Selected 3 – Language Detection Project

Team Members:

Level	ID	Name
4	20170412	محمد احمد محمد على
4	20170396	ماريو عماد كتشنر خير
4	20170399	مايكل اسحاق جورجي اسحاق
4	20170414	محمد أسامة نبوي عاشور
4	20170446	محمد صلاح عبدالقادر ابراهيم
4	20170458	محمد عزت محمود عيسى عيد صالح الجمل

Problem description:

Language Identification is one of the Natural Language Processing problems and it is to predict the natural language that is written into a document or a text, Language identification has been a very important and research-intensive idea for over fifty years.
We have obtained the WiLi-2018 Dataset which contains around 250+ languages, we have only worked on 25 and 100 language because of the weak resources that we have, and we obtained accuracies along with confusion matrices, all of this is explained in detail below.

Model design:

We have used the SKLearn library in python to build the model up, we did language modelling using N-Grams, we have used the CountVectorizer to slice the strings up into N-Grams Vector and then use the pipeline architecture to feed the data into the models.

Experimental results:

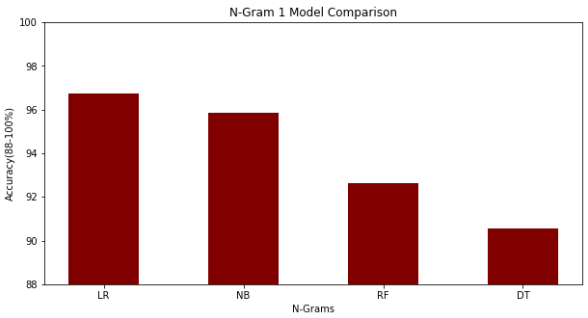
1-25-Language:

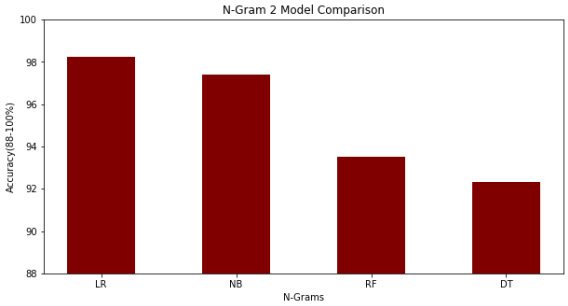
We have used 4 models, each model has proceeded 6 phases, one with the raw string and 5 with the n-grams, we have used the n-grams in the form of 1,2,3,4,5 grams and then classified them accordingly, the models we used were Decision Tree, Random Forest (Ensemble Learning), Logistic Regression (Linear Model), Naïve Bayes (Multinomial), the best model in the 25-language trial was the Logistic Regression with the accuracy of 98.464% in the n-gram = 3.

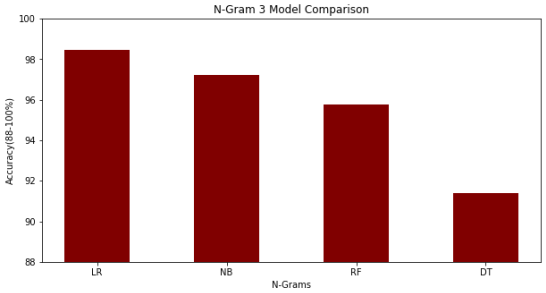
Accuracy Scores:

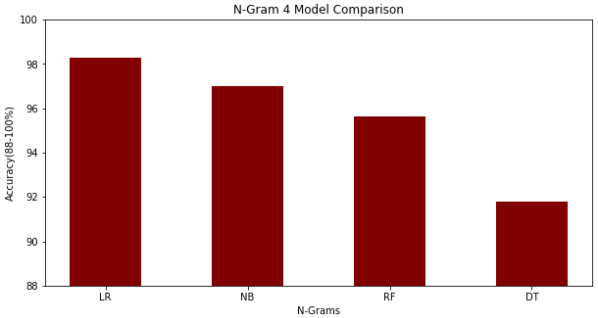
	Strings	N-1	N-2	N-3	N-4	N-5
Naïve Bayes	93.816%	95.856%	97.408%	97.216%	97.016%	96.328%
Logistic Regression	94.56%	96.744%	98.216%	**98.464%**	98.264%	97.92%
Random Forest	94.56%	92.608%	93.512%	95.76%	95.608%	94.024%
Decision Tree	90.016%	90.568%	92.296%	91.384%	92.08%	89.928%

