Nonnegative Matrix Factorization on Vehicle Crash Summaries

COSC 501 Pilot Project Report

Ye Sun
Master candidate, Electrical Engineering and Computer Science, University of Tennessee at Knoxville
Email: ysun22@utk.edu

ABSTRACT

In this study, a nonnegative matrix factorization approach is used to extract topics from the summary descriptions of motor vehicle crashes in National Automotive Sampling System. Summaries of single-vehicle crashes from years 2008 to 2012 are collected from National Automotive Sampling System. Vector space model is used to extract meaning from the summary file and generate term-by-document matrix, which is subsequently analyzed using nonnegative matrix factorization (NMF). The resulting feature vectors are analyzed for topic detection. The documents are clustered according to the topics. The quality of the information retrieval using NMF is evaluated. This work attempts to produce a parts-based factorization from the narrative summaries, generate topics directly interpretable, and find out the relations among different features and possible new variables that may be useful for further investigation.

INTRODUCTION

National Automotive Sampling System (NASS) is a national traffic accident database created by National Highway Traffic Safety Administration (NHTSA). Information on motor vehicle crashes was collected to develop, implement, and evaluate motor vehicle and highway safety needs. In the database, each dataset contains ten fields to describe the accident including summary, events, scene diagram, scene photos, general vehicle, vehicle exterior, vehicle interior, safety systems, occupants, images. Variables related to vehicles (vehicle make, model, body category, start/end model year, plan of impact, plan sub-section, rollover), occupant (age, sex, height, weight, seat position), injury (body region, etc.), restraint use (manual/automatic belt, automatic air bag availability/location/deployed, child seat used) are downloadable from the website http://www-nass.nhtsa.dot.gov/nass/cds/SearchForm.aspx. In the summary field of each dataset, there is one-paragraph text from the police report that briefly summarizes the crash with information about roadway, vehicle movement, environmental conditions, damages, etc. In this study, we are using nonnegative matrix factorization (NMF) to retrieve the information from the narratives, find out the topics described, cluster the documents, and discover potential new variables for further study.

METHOD

1. Data Preparation

NASS data were kindly provided by Jun Liu (PhD candidate in Civil & Environmental Engineering at UTK) collected from the NASS website. The dataset collection contains single-vehicle and multiple-vehicle crashes saved in Microsoft Excel format with each dataset (one crash case) in one spreadsheet. In this project, the summary data of single-vehicle crashes from years 2008 to 2012 are used for analysis purpose. The case number and its corresponding summary for each dataset were extracted from the original data and then combined into one spreadsheet using MATLAB. The spreadsheet was then converted by MATLAB to a text file that contains summaries of cases with each case separated by an empty line.

2. Construction of term-document matrix

In this study, we use Text to Matrix Generator (TMG) software developed by D. Zeimpekis and S. Gallopoulos at University of Patras, Greece [1] for the construction of term-by-document matrix and the following NMF analysis. As a MATLAB toolbox, TMG provides a wide range of tools for indexing, non-negative matrix factorizations, dimensionality reduction, retrieval, clustering, and classification. The Indexing module of TMG was used to build the term-by-document matrix in this study. This module uses vector space model to extract meaning from the text file. After removing the commonly used terms (stopwords) and excluding the terms that do not meet a certain threshold on frequency or length, the remaining non-alphanumeric terms in the text file form a dictionary. Based on the dictionary, an m x n term-by-document matrix A = $[w_{ij}]$ is created to represent the text, where w_{ij} is the weight associated with term i in document j, m is the number of terms in the dictionary, and n is the number of documents. In this case, each column of A represents a document vector of length m, which corresponds to the weights of all terms in a summary extracted from a particular case. Each row of A is a term vector of length n, which represents the weights of this term in all of the summaries.

In this study, the log-entropy weighting scheme was used to generate the term weight w_{ij}, which is calculated as $w_{ij} = l_{ij} g_i d_j$. Here l_{ij} denotes the local weight of term i in document j, which is calculated using logarithmic function $l_{ij} = \log{(1 + f_{ij})}$, where f_{ij} is the frequency of term i in document j. In addition, g_i corresponds to the global weight of term i, which is calculated as $g_i = 1 + \left(\frac{(\sum_j (p_{ij} log p_{ij}}{log n}\right)$. Here $p_{ij} = f_{ij} / \sum_j f_{ij}$ is the probability of term i occurring in document j. d_j is a document normalization factor for the columns of A.

