To complete this task I used MeTA toolkit.

I used **ngram-word** analyzer method to preprocess data. This method was used with the next filter chain:

- whitespace tokenizer
- converting to lowercase
- selecting words with length from 2 to 35 characters
- stop word removal (lemur-stopwords.txt was used)
- stemming

After this I build forward and inverted indexes to represent preprocessed data as features and be able to use it with learning classifier algorithms.

I used multiple classifier algorithms, with mostly default configuration given by MeTA, with some minor tweaks to the parameters, if the method was supporting it. Most of them performed with almost the same results:

1. Support vector machine

```
[classifier]
method = "one-vs-all"
    [classifier.base]
    method = "sgd"
    loss = "hinge"
    prefix = "sgd-model"
```

As SVM is a non-parametric method it means you cannot really tweak it and should rely on what it can get you by default. Theoretically this method should be the best one in handling and not being overwhelmed by the huge amounts of noise in text data, but in reality it gave almost the same results as the naive bayes algorithm, which is probably due to the base classifier i used.

2. Naive Bayes

```
[classifier]
method = "naive-bayes"
```

This method gave the best results, and finished in pretty fast amount of time. As for the good results I assume it's due to the fact that this method is better in separating signals from noise than two other methods that i used.

3. K-nearest neighbor

```
[classifier]
method = "knn"
k = 10
    [classifier.ranker]
    method = "bm25"
```

This method was the slowest out of 3 methods i used, and gave the worst results, which may be related to the fact that this method do not handle irrelevant features in the data. Which probably was the main issue here, because even after stop-word removing initial data had ton of irrelevant words to the hygiene topic.

Results

Part	Name	Last Submission	Score	Feedback	
1 / 1	Task 6	Sun 11 Oct 2015 10:29 AM PDT	15.00 / 15	View	Submit
Total Score		**15 / 15**			