k Nearest Neighbor (kNN)

Topics

- Windows_Phone
- Astronomy
- Coffee
- Chess
- Cooking
- Wood_Working
- Law
- Space
- Arduino
- Biology
- Anime

Training Data Size

- 5500 documents

Validation Data Size

- 2200 documents

Methodologies and k Matrix

	1	3	5
Hamming Distance	40.91%	41.59%	41.27%
Euclidean Distance	57.27%	57.64%	57.32%
Cosine Similarity	81.23%	83.50%	**83.68%**

Best Performing kNN Parameters

> $k = 5$
> Methodology : **Cosine Similarity**

Naive Bayes

Topics

- Anime
- Astronomy
- Arduino
- Windows_Phone
- Biology
- Chess
- Wood_Working
- Coffee
- Space
- Law
- Cooking

Training Data Size

- 5500 documents

Validation Data Size

- 2200 documents

Accuracy for Different Smoothing Factors (α)

Serial	Smoothing Factor (α)	Accuracy
1	0.10	71.59%
2	**0.20**	**71.64%**
3	0.30	71.55%
4	0.40	71.23%
5	0.50	71.23%
6	0.60	71.18%
7	0.70	71.09%

Serial	Smoothing Factor (α)	Accuracy
8	0.80	70.82%
9	0.90	70.59%
10	1.00	70.55%

Best Performing NB Parameters

> Smoothing Factor, $\alpha = 0.2$

kNN vs. NB Accuracy

Serial	kNN ($k = 5$,Cosine Similarity)	NB ($\alpha = 0.2$)
1	86.36%	70.91%
2	88.18%	79.09%
3	87.27%	77.27%
4	82.73%	70.91%
5	86.36%	64.55%
6	88.18%	72.73%
7	87.27%	73.64%
8	87.27%	76.36%
9	81.82%	72.73%
10	82.73%	70.00%
11	81.82%	71.82%
12	85.45%	75.45%
13	77.27%	73.64%
14	80.00%	67.27%
15	83.64%	70.00%
16	76.36%	71.82%
17	80.91%	75.45%

Serial	kNN ($k=5$,Cosine Similarity)	NB ($\alpha=0.2$)
18	78.18%	71.82%
19	75.45%	70.91%
20	78.18%	72.73%
21	80.00%	70.00%
22	88.18%	75.45%
23	86.36%	78.18%
24	79.09%	76.36%
25	83.64%	77.27%
26	83.64%	78.18%
27	82.73%	74.55%
28	78.18%	77.27%
29	80.91%	71.82%
30	80.00%	78.18%
31	84.55%	78.18%
32	81.82%	77.27%
33	85.45%	67.27%
34	82.73%	70.00%
35	82.73%	70.00%
36	80.00%	72.73%
37	80.91%	69.09%
38	78.18%	73.64%
39	88.18%	77.27%
40	89.09%	82.73%
41	78.18%	69.09%
42	88.18%	79.09%
43	80.00%	71.82%

Serial	kNN ($k = 5$,Cosine Similarity)	NB ($\alpha = 0.2$)
44	80.91%	69.09%
45	84.55%	74.55%
46	83.64%	80.00%
47	82.73%	70.91%
48	80.91%	69.09%
49	80.91%	74.55%
50	84.55%	74.55%

T-test

Significance Level	T Statistics	T Critical Value	Result
0.005	15.439037	2.679952	*t_critical* < *t_stat* **kNN** better
0.010	15.439037	2.404892	*t_critical* < *t_stat* **kNN** better
0.050	15.439037	1.676551	*t_critical* < *t_stat* **kNN** better

Justifiation

- **Cosine Similarity with TF-IDF** not only considers the term frequency but also takes an account of the **differentiating power** of a particular word. If a word appears in all the documents, it is considered less important.
  NB does not consider differentiating power of words.
- **kNN** implementation **discards any new word** in test document.
  NB implementation does not discard the new words, rather smooths the probability of any new word to a small probability value. Consequently, if a document has a lot of new words, those words may divert the probability from the correct class.