)
  - Content analysis (Keywords, Latent Topics)





Agenda:

- Describe LDA Topic Modeling via case study
- Define LDA Topic Modeling concepts
- Discuss ideal candidates for LDA Topic Modeling
- Hands-on exercise: Build a Topic Model using LDA and Excel

- You can play long with the exercise, download the workbook at http://www.cgconsult.com/CAS2019

Topic Modeling Case Study

- Capital Insurance Group
- Challenge: Identify text in claim notes that help predict whether or not a claim will go into litigation
- Data: Three years of claim information as of 90 days
  - Typical exposure, policy, claim fields
  - Claim paid amounts at 90 days
  - Free-Form Text: Short claim description (255 char)
  - Free-Form Text: Full text of claim adjuster notes (2000+ char)

Topic Modeling Case Study

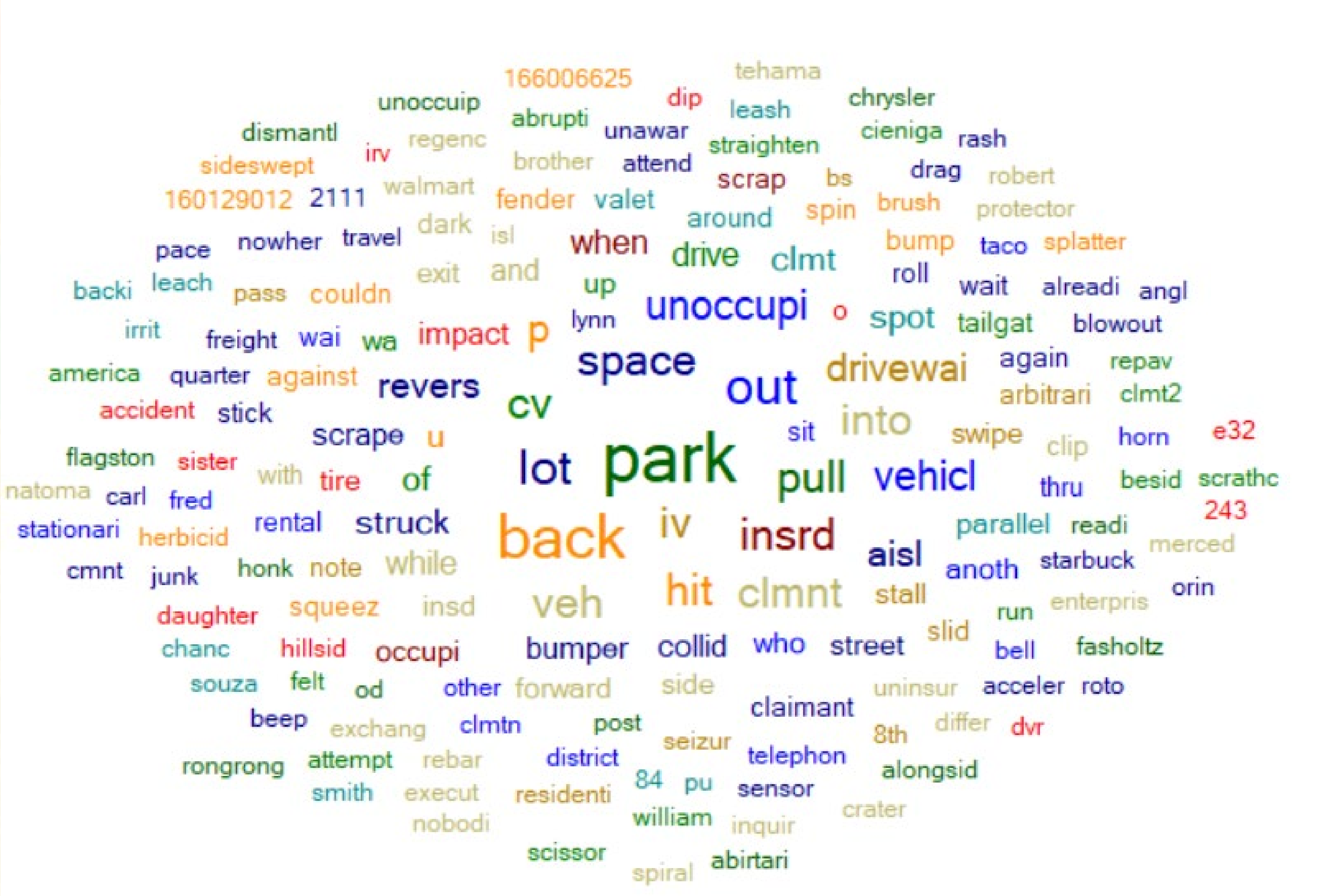
The Challenge:

