Task 6: Hygiene Prediction

# Description and Comparison

Based on the same StringToWordVector configurations demonstrated in the next section. I tested the different classification algorithms in Weka. Here below is the result:

# Top Ranked Method (F1: 0.551731949446, 15/15)

## Tools

Python: pre-process and format data

Weka 3.7: classification and prediction

## Algorithms

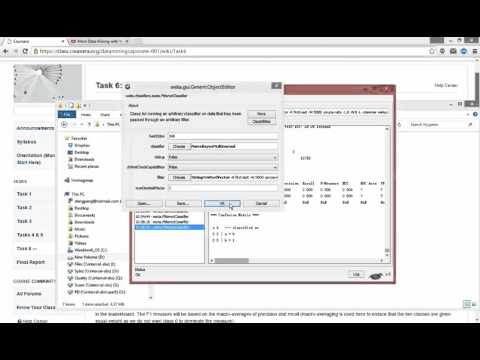
* Classification:
  + Naïve Bayes Multinominal
* Filter:
  + String to Word Vector
    - Lower Case Tokens: True
    - Stemmer: SnowballStemmer
    - StopwordsHandler: Rainbow
    - Tokenizer: Weka WordTokenizer
    - WordsToKeep: 5000

## Tutorial

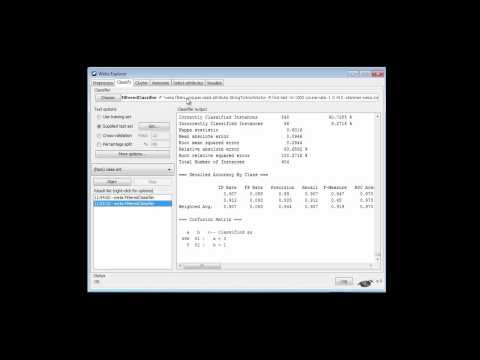
## Steps

1. Download raw data
2. Pre-process the data to Weka ARFF format using a simple python script.   
   <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%206/data2arff.py>   
   The labelled 546 instances are split into a training arff file and the rest 12753 are in test arff file labled with ? mark.
3. Open Weka 🡪 Explorer 🡪 Classifier 🡪 Choose “Meta: Filtered Classifier” 🡪 Classifier:Bayes: Naïve Bayes Multinominal 🡪 Filter:Unsupervised:Attribute:StringToWorldVector 🡪 Supplied test set to load test arff file 🡪 “More options…” 🡪 Output predictions to CSV file
4. Extract predicted label from output predictions from Weka by a python script  
   <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%206/csv2submission.py>
5. Submit

## Video

[](https://www.youtube.com/watch?v=wg5KOhvTqmE)

# Reference

[](https://www.youtube.com/watch?v=5_B_LLFidFE)