The parameters used in TMG Indexing for this study are listed below. The input file is the text file prepared above. A 455-term stoplist with unimportant words was used to be excluded from indexing. The stoplist was customized that includes the stop words provided by TMG package supplemented with most common words for vehicle crashes. The minimum and maximum term length was set to be 2 and 30, respectively. The minimum local or global frequency was set to be 1. The maximum local or global term frequency was set to be infinity. In addition, the columns of matrix A were normalized. The alphanumerics and numbers were removed from the indexing.

3. NMF analysis

NMF, first developed by Lee and Seung [2], is very useful in dimension reduction where data are comprised of non-negative components in a high-dimensional matrix. Given the m × n nonnegative matrix A obtained from parsing the text file, NMF attempts to find two nonnegative factor matrices W and H to minimize the squared Frobenius norm $\|A - WH\|^2 = \sum_{ij}(A_{ij} - (WH)_{ij})^2$. W and H are m × k and k × n matrices, respectively. Here k represents the number of features. The optimal choice of k is application dependent and is typical chosen so that k<< min(m,n). W is a term-by-feature matrix that can be referred to as the feature matrix containing feature vectors describing the themes inherent within the data. H is a feature-by-document matrix referred to as a coefficient matrix since its columns describe how each document spans each feature and to what degree. Since k is much smaller than m and n, WH is a low-rank approximation of A generated by NMF algorithm for feature extraction and cluster identification.

3.1 Initialization of W and H

NMF is an iterative algorithm in which W and H are initialized either randomly or by other optimization methods and then repeatedly updated until it converges to a local minimum. The initialization method affects the final solution and speed of NMF convergence. In the standard NMF algorithm, W and H are initialized with random nonnegative values. To speed up the convergence, various algorithms have been reported [3]. In this study, the initialization was performed using Nonnegative Double Singular Value Decomposition (NNDSVD) [4]. This algorithm contains no randomization and is based on two SVD processes, one approximating the data matrix, the other approximating positive sections of the resulting partial SVD factors utilizing an algebraic property of unit rank matrices, which avoids the negative elements of the SVD by enforcing nonnegativity. The advantage of NNDSVD is: (1) It contains no randomization and therefore provides NMF with a static initialization, which makes NMF converge to the same minima if NMF converges. (2) By providing a structured initialization, the convergence is faster due to the reduced number of iterations to achieve a stationary point.

3.2 Update Rule

Once both W and H have been initialized, those initial estimates are iteratively improved using multiplicative update method [2] given by the following equations:

$$H_{cj} \leftarrow H_{cj} \frac{(W^T A)_{cj}}{(W^T W H)_{cj}}$$

$$W_{ic} \leftarrow W_{ic} \frac{(A H^T)_{ic}}{(W H H^T)_{ic}}$$

As we know that the elements of A, W, and H are nonnegative, the updated matrices are guaranteed to be nonnegative. In addition, Lee and Seung also demonstrated that the Frobenius norm $\|A - WH\|$ is nonincreasing during the iterations and is invariant if and only if W and H are at a stationary point, in which case the closest approximation of A is achieved [2].

In this study, the number of iterations is set up to be 100. MATLAB is used to calculate SVD in the initialization of W and H. The number of feature vectors k was assigned the values of 5, 10 15, 20, 25, and 30. For each k value, the Frobenius norm $\|A - WH\|$ is calculated in each update iteration to show if convergence is achieved.

