Question 1 (3 points)

Identify 10 non-word tokens in the passage.

</

3

<font

&quot

<img

url

alt

src

Question 2 (2 points)

Suppose this passage constitutes a document to be classified, but you are not certain of the business goal of the classification task. Identify material (at least 20% of the terms) that, in your judgment, could be discarded fairly safely without knowing that goal.

and

will

be

for

to

on

their

Do

that

If

the

your

Question 3 (3 points)

Suppose the classification task is to predict whether this post requires the attention of the instructor, or whether a teaching assistant might sufice. Identify the 20% of the terms that you think might be most helpful in that task.

John

Illustrations

demos

provided

students

work

own

finish

project

where

find

demos

help

Question 4 (3 points)

What aspect of the passage is most problematic from the standpoint of simply using a bag-of-words approach, as opposed to an approach in which meaning is extracted?

If we simply use a bag-of-words, one document is not enough. Also, bag-of-words approach loses the ordering of the words, questions are treated the same as statement. Semantics loses too.