Can we predict which claims will be litigated?

What data is predictive?

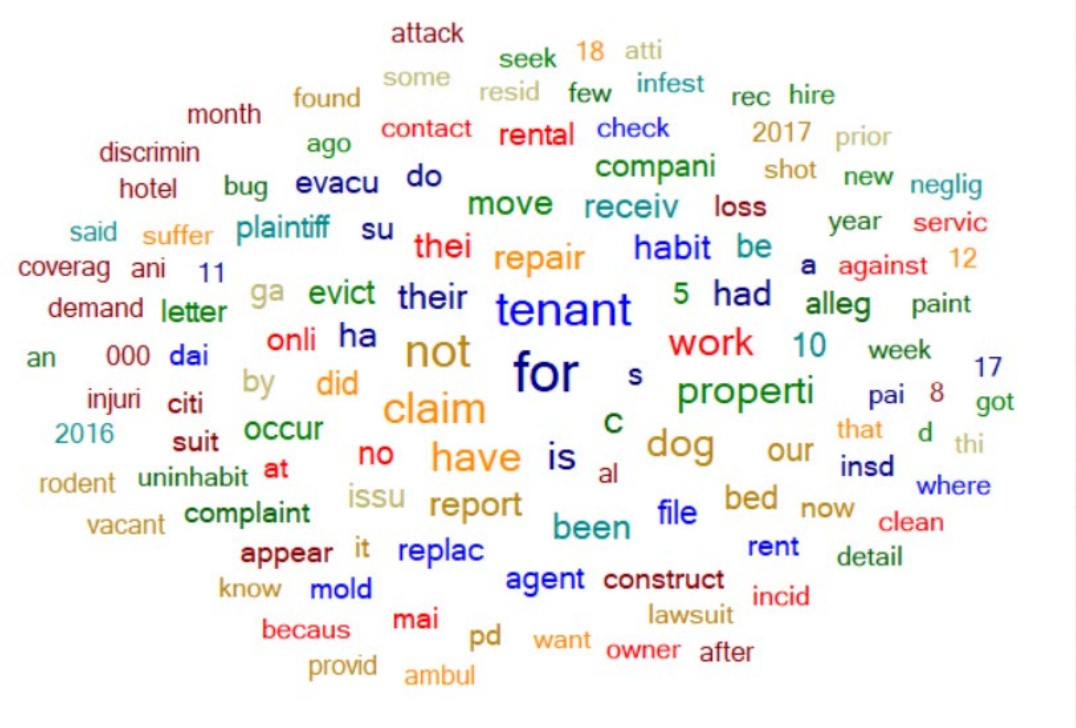
Topic Modeling: Short Topic 8

Topic Modeling: Short Topic 5

