CSE 523: Machine Learning

Group 17 - Hardly Humans
Weekly Project Report – 8

Quora Insincere Questions Classification

Name	Enrolment Number
Malav Doshi	AU1940017
Parth Shah	AU1940065
Sanya Zaveri	AU1920064
Mihir Pathak	AU1920138

1) Tasks Performed in the week.

- Implemented TF-IDF
- Implemented word2vec

2) Outcomes of the tasks performed.

- The new features added by T F-IDF
    1) skew_question_text_vec
    2) kur_question_text _vec

	qid	question_text	target	num_words	num_unique_words	len_text	len_char_question_text	len_word_question_text	num_stopwords	common_words	len_mean_words	skew_question_text_vec	kur_question_text_vec
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0.0	13	13	72	24	13	7	13	4.615385	8.166792	89.894743
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0.0	16	15	81	20	16	10	15	4.125000	3.332967	28.185428
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0.0	10	8	67	20	10	3	7	5.800000	-0.510557	22.371391
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0.0	9	9	57	24	9	3	9	5.444444	4.064800	35.350999
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0.0	15	15	77	23	15	7	15	4.200000	3.494224	39.028359

3) Cosine distance

4) Jaccard distance

5) Euclidean distance

cosine_distance	jaccard_distance	euclidean_distance
0.958788	1.0	1.384766
1.044776	1.0	1.445528
0.975169	1.0	1.396545
0.974808	1.0	1.396287
0.998848	1.0	1.413398

3) Tasks to be performed in the upcoming week.

- Implementing different balancing methods.
- Implementing different models like SVM, Random Forest and LDA.