Abstract:

This report describes the text classification problem about questions on Reddit.com.

We concerned on extracting the topic words which decided by our own. We first achieved data from the homepage in Reddit.com. After, we construct the keyword bag for different labels. Then, we used the provided glove twitter 50d dataset as pre-trained data. After the word passes through the embedding layer, the local information, global information or context information is extracted by using the structure such as RNN.

Keyword: Reddit, text classification, embedding, word2vec, RNN

Introduction:

* Preparation:

Read the relevant library functions, read in the data set, and divide the training set and test set.

* Data serialization:

converting text to a sequence of numbers

* Get the word embedding matrix:

read the love model and get the word embedding matrix

* Training network:

divide the training set and verification set, build a simple RNN network, train the network