NLP Assignment 2

1. Does naive Bayes outperform the baseline? Does naive Bayes outperform logistic regression? For this question consider only naive Bayes without using start- and end-of-word information.

   Yes, Naïve Bayes outperforms the baseline.
   No, Naïve Bayes does not outperform Logistic Regression.

   Both Naïve Bayes and Logistic Regression significantly outperform the baseline in terms of F1-score, indicating that they are better at identifying positive cases. However, between Naïve Bayes and Logistic Regression, Logistic Regression has a slightly higher F1-score, Precision, and Recall.

	A	P	R	F
Baseline	0.519	0.029	0.056	0.038
Naïve Bayes	0.760	0.522	0.278	0.321
Logistic Regression	0.784	0.554	0.299	0.347

2. Does incorporating start-and-end of word information in naive Bayes give improvements over (vanilla) naive Bayes?

   Yes, Naïve Bayes start end works better than Naïve Bayes in terms of accuracy by approx. 5%, recall by 9%, and F1-score by 8%.
   Specifically, NBSE has higher Recall and F1-score, which indicates that it performs better at identifying positive cases (higher Recall) while maintaining a balance between Precision and Recall (higher F1-score).

	A	P	R	F
Naïve Bayes	0.760	0.522	0.278	0.321
Naïve Bayes SE	0.811	0.509	0.365	0.401

3. Is performance better on names originating from some languages than others? If so, why might this be the case?

   Yes, the performance on names originating from some languages is better than others.

'Russian': 6558	'Italian': 491	'Irish': 144
'English': 2526	'Arabic': 66	'Vietnamese': 47
'Spanish': 216	'Korean': 60	'German': 519
'Japanese': 701	'Chinese': 186	'Dutch': 207
'Czech': 352	'French': 193	'Portuguese': 52
'Greek': 130	'Polish': 93	'Scottish': 69

   There are several factors contributing to this variance in the performance:
   a. Data Imbalance – We have significantly more data in Russian and English than others, which leads to better performance that these two languages.
   b. Similarity between languages – Spanish and Portuguese may have similarities due to the shared Latin alphabet.

4. We considered a bag-of-character trigrams representation in this assignment. The choice of trigrams was somewhat arbitrary. Can you achieve better performance with a different choice, e.g., bigrams (2-grams) or 4-grams?

Naïve Bayes	A	P	R	F
2-grams	0.749	0.522	0.294	0.338
3-grams	0.760	0.522	0.278	0.321
4-grams	0.740	0.526	0.225	0.265

   From above table, we can conclude that using 2-grams results in the highest F1-score among the three representations (2-gram, 3-gram, and 4-gram). It could be because it captures some character-level patterns.
   The 3-gram representation is similar to 2-grams but with slightly lower Recall and F1-score, whereas 4-grams results in the lowest Recall and F1-score among the three representations indicating that adding one more character to each n-gram may not necessarily improve performance significantly.