RESULTS AND DISCUSSION

1. Single-vehicle crashes

In this study, a term-document matrix A is constructed using the Indexing module of TMG package as described in the Method section. Summaries of 4576 single-vehicle crash were extracted from the NASS database and included in a text file using MATLAB. After processed with the indexing module of TMG, a dictionary with 2218 terms was formed. The matrix A is a 2218 x 4576 matrix with its rows representing the terms and columns representing the summaries extracted from a crash case. Each component w_{ij} in matrix A represents the weight associated with term i in summary j calculated by log-entropy scheme. Each column of A is normalized to a length of 1. Using the NNDSVD initialization and multiplicative update algorithms described in Method section 3.1 and 3.2, matrix A was approximated via A ≈ WH, where W is 2218 x k matrix and H is k x 4576 matrix. The goal of NMF is to approximate the original matrix A with factor matrices W and H as accurately as possible. Frobenius norm $\|A - WH\|$ is used to measure the deviation, which is zero if A=WH. Figure 1 shows the Frobenius norm $\|A - WH\|$ during the 100 iterations of multiplicative update at different number of features k = 5, 10, 15, 20, 25, 30. From the figure, we can see that the Frobenius norm decreases rapidly when the update iteration is below 10 and gradually levels off when update process reaches 20 iterations. Therefore, convergence is achieved after 20 iterations of multiplicative update. As k increases from 5 to 30, $\|A - WH\|$ decreases, which means WH is closer to A. This behavior could be explained since increasing the number of features also increases the size of the effective labeling vocabularies, thus enabling a more robust labeling. By analyzing the W matrix, k = 25 is used for the following analysis due to its classification of feature is more explainable with higher purity. Further increase of k to 30 does not provide more valuable information.

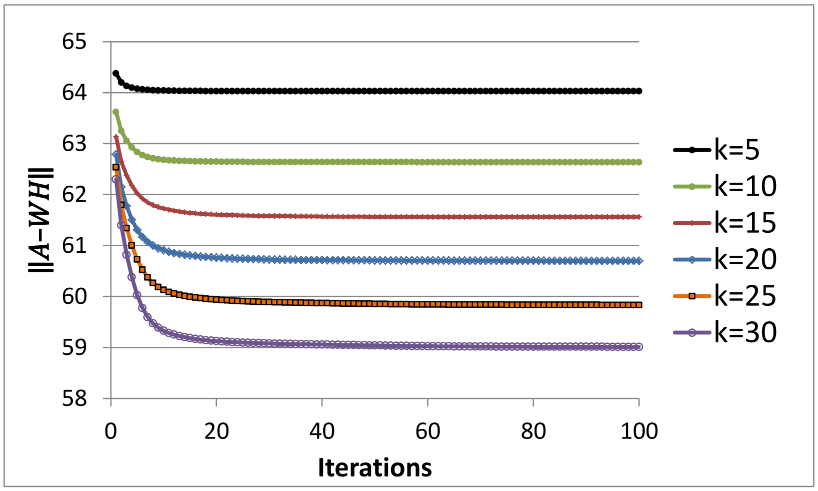

Figure 1. Convergence graph that shows the Frobenius norm $\|A - WH\|$ with iterations in multiplicative update for different number of features.

1.1 Feature Extraction

Twenty-five feature vectors (W_k) corresponding to the k-th column of the matrix W (2218 x 25) were output from NMF module. The 25 features are used to generate 25 topics, which is also the column dimension of the W matrix and row dimension of the H matrix. Table 1 illustrates some of the explainable topics. Ideally each cluster of terms would include the documents by a specific topic. However, there are five features that are too broad to form a topic. The ten dominant terms having the largest magnitude in each selected feature are listed to explain and derive the context of the topics. In table 1, these topics belong to five categories:

(1) The first nine topics are related to the objects that the vehicle contacted or impacted, which include tree, concrete barrier, utility pole, guardrail, median curb, post sign, ditch, wall, and embankment.

(2) The following five features with index 7, 12, 15, 16, 20 related to the road conditions that include intersection, negotiating a curve, turns, divided highway, and ramp exit or entrance.

(3) Feature 5 is related to the appearance of deer, which can be categorized as collision due to animals.

(4) Features 3 and 11 describe the vehicle movement in the collisions. Feature 3 is the rollover with intensity measured by the number of quarter-turns. Feature 11 is related with lost control, which may cause rotation and may happen in icy or wet roadway.

(5) Features 17, 19 and 21 are related to the directions. By looking at the summary collection containing 4576 single-vehicle crashes, the driving direction when the crash occurs is usually described at the first sentence, which is most likely the format required or human instinct on how to describe the crashes. The final rest facing direction is sometimes described in the summary. The eastern driving direction is not extracted as features, but it did appear as dominant terms in features 1, 14, and 15 in Table 1.

