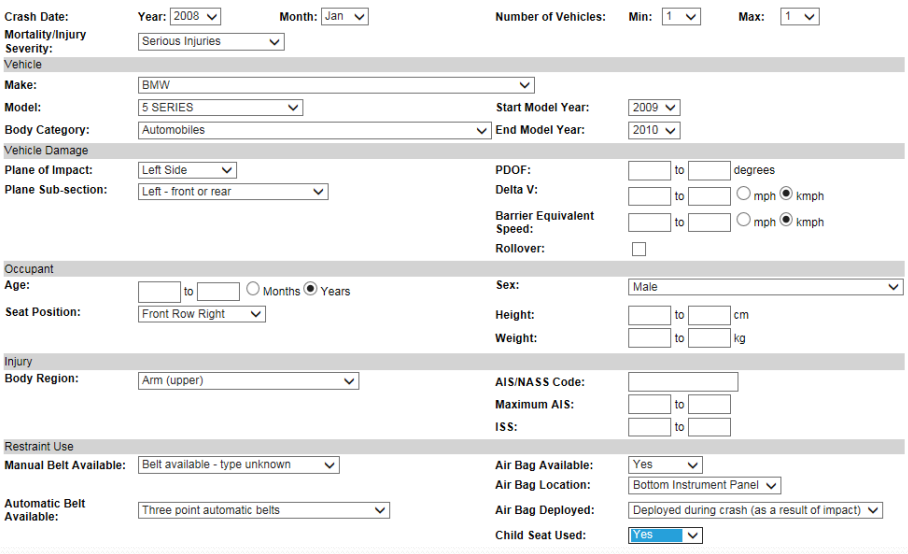
Nonnegative Matrix Factorization on Vehicle Crash Summaries

Pilot defense
Ye Sun
October 31, 2014

NASS database



http://www-nass.nhtsa.dot.gov/nass/cds/SearchForm.aspx

Narrative Summary

Crash Overview - Summary

Crash

PSU	2
Case Number	003
Stratum	G
Weight	686.145
Crash Date	01/2008
Day Of Week	Tuesday
Crash Time	11:30

Case Summary

Crash Type	Vehicle to Object(s)
Configuration	Rollover
Summary	V1 traveling east when it lost control on snow/ice covered roadway and went off the right side of the roadway. V1 continued and struck a metal guardrail end with its front. V1 continued up the guardrail and rolled over 3/4 turns to the right down a small embankment and came to rest facing east on its left side.

Vehicles

Vehicle	Year	Make	Model	Damage Plane	Severity
1	2007	HYUNDAI	ACCENT	Unknown	Unknown

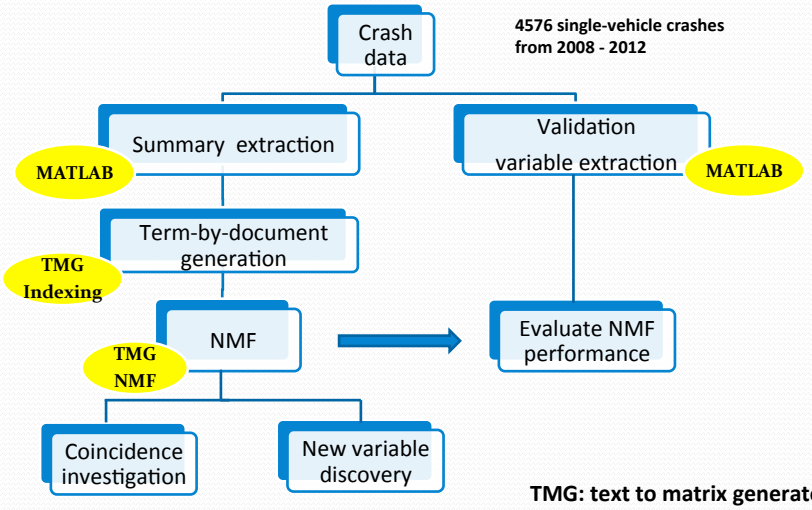
Persons

Vehicle	Occupant	Role	Seat	Restraints	Max Injury	Max Severity	Injury Source
1	1	Driver	Front-Left	Air Bag/ Manual-Used		Unknown	

Objectives

1. Use nonnegative matrix factorization (NMF) to extract topics from narrative summaries of 4576 single-vehicle crashes in 2008 - 2012.
2. Evaluate the performance of NMF for topic extraction.
3. Develop new variables from the narrative summaries that are not available in the NASS database.

