['CDC', 'TEXT', 'CLASSIFICATION']

Yasaman Shirian, Sara Abdali

['Overview']

We analyzed CDC's dataset for 48 events of injury along with a text description of them, and built a multi-class classifier to associate every event with its description.

['Feature', 'Extraction']

- **Word2Vec**: this method uses the **similarity** of the words and attributes a number to each word. If two words are more likely to appear together, their value will be higher.

- **WordNetLemmatizer**: This method uses the **lemma** and **frequency** of each word for classification. i.e. it determines if they are adjective, verb, or noun.

An array of selected size of features that describe the distance of each word in the corpus with a word in the feature vector.

Word2vec

57YOM WITH CONTUSION TO FACE AFTER STRIKING IT WITH A POST POUNDER WHILE SETTING A FENCE POST

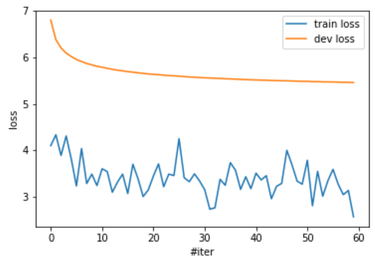
['FENCE', 'STRIKING', 'FACE', 'POUNDER', 'SETTING', 'POST', 'CONTUSION']

WordLemmatizer

An array of selected size of features that describes the importance of a word in the corpus based on frequency of appearance.

We have probed the dev set accuracy with 500,1000 features size vector

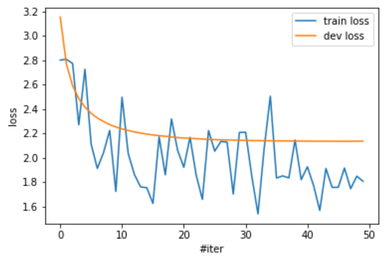
['Support', 'vector', 'machine']



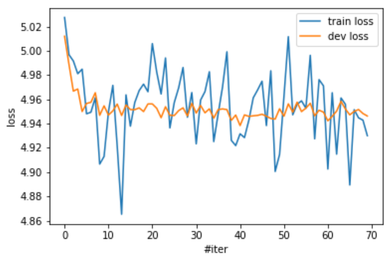
L1 regularization

word2vec

L2 regularization



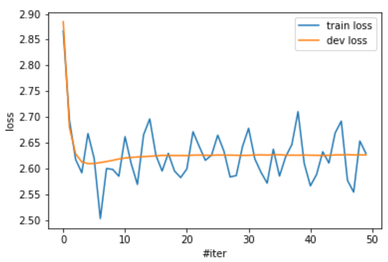
['Linear', 'Regression']



L1 regularization

WordLemmatizer

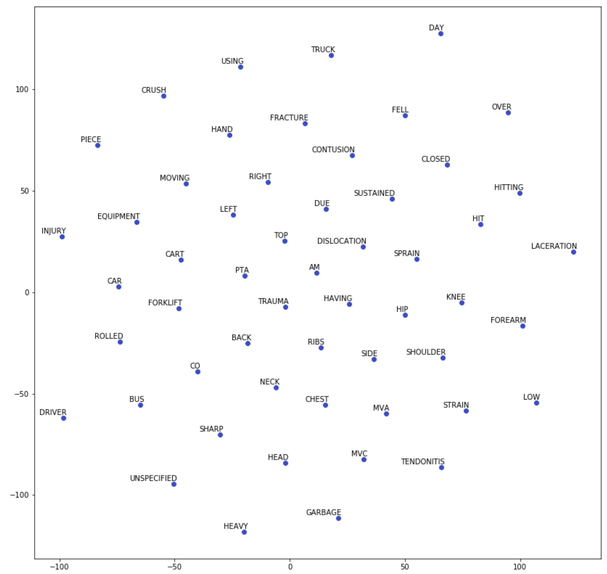
L2 regularization



error	L1	L2
LR	train : 82% dev:82%	train : 44% dev:48%
SVM	train : 36% dev:37%	train : 29% dev:29%
NB	in progress	in progress

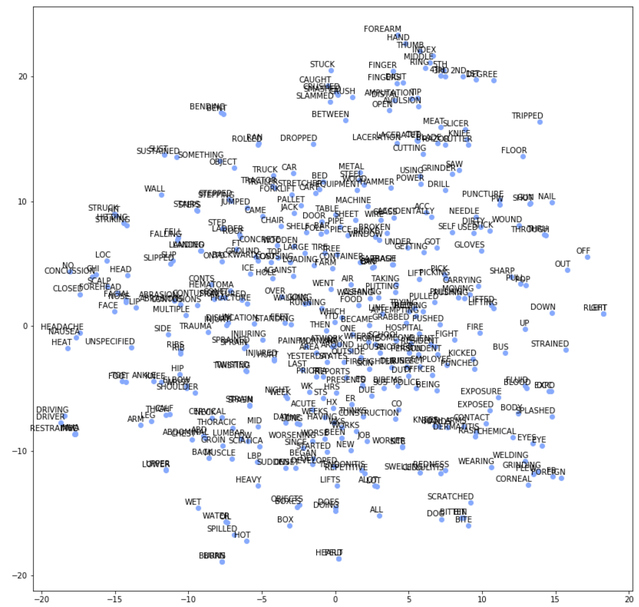
['Error', 'Analysis']

Why some events are easier to classify?



95% error for this class prediction

8% error for this class prediction



['References']

- *Tom Kenter, Alexey Borisov, Maarten de Rijke 2016. Siamese CBOW: Optimizing Word Embeddings for Sentence Representations.*
- *Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.*