Table 1: Selected features and topics from the feature vectors of W (columns of W).

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
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
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

1.2 Analysis of H matrix

A given row of H matrix can be used to reveal the crashes that share common feature vector, i.e. semantic features. A column of H matrix represents the distribution of 25 features in one crash summary. In Figure 2, the number of crashes in each row of H matrix with magnitude greater than $row_{max}/10$ is shown. The utilization of a threshold ($row_{max}/10$) is to remove these elements with less weight and therefore reduce the noises. We can see that the cluster 1 with topic "tree" has maximum number of cases, which indicates that tree is the number one object contacted in single-vehicle crashes. The 2nd largest cluster is in feature 4, which has the top ten high weight terms as "struck, traveled, undivided, tree, side, trafficway, roadway, front, divided, ccw). This cluster is too broad to be meaningful. The clusters 3 (rollover) and 11 (lost control) are ranked as the 3rd and 4th high number of clusters. The cluster 5 represents the topic "deer", which is the smallest cluster. The cluster 18 has two description terms "diameter, cm", which is the description of the size of a tree, pole, or other objects that a vehicle impacted. It is not listed in Table 1 as a topic because it provides supplementary information for the objects, but is not independent.

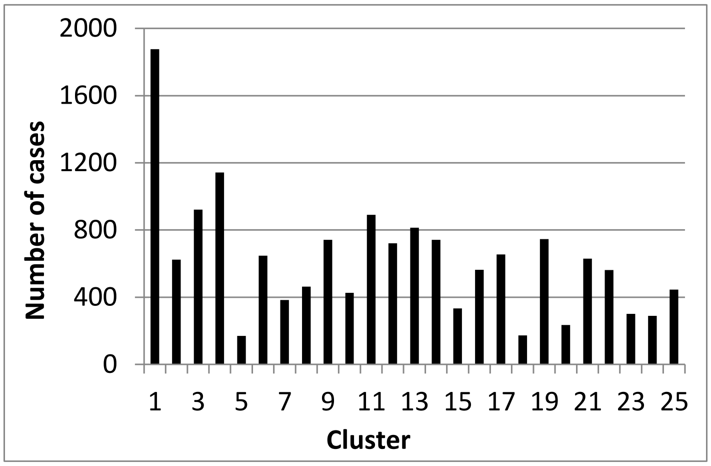

Figure 2. Number of crash cases in each cluster. Log-entropy weighting scheme is used.

1.3 Quality of topics

The summary from each crash is a brief narrative. In the NASS databases, each dataset contains ten fields to describe the accident. The quality of the NMF for topic extraction can be validated using the data available in the dataset. The topic descriptions listed in Table 1 are labels for each cluster of terms. Once labeling is produced for a given topic, a measure of "goodness" must be calculated to determine the quality of clusters. When dealing with simple return lists of documents that can classified as either relevant or not relevant to a specific topic, information retrieval methods typically use precision and recall to describe its performance. Precision is defined as $P = R/L$, where R represents the number of relevant documents within the return list and L is the number of returned documents. Recall, on the other hand, is denoted as $R=R/T$, where T is the number of all relevant documents. Therefore, precision measures how accurately a system returns relevant documents, while recall quantifies the system's coverage of all relevant documents. The goal is to have high levels of precision at high levels of recall. However, as the level of recall rises, precision tends to fall. To perform the validation, let's use the fist topic "Tree" as an example. This topic is related to the crashes that impacted a tree. From the NASS database, the case numbers of all single-vehicle crashes from years 2008 to 2012 that hit a tree are collected, which is referred as collection 1. Using the clusters generated in Figure 2, we get the case numbers for all the crashes that belong to cluster 1 (tree), which is referred as collection 2. The two collections are compared using their case number. R is the number of cases in collection 2 that is also in collection 1. L is the number of cases in collection 2. T is the number of cases in collection 1. In this study, ten topics in Table 1 were selected to validate the NMF results. Among the ten topics, nine topics are about the objects that a vehicle impacted and one topic is about rollover. The reason to select these topics for validation is that these variables are easily available from the NASS database and do not interact with other variables, which provides a clean collection. The recall and precision of each cluster are listed in Table 2. Injuries of crashes are recorded using Abbreviated Injury Scale (AIS): maximum, severe, moderate, minor, and no injuries in NASS database, which is coded as 4, 3, 2, 1, 0. The average crash severity for each selected cluster is calculated using the cases in collection 1.