Methods



Construction of Term-by-document Matrix

Vector space model:

1. Terms: extracted from documents to form dictionary

- Remove 455 stopwords, min length 2 chars, max length 30 chars.
- Min local and global frequency =1

2. Term-by-document matrix $A\downarrow m \times n = [w\downarrow ij]$

m: number of terms (2218) n: number of documents (4576)

$w\downarrow ij$: the weight associated with term i in document j

3. Log-entropy weighting scheme $w_{ij} = l_{ij}\, g_i d_j$

Local term weighting: $l\downarrow ij = \log(1+f\downarrow ij)$

Global term weighting: $g_i = 1 + \left(\frac{(\sum_j (p_{ij} \log p_{ij})}{\log n} \right)$ $p_{ij} = f_{ij} / \sum_j f_{ij}$

Term-by-document Matrix A 2218 x 4576

Documents (x4576)

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
plane	0.0000	0.0000	0.0000	0.0000	0.0000	0.2209	0.0000	0.1092	0.0000	0.0000
rotate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1963	0.1523
tree	0.0000	0.1449	0.0677	0.0000	0.0000	0.2120	0.0000	0.1048	0.0694	0.0853
road	0.0000	0.0588	0.0690	0.0000	0.0000	0.0000	0.0789	0.0000	0.0000	0.1099

Terms (x2218)

d2: Vehicle 1 was traveling south negotiating a sharp left curve on an upgrade. The driver lost control due to sand and salt on the roadway. V1 went off the right side of the **road**. The front of V1 struck a 9 cm diameter **tree**, which uprooted the tree.

d6: Vehicle #1 was traveling north on an undivided two-way, two-lane roadway. Vehicle #1 departed roadway on the east side and impacted a 25cm **tree** with the frontal end **plane**.

d10: Vehicle 1 was traveling south on an undivided two-lane **road** negotiating a slight right curve. V1 started to **rotate** and went off the right side of the **road**. The vehicle dropped off the edge of the pavement and the front of the vehicle struck a mound of earth. The vehicle deflected to the right slightly and the left side of the vehicle struck a 28 by 31 cm diameter **tree**. V1 came to rest off the right side of the road.

Nonnegative matrix Factorization (NMF)

- $A_{m \times n} \approx W_{m \times k} H_{k \times n}$

  Minimize $\|A - WH\|_F = \sqrt{\sum_{ij} (A_{ij} - (WH)_{ij})^2}$

  where W_{ij} and H_{ij} are nonnegative elements

  $k \ll \min(m, n)$, number of features

- **Advantage:** "Parts-based" representation of the data

  Dimension reduction $A_{2218 \times 4576} \approx W_{2218 \times} H_{10 \times 4576}$

- **W – feature vectors or basis vectors (term x feature)**

  Each column is a feature vector representing one theme

  W_{ij} is the weight of term i in feature j

- **H – coefficient vector (feature x document)**

  Each column shows how a document spans each feature and to what degree

  H_{ij} is the weight of feature i in document j

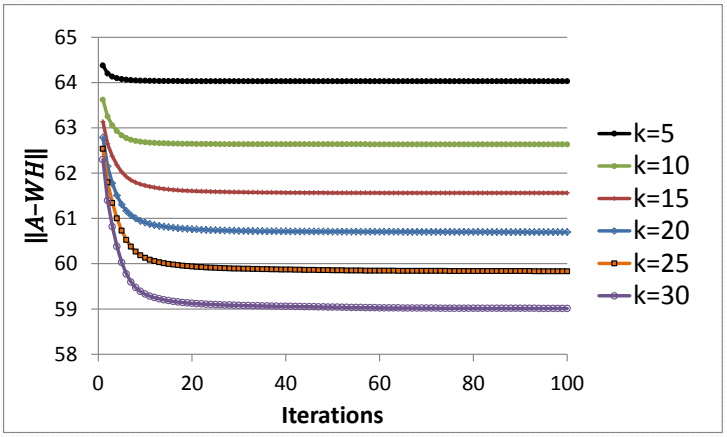
NMF Algorithm

- Initialization:
  Nonnegative Double Singular Value Decomposition
- Multiplicative Update