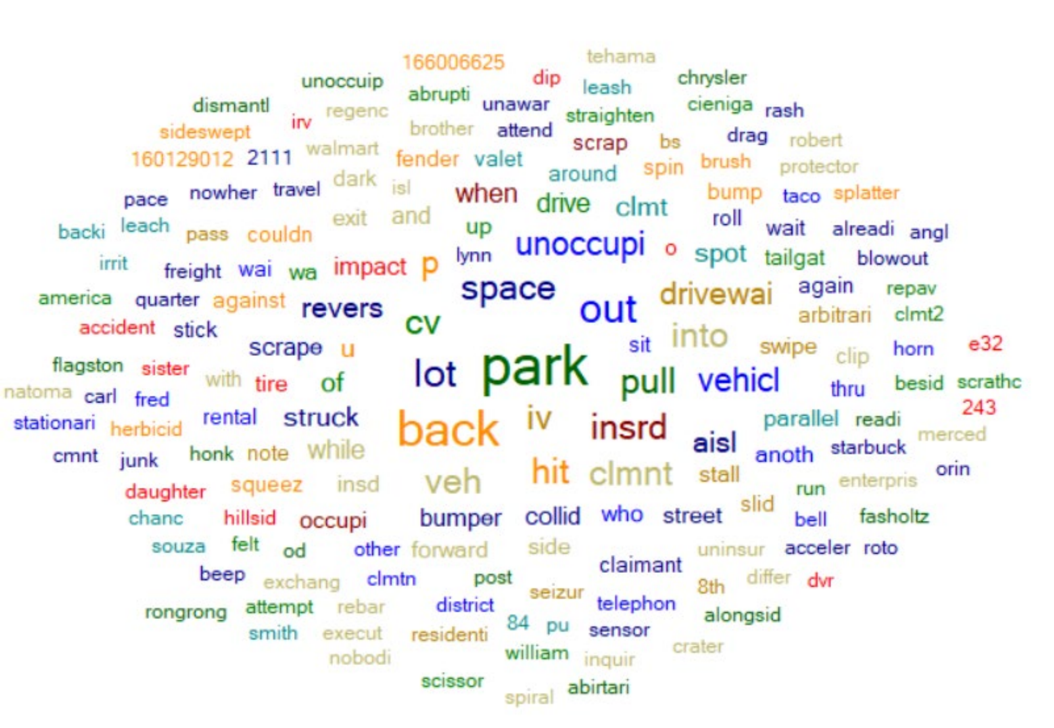


Topic Modeling Illustration: Topic 8

TOPIC 8



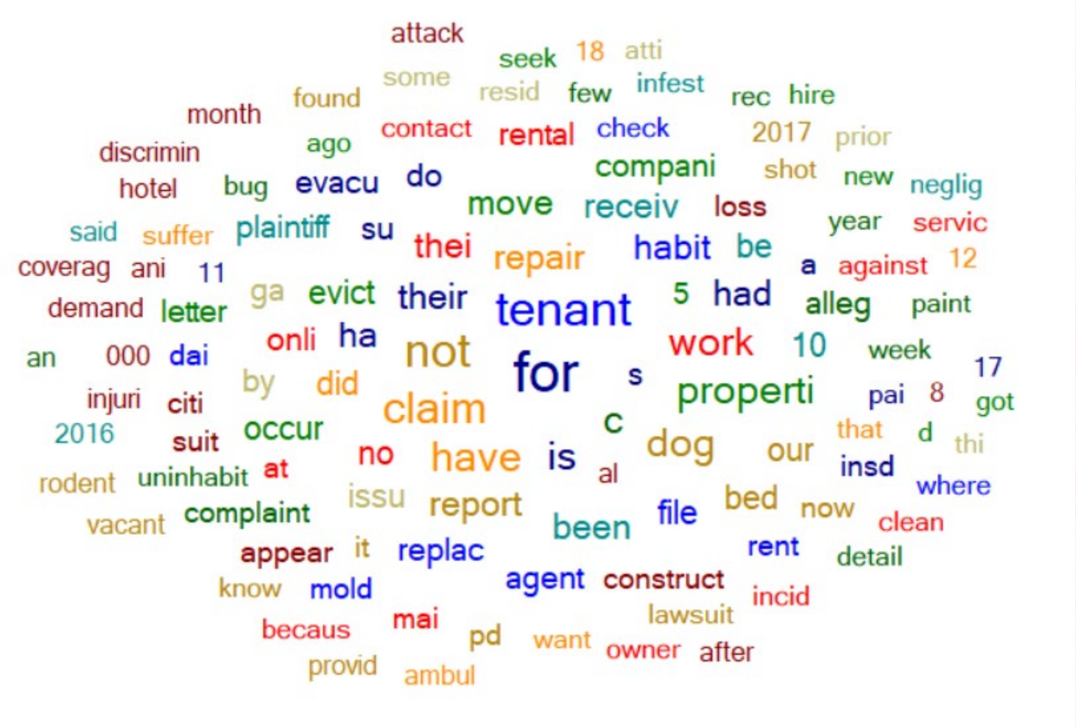
TOPIC 5



