Meaning of the fields in each line of training and test data.

Let's use the top five line in train.txt as an example:

1 qid:1 26:1 27:1 28:1 29:1 30:1 ...

11 qid:1 9:1 17:1 25:1 28:1 29:1 ...

5 qid:1 36:1 37:1 42:1 43:1 44:1 ...

15 qid:2 20:1 21:1 22:1 26:1 27:1 ...

13 qid:2 43:1 44:1 46:1 47:1 49:1 ...

Each line stands for a letter. In the first line:

1 is the class label of the letter. Numbers start from 1, i.e. 1 means 'a', 11 means 'k'.

qid:1 means the WORD id is 1. So the lines 1-3 correspond to the first word,

and then the second word starts from line 4.

26:1 means the 26-th feature is 1. The rest of the line is the feature description of the letter.

If the feature id does not appear (e.g. there is no 10:1),

that means the corresponding feature (e.g. 10) is 0.

Since a lot of feature are valued 0, this is a very compact way of representing the features.

When the program reads the data, it will add a constant feature 1 to the feature vector.

This leads to 16\*8 + 1 = 129 features, which applies to both training and test data.

Some statistics:

train.txt:

3438 words

25953 letters

test.txt:

3439 words

26198 letters