$$H_{cj} \leftarrow H_{cj} (W^T A)_{cj} / (W^T WH)_{cj}$$
$$W_{ic} \leftarrow W_{ic} (AH^T)_{ic} / (WHH^T)_{ic}$$

- $\|A-WH\|$ is nonincreasing during iterations
- Convergence is achieved if and only if $\|A-WH\|$ is invariant.

NMF convergence – Select k



$A_{2218\times4576} \approx W_{2218\times25}\, H_{25\times4576}$

$W\downarrow 2218\times 25$

Features →

	f1	f2	f3	f4	f5	f6	f7	f8	f9	…	f25
plane	0.3251	0.0551					0.0167		0.1838	…	
rotate	0.0076	0.0626		0.0731				0.0610	0.0131	…	0.0497
tree	0.3736			0.2171					0.1730	…	
road	0.2936		0.0914		0.0417	0.0658		0.0709		…	0.0976

↓ *Terms*

The *top ten dominant terms with the largest magnitude were selected for topics*

Features - Objects

Feature Index	Topic Description	Dominant Terms
1	Tree	contacted, tree, plane, east, road, edge, north, west, two, lane
2	Concrete barrier	barrier, concrete, traffic, lanes, divided, median, trafficway, jersey, crossed, expressway
6	Utility pole	Utility, pole, wooden, struck, roadway, south, road, two, north, front
8	Guardrail	Guardrail, metal, face, contacted, lanes, end, crossed, road, rotate, number
14	Curb, median	Curb, median, raised, tire, wheel, light, center, eastbound, pole, lanes
22	Post sign	Post, sign, fence, wooden, street, continued, metal, striking, mailbox, contacted
23	Ditch	Ditch, undercarriage, culvert, drainage, north, road, driveway, shallow, struck, ground
24	Wall (retaining, concrete, brick)	Wall, retaining, concrete, brick, counterclockwise, stone, center, median, rotated, cement
25	Embankment	Embankment, south, traveled, steep, edge, rock, road, rest, roof, dirt

Features - Road

Feature Index	Topic Description	Dominant Terms
7	Intersection	Intersection, turned, corner, turning, curb, controlled, light, traffic, passed, stop
12	Negotiating curve	Negotiating, curve, left, event, driver, impacting, towed, hand, entered, roadway
15	Turn	Turn, attempted, eastbound, attempting, intersecting, left, quarter, rolled, intersection, hand
16	Highway, divided	highway, divided, interstate, trafficway, elk, non, plane, south, median, unknown
20	Ramp, highway, exit/entrance	Ramp, exit, entrance, highway, attenuator, interchange, freeway, expressway, gore, impact

Features

Feature Index	Topic Description	Dominant Terms
5	Deer	Deer, entered, ran, swerved, avoid, contacted, front, west, path, steered
3	Rollover	Turns, quarter, rolled, tripped, roof, top, wheels, entered, clockwise, rotation
11	Lost control	Lost, control, clockwise, driver, counter, rotated, traction, icy, rotate, wet
17	Northbound	Northbound, end, lane, number, unknown, roadway, expressway, front, street, driver
19	Westbound	Westbound, end, event, roadside, street, elk, final, damage, struck, pole
21	Southbound	Southbound, end, veered, unknown, number, roadside, reasons, front, steel, grass