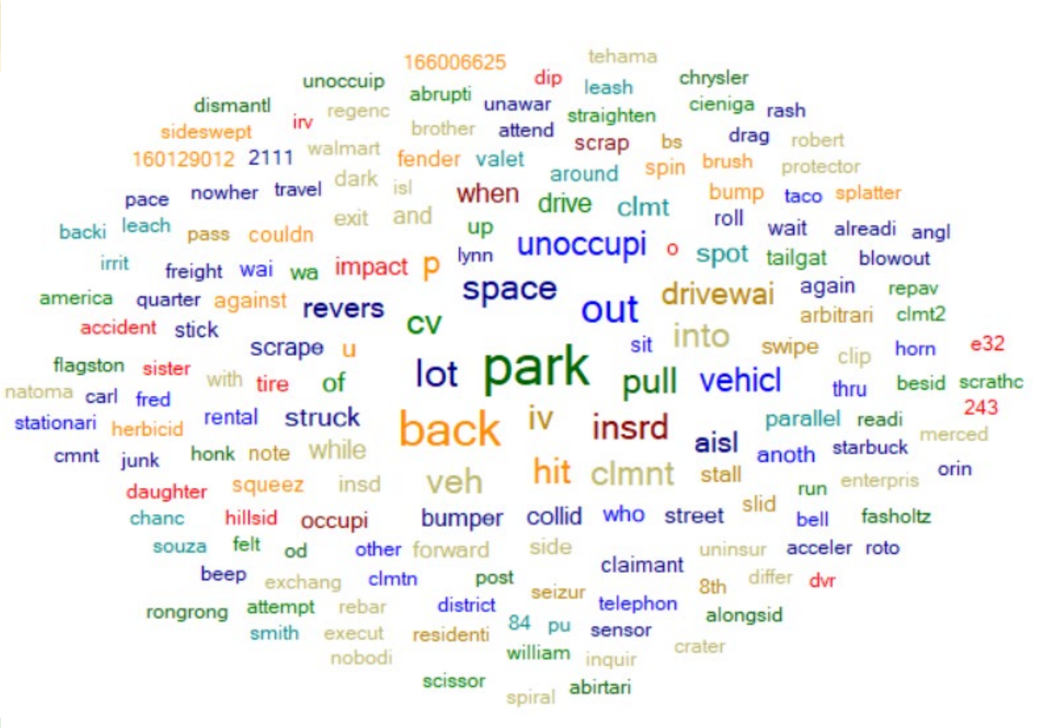
- Tenant is claiming inhabitability of property and lost wages due to health problems caused by landlord having bldg. retrofitted for EQ.

