

### Document classes :

1. Anime
2. Arduino
3. Astronomy
4. Biology
5. Chess
6. Coffee

**Reason of choice:** Laptop crushed while working for larger classes numbers than 6.

## K Nearest Neighbour:

	K=1	K=3	K=5
<b>Hamming</b>	60.77396657871592	58.83905013192612	57.34388742304309
<b>Euclidean</b>	70.2726473175022	70.09674582233949	69.74494283201408
<b>TF-IDF</b>	89.62181178540017	90.67722075637643	91.38082673702726

Table 1: the validation accuracy under different types and neighbour count

As the accuracy for tf-idf and k=5 is the highest. We will use it as the final configuration and report our test accuracy.

93.33333333333333,	78.33333333333333,	88.33333333333333,	91.66666666666667,
88.33333333333333,	85.0,	88.33333333333333,	80.0,
98.33333333333333,	91.66666666666667,	90.0,	93.33333333333333,
93.33333333333333,	96.66666666666667,	91.66666666666667,	90.0,
96.66666666666667,	98.33333333333333,	93.33333333333333,	90.0,
98.33333333333333,	90.0,	83.33333333333333,	78.33333333333333,
88.33333333333333,	71.66666666666667,	58.33333333333333,	60.0,
73.33333333333333,	71.66666666666667,	78.33333333333333,	96.66666666666667,
95.0,	91.66666666666667,	95.0,	93.33333333333333,
96.66666666666667,	98.33333333333333,	90.0,	100.0,
95.0,	98.33333333333333,	95.0,	100.0

Table 2: the test accuracy of 50 non-overlapping segments that consist of the whole testset.

## Naive bayes:

Smoothing Factor, $\alpha$	Validation accuracy
0.1	96.48197009674583
0.2	96.21811785400176
0.3	96.39401934916447
0.4	96.3060686015831
0.5	96.3060686015831
0.6	96.3060686015831
0.7	96.3060686015831
0.8	96.1301671064204
0.9	96.04221635883906
1	95.9542656112577

Table 3: the validation accuracy under different smoothing factors for Naive Bayes.

As the accuracy for smoothing factor,  $\alpha = 0.1$  is the highest. We will use it as the final configuration and report our test accuracy.

95.0, 89.65517241379311, 98.27586206896552, 96.55172413793103, 94.54545454545455, 96.49122807017544, 100.0, 100.0, 95.0, 100.0, 98.27586206896552, 98.14814814814815, 94.91525423728814, 96.49122807017544, 96.66666666666667, 96.61016949152543, 93.22033898305085, 91.66666666666667, 98.33333333333333, 98.33333333333333, 91.66666666666667, 96.66666666666667, 94.82758620689656, 96.66666666666667, 91.2280701754386, 89.65517241379311, 93.10344827586206, 98.24561403508773, 96.66666666666667, 96.61016949152543, 91.52542372881356, 91.66666666666667, 94.91525423728814, 94.91525423728814, 93.33333333333333, 91.66666666666667, 98.27586206896552, 94.91525423728814, 98.30508474576271, 94.91525423728814, 96.66666666666667, 98.30508474576271, 93.33333333333333, 94.91525423728814, 96.66666666666667, 95.0, 96.66666666666667, 93.33333333333333, 90.0, 93.33333333333333
---

Table 4: the test accuracy of 50 non-overlapping segments that consist the whole testset.

## T-statistics:

We calculate t-test using the values illustrated in table 2 and 4.

Combination	Statistics	Pvalue
(Naive Bayes, KNN)	3.8906065136808343	0.00030169724059368736

Table 5: Statistical Hypothesis Testing Result

Significance Level	Decision	Which one is better?
0.005	Reject null hypothesis, Statistically significant	Naive Bayes
0.01	Reject null hypothesis, Statistically significant	Naive Bayes
0.05	Reject null hypothesis, Statistically significant	Naive Bayes