$H_{\downarrow 25\times 4576}$

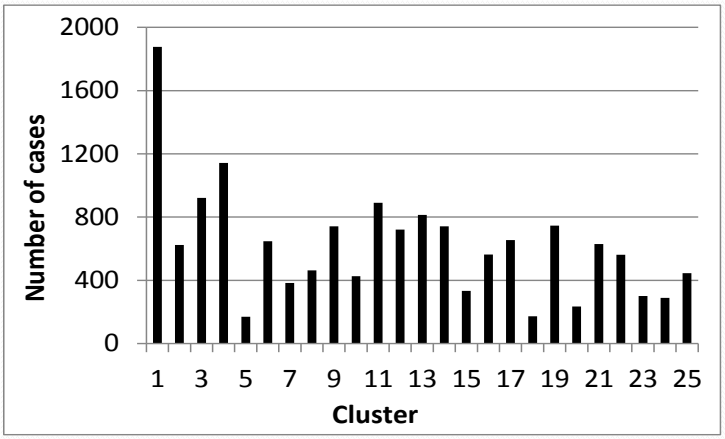
documents →

features ↓

	d1	d2	d3	d4	d5	d6	d7	d8	…
f1	0.0607	0.0692	0.0712	0.0300	0.0485	0.3010	0.0483	0.0950	…
f2									…
f3	0.1484						0.0506		…
f4	0.0901	0.0718	0.1167	0.0823	0.1375		0.0926	0.0149	…
f5		0.0001			0.0091		0.3442	0.2744	…
f6					0.0107		0.0000		…
f7							0.0059		…
f8	0.3689		0.2530						…
f9			0.0001		0.0114	0.3689		0.0319	…
f 10		0.0100		0.0158	0.0072	0.0001		0.0464	f10
…	…	…	…	…	…	…	…	…	…

Cluster *by feature: Each row of H matrix with magnitude greater than rowmax/10* is put into a cluster.

Cluster size



Quality of clusters

- Precision:
  P = R/L
  R – number of relevant documents within the cluster
  L – cluster size
- Recall:
  R=R/T
  R – same as above
  T – number of all relevant documents in the NASS database
- R: compare the crash cases within a specific cluster with the relevant crashes validated in the database

Quality of clusters

Feature index	Topics	Recall	Precision	Average Severity
1	Tree	0.64	0.41	1.88
2	Concrete barrier	0.86	0.62	1.5
3	Rollover	0.78	0.90	1.70
6, 22	Utility pole & Post	0.55	0.68	1.56
8	Guardrail	0.84	0.76	1.29
14	Median curb	0.60	0.60	1.72
23	Ditch	0.65	0.47	1.69
24	Wall (retaining, concrete, brick)	0.56	0.55	1.59
25	Embankment	0.53	0.44	1.63

Severity: maximum, severe, moderate, minor, and no injuries in NASS database, which is coded as 4, 3, 2, 1, 0.

Feature Coincidence

- Features have high coincidence
  - Case Number 2008-03-102 summary:

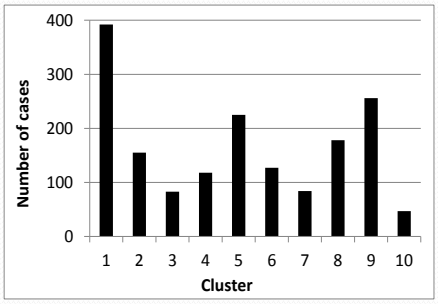
V1 was traveling westbound (C19) on a physically divided two-way roadway with four lanes of travel in the westbound directions. The driver lost control (C11) and the front plane of V1 impacted the center concrete traffic barrier (C2), and then rolled over two quarter turns (C3).

  - Case Number 2008-11-164 summary:

V1 was traveling on an exit ramp (C20) from a fully controlled access westbound (C19) highway (C16) negotiating a curve (C12) to the right. V1 left the roadway on the right and rolled 4 quarter turns (C3) to the left. V1 came to rest off of the road on it's wheels facing south.