Topic Modeling Illustration: Topic 5

TOPIC 8



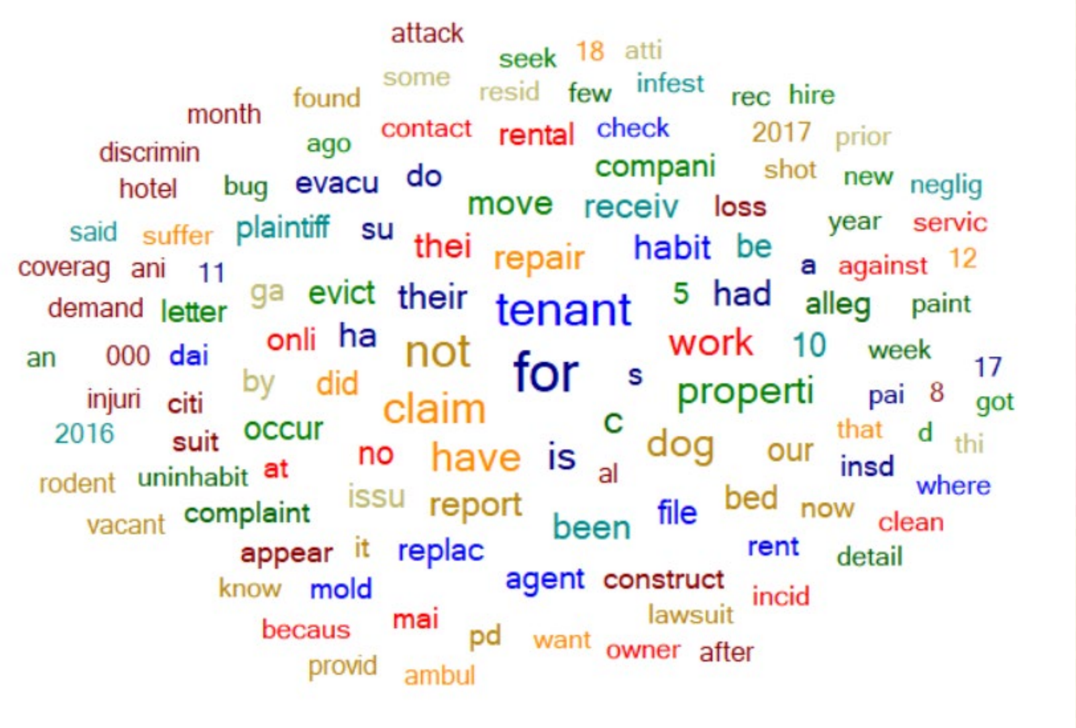
TOPIC 5



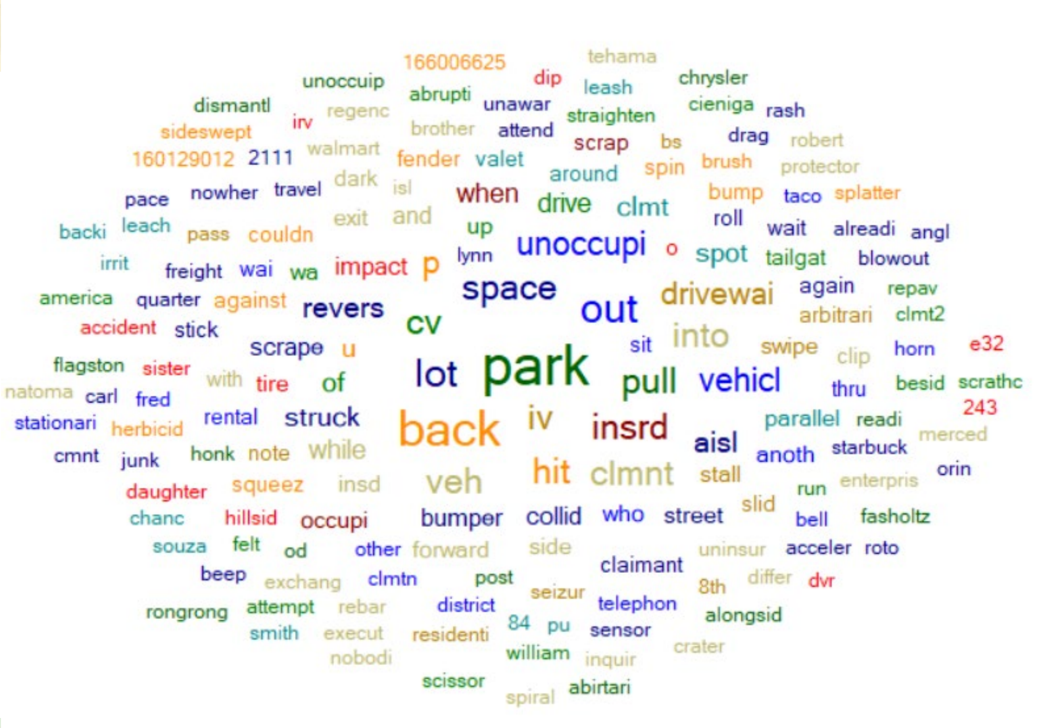
- IV pulling into parking lot of B&B and hit parked and unoccupied claimant vehicle.

