Tutorial 10

Data Cleaning

```
{'your', 'were', 'themselves', 'have', 'the', 'o', 'he', 'myself', "it's", 'they', "she's", 've', 'mightn', 'are', "you'd", 'during', 'then', 'so', 'doesn', 'himself', 'hav
Jim stole my tomato sandwich.
['jim', 'stole', 'my', 'tomato', 'sandwich']
"Help!," I sobbed, sandwichlessly.
['help', 'i', 'sobbed', 'sandwichlessly']   Is i a stopword?  True
"Drop the sandwiches!," said the sandwich police.
['drop', 'the', 'sandwiches', 'said', 'the', 'sandwich', 'police']   Is drop a stopword?  False
['drop', 'help', 'i', 'jim', 'my', 'police', 'said', 'sandwich', 'sandwich', 'sandwiches', 'sandwichlessly', 'sobbed', 'stole', 'the', 'the', 'tomato']
```

Vectorizer | TFIDF

```
['jim', 'stole', 'my', 'tomato', 'sandwich']            [0 0 1 0 0 1 0 0 1 1]
['help', 'i', 'sobbed', 'sandwichlessly']                  [0 1 0 0 0 0 1 1 0 0]
['drop', 'the', 'sandwiches', 'said', 'the', 'sandwich', 'police'] [1 0 0 1 1 2 0 0 0 0]
vocabulary word count vector ------------------->  [1 1 1 1 1 3 1 1 1 1]
[[1. 0.]
 [0. 1.]]  <------------cosine similary array vocabulary_vector[0] vs vocabulary_vector[1]
[[1.         0.37796447]
 [0.37796447 1.        ]]  <------------cosine similary array vocabulary_vector[0] vs vocabulary_vector[2]
[[1. 0.]
 [0. 1.]]  <------------cosine similary array vocabulary_vector[1] vs vocabulary_vector[2]
['jim', 'stole', 'my', 'tomato', 'sandwich']            [0.   0.   0.25 0.   0.   0.25 0.   0.   0.25 0.25]
['help', 'i', 'sobbed', 'sandwichlessly']                  [0.   0.33 0.   0.   0.   0.   0.33 0.33 0.   0.  ]
['drop', 'the', 'sandwiches', 'said', 'the', 'sandwich', 'police'] [0.2 0.  0.  0.2 0.2 0.4 0.  0.  0.  0. ]
vocabulary TF vector --------------------------->  [0.2  0.33 0.25 0.2  0.2  0.65 0.33 0.33 0.25 0.25]
```

NLP Example with Spam

```
 [0. 1.]]  <------------cosine similarity vocabulary_TFvector[0] vs vocabulary_TFvector[1]
[[1.         0.37796447]
 [0.37796447 1.        ]]  <------------cosine similarity vocabulary_TFvector[0] vs vocabulary_TFvector[2]
[[1. 0.]
 [0. 1.]]  <------------cosine similarity vocabulary_TFvector[1] vs vocabulary_TFvector[2]
['drop', 'help', 'i', 'jim', 'my', 'police', 'said', 'sandwich', 'sandwich', 'sandwiches', 'sandwichlessly', 'sobbed', 'stole', 'the', 'the', 'tomato']
['jim', 'stole', 'my', 'tomato', 'sandwich']            [0.   0.   1.09 0.   0.   0.4  0.   0.   1.09 1.09]
['help', 'i', 'sobbed', 'sandwichlessly']                  [0.   1.09 0.   0.   0.   0.   1.09 1.09 0.   0.  ]
['drop', 'the', 'sandwiches', 'said', 'the', 'sandwich', 'police'] [1.09 0.   0.   1.09 1.09 0.4  0.   0.   0.   0.  ]
vocabulary IDF vector ------------------------->  [1.09 1.09 1.09 1.09 1.09 0.8  1.09 1.09 1.09 1.09]
[[1. 0.]
 [0. 1.]]  <--------cosine similarity IDFV[0] vs IDFV[1]
[[1.         0.04296109]
 [0.04296109 1.        ]]  <--------cosine similarity IDFV[0] vs IDFV[2]
[[1. 0.]
 [0. 1.]]  <--------cosine similarity IDFV[1] vs IDFV[2]
```

```
[[1. 0.]
 [0. 1.]]  <--------cosine similarity IDFV[1] vs IDFV[2]
['drop', 'help', 'i', 'jim', 'my', 'police', 'said', 'sandwich', 'sandwich', 'sandwiches', 'sandwichlessly', 'sobbed', 'stole', 'the', 'the', 'tomato']
['jim', 'stole', 'my', 'tomato', 'sandwich']            [0 0 1 0 0 1 0 0 1 1]
['help', 'i', 'sobbed', 'sandwichlessly']                  [0 1 0 0 0 0 1 1 0 0]
['drop', 'the', 'sandwiches', 'said', 'the', 'sandwich', 'police'] [1 0 0 1 1 2 0 0 0 0]
```

```
vocabulary word count vector ------------------>  [1 1 1 1 1 3 1 1 1 1]
vocabulary TF vector -------------------------->  [0.2  0.33 0.25 0.2  0.2  0.65 0.33 0.33 0.25 0.25]
vocabulary DF vector -------------------------->  [0.33 0.33 0.33 0.33 0.33 0.66 0.33 0.33 0.33 0.33]
vocabulary IDF vector ------------------------->  [1.09 1.09 1.09 1.09 1.09 0.8  1.09 1.09 1.09 1.09]
vocabulary TFIDF vector ----------------------->  [0.21 0.35 0.27 0.21 0.21 0.52 0.35 0.35 0.27 0.27]
[[1.        0.9836604]
 [0.9836604 1.       ]]  <--------cosine similarity ~ vocabulary count vs vocabulary TF
[[1.         0.98058068]
 [0.98058068 1.        ]]  <--------cosine similarity ~ vocabulary count vs vocabulary DF
[[1.         0.85488733]
 [0.85488733 1.        ]]  <--------cosine similarity ~ vocabulary count vs vocabulary IDF
[[1.         0.95993667]
 [0.95993667 1.        ]]  <--------cosine similarity ~ vocabulary count vs vocabulary TFIDF

Process finished with exit code 0
```