Feature Coincidence – Rollover

Cluster 3: Rollover

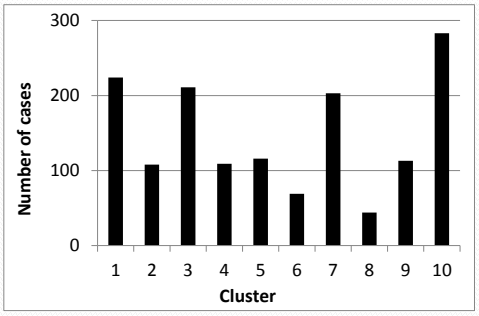


Index	Topics
C1	embankment
C2	tripped, yaw
C4	divided median, guardrail, highway
C5	pole, curb
C6	lost control
C7	tree
C9	driver, steered, avoid, swerved
C10	exit highway ramp, shoulder

Embankment is the largest cluster

Feature coincidence – lost control

Cluster 11: lost control



Index	Topics
C1	Negotiating curve
C2	concrete barrier
C3	roll, turn
C4	guardrail
C6	wall
C7	utility pole
C8	snow, ice
C9	curb
C10	Tree

Tree is still the top one object (C10), negotiating curve (C1) is a highly possible reason.

New Variables – Ramp exit/entrance

- Cluster 20 (Ramp, highway, exit/ entrance):

  206 crashes out of 235 crashes are related with ramp (88% precision). Ramp exit includes 124 cases, while only 39 involves ramp entrance. There is no data about whether the crash occurred at ramp entrance or exit.

  Summary: V1 was traveling on an exit ramp from one highway to another. V1 exited the roadway on the left side and rolled over 6 quarter turns coming to rest on the top of the vehicle off the roadway.

Case Number: 2008-11-125

Environment

Trafficway	
Relation To Junction	Interchange area related
Relation To Flow	One way traffic
Roadway	
Travel Lanes	One
Alignment	Curve right
Profile	Downhill grade (>2%)
Surface Type	Bituminous (asphalt)
Surface Condition	Dry
Conditions	
Light	Dark
Weather	Cloudy
Traffic Control Devices	
Device	No traffic control(s)
Functioning	No traffic control device

New Variables - Highway

Summary: V1 was traveling westbound in lane three on a three-lane highway divided by a metal guardrail. V1 switched lanes from lane three to lane two to avoid a vehicle disabled in her lane. V1 went off the roadway on the left side in a counterclockwise yaw contacting a guardrail with the front and back of the vehicle. V1 at final rest was half in lane three and half on the shoulder facing northwest.

Environment

Trafficway	
Relation To Junction	Non-interchange area and non-junction
Relation To Flow	Divided trafficway-median strip with positive barrier
Roadway	
Travel Lanes	Three
Alignment	Straight
Profile	Level
Surface Type	Bituminous (asphalt)
Surface Condition	Wet
Conditions	
Light	Dark
Weather	Blowing Snow
Traffic Control Devices	
Device	No traffic control(s)
Functioning	No traffic control device

New Variables

- Road is rural (117 summaries) or city (17 summaries).
- Directions:

  Summary of case 2008-11-201: V1 was traveling eastbound on a two-way, two lane roadway. V1 left the road on the left. The front of v1 contacted a 7cm diameter tree. V1 came to rest against the tree facing east.

  For the single-vehicle collision, the initial driving direction and final resting direction may provide some clue about the turns and rotation occurred during the crashes.

Conclusions

- NMF can be used to extract topics from narrative summaries
- Twenty features out of the 25 features extracted are interpretable as topics including the objects impacted in vehicle crashes, road condition, vehicle movement, vehicle directions.
- NMF performance: Recall 53% - 86%
  Precision 44% - 90%
- New variables are observed

References

- D. Zeimpekis, E. Gallopoulos, Text to Matrix Generator User's Guide. http://scgroup20.ceid.upatras.gr:8000/tmg/
- D.D. Lee, H.S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755), 788-791.
- M.B. Berry, M. Browne, A.N. Langville, V.P. Pauca, R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis, 2007, 52, 155-173.
- C. Boutsidis, E. Gallopoulos. SVD based initialization:Ahead start for nonnegative matrix factorization. Pattern Recognition, 2008, 41, 1350-1362.

Acknowledgements

- Jun Liu, Xin Wang in Civil & Environmental Eng.
- Committee members Dr. Berry, Dr. Cao, Dr. Khattak for suggestions and support on this project.