Topic Modeling Illustration: Both Topics

TOPIC 8



TOPIC 5



- Both Calvin & claimant were backing out of spaces at same time. Veh`s collided. Claimant has ALOT of pre-existing/PRIOR damage. Claimant not willing to accept 50/50.

Topic Modeling Case Study

Short Text Topic	True Positives	Topic Frequency	% of Lit_CLM identified	Positive Rate
5	**77**	**11,616**	**3.7%**	**0.7%**
1	94	12,254	4.5%	0.8%
7	61	6,706	2.9%	0.9%
2	112	7,771	5.4%	1.4%
4	155	10,463	7.4%	1.5%
9	186	10,139	8.9%	1.8%
10	227	10,303	10.9%	2.2%
3	183	7,899	8.8%	2.3%
8	**1,060**	**7,501**	**50.7%**	**14.1%**
6	855	5,508	40.9%	15.5%
	2,090	Total Litigated Claims		
	102,991	Total claims		
	Overall litigation rate		**2.0%**	

- Key point: Not all topics are related to litigation %. The determination of topic takes place before assessment of whether or not topic is related to an external target variable.
- Topic modeling is solely a function of the aggregate body of text.

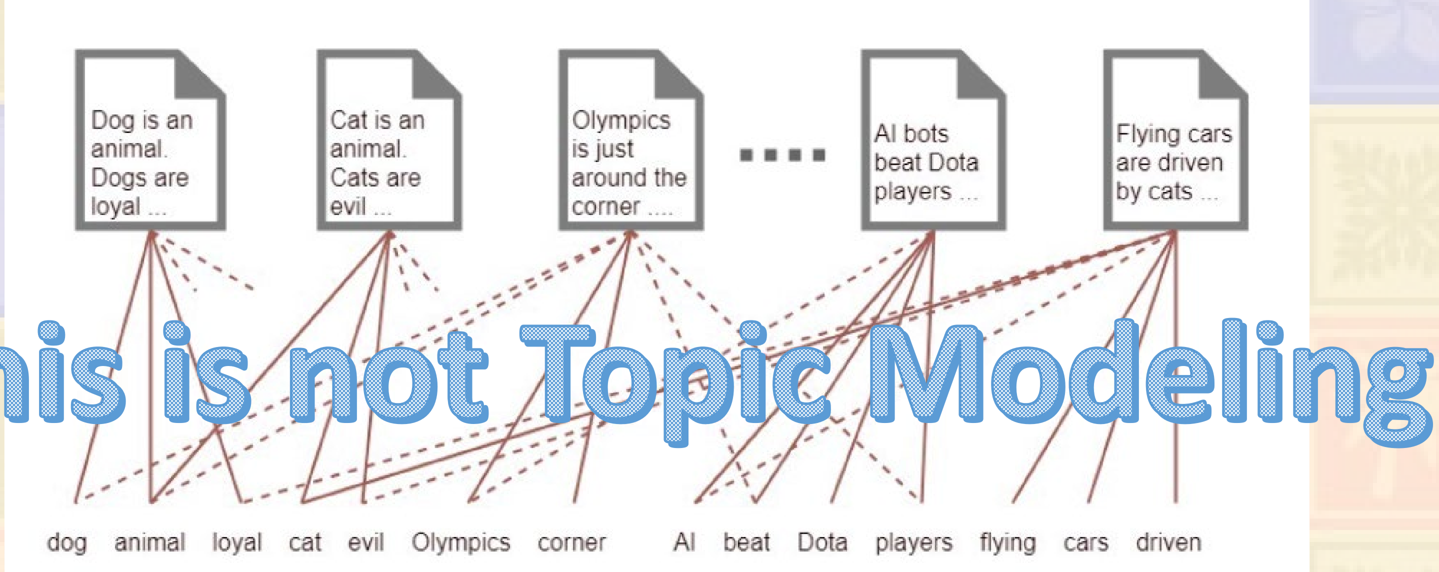
LDA Topic Modeling Terminology

- Document, record, composite: A collection of Words
  - “The OV collided with IV”
- Word, n-gram, part: The smallest unit of text
  - “The”, “OV”, “collided”, “with”, “IV”
- Corpus: A matrix of the frequency of each word by document
  (order of words is not important, this is not semantic analysis)

	Document				
Word	**1**	**2**	**3**	**n**	**…**
The	1			1	…
OV		1			…
collid	1	1			…
other			1		…
vehicle	2		1		…
..	…	…	…	…	…

LDA Topic Modeling Terminology

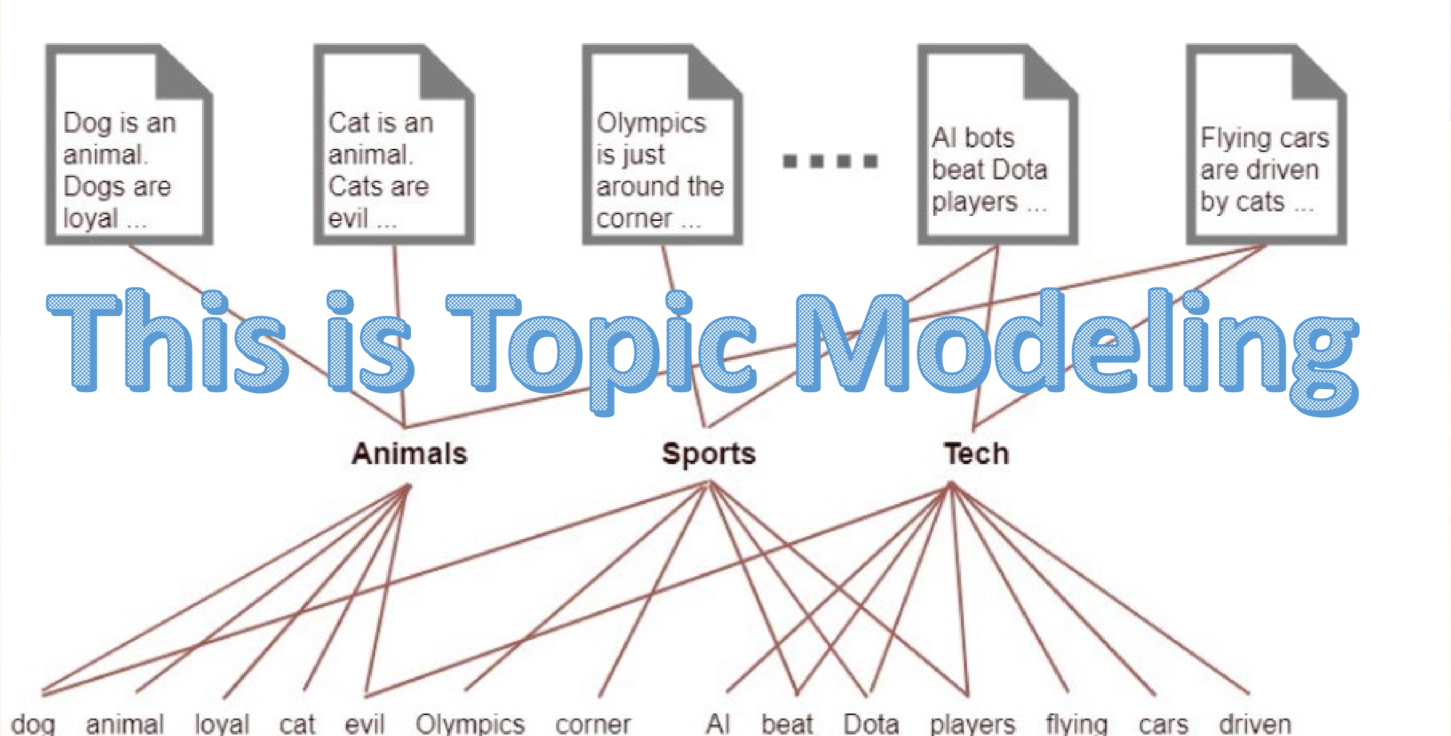
- LDA: Latent Dirichlet Allocation
  - Latent: the topics are hidden, or unknown
  - Dirichlet: a probability distribution; the conjugate prior of the categorical and multinomial distributions