Table 2. Recall, precision, and severity of each cluster.

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	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

In Table 2, we can see that the selected topics generated from NMF have relatively good recall and precision. The performance of NMF may be affected by how close the approximation of WH to A matrix, the selection of threshold to filter the documents, and the way the narrative was written. From Table 2, we can see that the first topic labeled as tree has relative low recall and precision (below 50%). If the threshold increases, the documents included in the cluster will decrease, which will increase precision and decrease recall, vice versa.

1.4 Coincidence

From the Table 2, we can see the clusters (topics) generated from the feature vectors categorize the documents with relatively good recall and precision. The cluster 3 (rollover) is the 2nd largest interpretable cluster and has relatively high recall (0.78) and precision (0.90). Rollover with intensity expressed by the number of quarter turns is an important factor on crash severity. Further analysis of this cluster is performed. Using the method described in 1.2, the summaries from cases of cluster 2 were isolated into a text files. Then TMG indexing and NMF modules (k = 10) were used to process the data. Ten features were extracted from the W matrix. By observing the ten dominant terms having the largest magnitude in each feature, the labels are produced from the features that are interpretable: cluster 1 (embankment), cluster 2 (tripped and yaw), cluster 4 (divided median, guardrail, highway), cluster 5 (pole, curb), cluster 6 (lost control), cluster 7 (tree), cluster 9 (driver, steered, avoid, swerved), cluster 10 (exit highway ramp, shoulder). Figure 3 shows the number of cases in each feature using a threshold to filter the H matrix. Feature 1 (embankment) contains the largest number of crashes. This can be explained that embankment is usually steep and may cause more rollover. The study of rollover indicates that there exists a coincidence between the cluster 3 (rollover) and cluster 25 (embankment). In addition, the vehicles impacted various objects including median, guardrail, tree, pole, and curb in the collision. The rollover may occur when the drive tried to avoid or swerve something (feature 9). The road may be highway, ramp, shoulder, trafficway as can be seen from the extracted features.

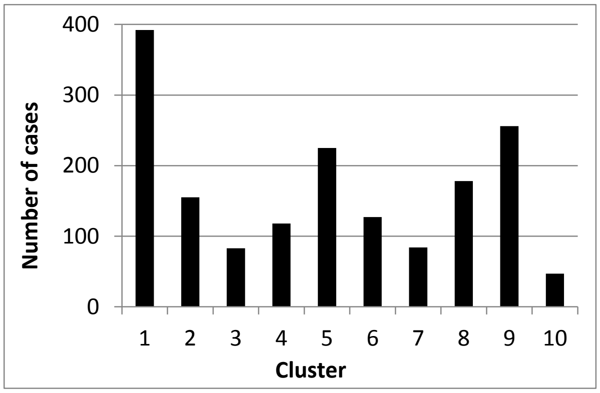

Figure 3. Number of cases in each feature in rollver.

The cluster 11 labeled as "lost control" is the 3rd largest interpretable cluster in Table 1. This cluster is further analyzed following the procedures for the cluster 3 (rollover) as described above. Ten features are extracted from the NMF and clusters can be formed based on the features. The label of each feature based on the top ten dominant terms in W matrix is listed below: Feature 1 (negotiating curve), feature 2 (concrete barrier), feature 3 (roll, turn), feature 4 (guardrail), feature 6 (wall), feature 7 (utility pole), feature 8 (snow, ice), feature 9 (curb), feature 10 (tree). Figure 4 shows the document size of each feature. By analyzing the documents in each feature, we can see that tree is still the top one object impacted in crash that accounts for 35% in 890 crashes in cluster 11. The other objects impacted by a vehicle crash include pole (203 cases), guardrail (109 cases), concrete barrier (108 cases), wall (69), and curb (113 cases). The crashes that occur on negotiating a curve are 224 cases, which indicate a high rate of accident. The coincidence of features is observed here. A vehicle may lost control (feature 11) when negotiating a curve (feature 12 describing road condition) and hit an object (clusters 1, 2, 6, 8, 14, 22, 24). Feature 11 has some coincidence with weather condition, such as ice (44 cases) or rain (30 cases).

