Assignment 1: Model Answers

Q.1 Answer:

The first thing to do is to create a TextMiner object as:

```
>  tm_train = tm.TextMiner('tmsk.properties')
```

This will load the properties file and parse all the required options. This ends up the pre-processing stages. We are now ready to answer the questions.

Q1.a)

We start by tokenizing the document:

```
> tm_train.tokenize()
```

Finally we generate a dictionary of 500 words

```
> tm_train.mkdict(500)
```

Q1.b)

From the TextMiner object of the previous question we add stopwords removal and word stemming:

```
> tm_train.stopwords()
> tm_train.stem()
```

And now we repeat the same steps to get the new dictionary:

```
> tm_train.mkdict(500)
```

Q1.c)

We can change the properties file to add white spaces etc and run all the steps again:

```
> tm_train = tm.TextMiner('tmsk.properties')
> tm_train.tokenize()
> tm_train.stopwords()
> tm_train.stem()
> tm_train.mkdict('earn',50)
```

Q.2 Answer:

We use the TextMiner object created in previous question to create a vector after generating the dictionary.

```
> tm_train.vectorize()
```

Finally, we pickle the TextMiner object in a file for later use in future assignments.

```
> pickle.dump(tm_train,open( "train.p", "wb" ))
```