Modeling documents just with words. You can see that we can't really infer any useful information due to the large amount of connections

LDA Topic Modeling Terminology

- LDA: Latent Dirichlet Allocation
  - Latent: the topics are hidden, or unknown
  - Dirichlet: a probability distribution



LDA Topic Modeling Terminology

- Phi: array of word scores by topic
- Theta: array of document probabilities by topic
- Alpha, beta: hyperparameters used in building the topic model in range (0,inf)
  - **Alpha: Document concentration**(how many documents can be in each topic?)
  - **Beta: Topic concentration** (how many words define each topic)
  - For alpha, beta = 0, most documents will not map to any topic
  - For alpha, beta >10, most documents will map to most topics
  - **Suggestion: select alpha = 0.5, beta = 0.1** as a starting point and look at the number of non-trivial documents and words

Pop Quiz: LDA Concepts

1. Definition of Topic
2. Definition of Corpus
3. Definition of Document
4. I’m getting all of my documents assigned to most topics with equal probability. What do I do?

Where can Topic Modeling be useful?

In general, look for

- Documents with multiple words
- Many distinct words
- Significant word overlap among documents

Good Topic Modeling targets:

Description of loss

- Documents with multiple words
- Many distinct words
- Significant word overlap among documents

IV collided with OV at four-way stop intersection. No injuries.

Good Topic Modeling targets:

Claim notes

- Documents with multiple words
- Many distinct words
- Significant word overlap among documents

Spoke with ID; she complains of pain In lower back, has scheduled medical appointment for 9/16

Good Topic Modeling targets:

Description of insured operations

- Documents with multiple words
- Many distinct words
- Significant word overlap among documents

ACME Manufacturing is the leading supplier of hydraulic and pneumatic controls in Georgia and SC

Poor Topic Modeling targets:

Injured Body Part

- ~~Documents with multiple words~~
- Many distinct words
- ~~Significant word overlap among documents~~

Left Arm

Poor Topic Modeling targets:

Email signature disclaimers

- Documents with multiple words
- ~~Many distinct words~~
- Significant word overlap among documents

Caution: This email originated from outside the organization. Do not click links or open attachments…

Survey: Will This Work for Topic Modeling? Why or why not?

- Class Code Description
- ICD9 / ICD10 Description
- Text from the insured's website or twitter feed
- Name of Insured
- Agency Name
- Poll audience for two more examples to survey, or use
  - A list of expense payment amounts / payees
  - Full street address
- Question to audience: anyone want to volunteer any ideas not yet discussed for reaction?

Hands-on: Building a Topic Model

- Repeat: You can play long with the exercise, download the workbook at http://www.cgconsult.com/CAS2019
- Open Excel workbook

Hands-on: Building a Topic Model in R

- Repeat: You can play long with the exercise, download the R scripts and data at http://www.cgconsult.com/CAS2019
- Open R Studio

Final Quiz

- How confident are you that you could identify text list that would be suitable for Topic Modeling? (1-5)
- How confident are you that you could transform list of claim descriptions into Topic scores? (scale of 1-5)
- If someone else developed a Topic Model and scored some text, how confident are you that you could explain/interpret Topic Model scores to your boss? (scale of 1-5)