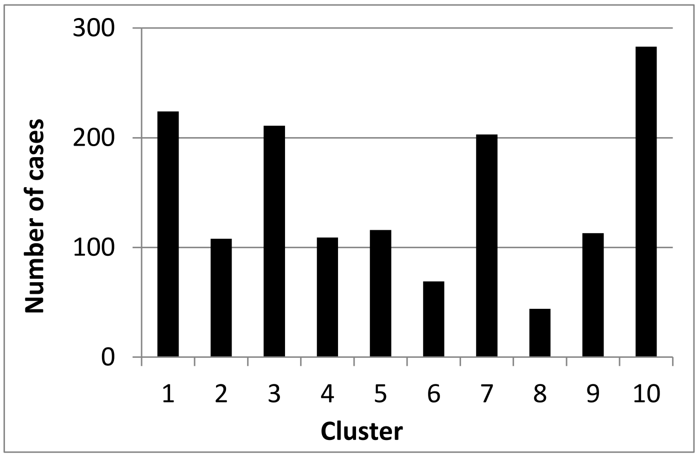

Figure 4. Number of cases in each feature in cluster 11 (lost control).

1.5 New variables for the narrative summaries

One goal of this study is to find out the potential new variables from the summary description that are not available in the NASS database. By analyzing the cluster 20 (Ramp, highway, exit/entrance), there are 206 crashes related with ramp among the 235 crashes in cluster 20, which provides 88% precision. Totally 124 cases involve ramp exit, while only 39 involves ramp entrance. In NASS database, ramp is categorized as Interchange. However, there are no data about whether the crash occurred at ramp entrance or exit. In addition, there are 75 cases in the 235 crashes related to negotiating a curve, which shows a coincidence of cluster 12 (negotiating curve) and cluster 20.

Another new variable that can be retrieved from the summary description is whether the road is highway or not. Among the 4576 case collection from years 2008 to 2012, there are 428 cases that are related to highway crashes according to the summary description. However, this information is not available from NASS database. The cluster 16 in Table 1 is labelled as “highway, divided”, which includes crashes on highway or divided roadway. This cluster contains 564 crash cases with 326 highway collisions, which provide 76% recall. This variable is directly related to the vehicle speed at the time of the crashes, which may play an important role in the crash severity.

Except from the two features discussed above (ramp exit/entrance, highway or not), from my observation of the summary description, there are 117 summaries that describe the road is rural and 17 summaries that describe the road is city road. This accounts for 2.6% of the total collection and therefore is not extracted as a feature by NMF analysis. However, it did give some information that is not available in the NASS database. There is one more observation that the vehicle driving direction when the crash occurs is usually described at the first sentence. It seems that all the summaries follow the format “V1 was traveling eastbound/westbound/southbound/northbound...”. The final rest facing direction is sometimes described in the summary. The driving direction is not collected as a variable in the NASS database. For the single-vehicle collision, the initial driving direction and final resting direction may provide some clues about the turns and rotation occurred during the crashes.

CONCLUSION

In this project, NMF has been used to extract topics from the summaries of vehicle crashes. Twenty out of twenty-five features generated by NMF are interpreted and labeled by topics. These topics provide information about the objects a vehicle impacted, the road conditions, vehicle movement, vehicle directions in the collision. Clusters are generated by features. The performance of NMF is evaluated by precision and recall. Coincidence of features is discussed. We have demonstrated that the parts-based representation of the summaries can retrieve information from vehicle crashes without requiring the reading of individual summaries, which is useful when large amount of text data need to be processed. New variables that are not available in NASS database have been observed in these topics. This research focuses on single-vehicle crashes. Future work is needed in exploring the summaries of multiple-vehicle crashes using NMF.

Reference:

1. D. Zeimpekis, E. Gallopoulos, Text to Matrix Generator User’s Guide. http://scgroup20.ceid.upatras.gr:8000/tmg/
2. D.D. Lee, H.S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755), 788-791.
3. 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.
4. C. Boutsidis, E. Gallopoulos. SVD based initialization:Ahead start for nonnegative matrix factorization. Pattern Recognition, 2008, 41, 1350-1362.