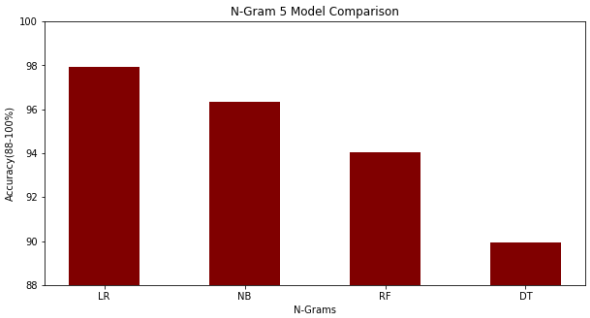
These Diagrams show the difference in accuracy between the models in different N-Grams:



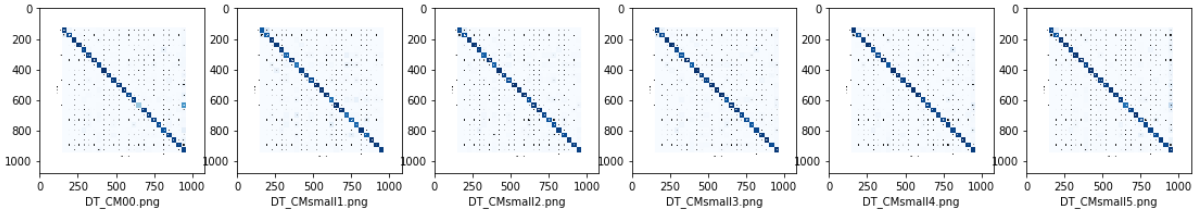


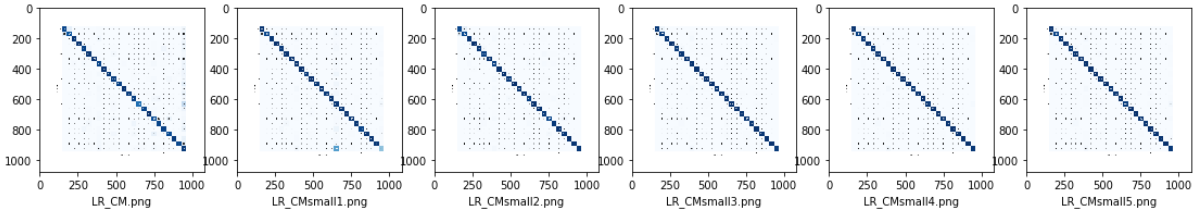


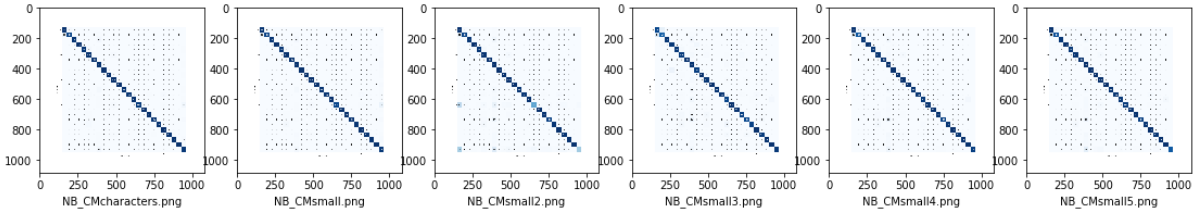


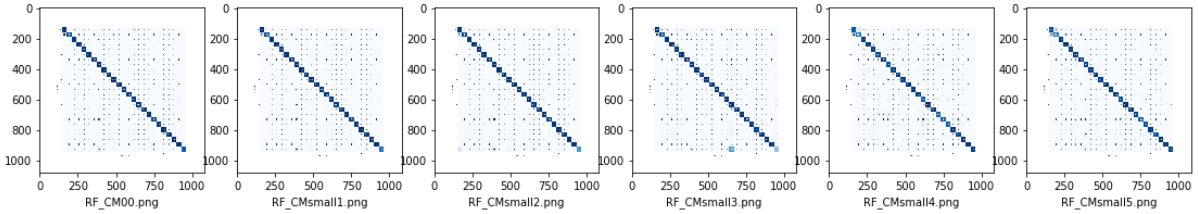


This is the collection of the confusion matrices of all the tests that we have done labelled with all of the models and N-grams used, if you want to check the actual confusion matrices you can check the github repo.









2-100-Language:

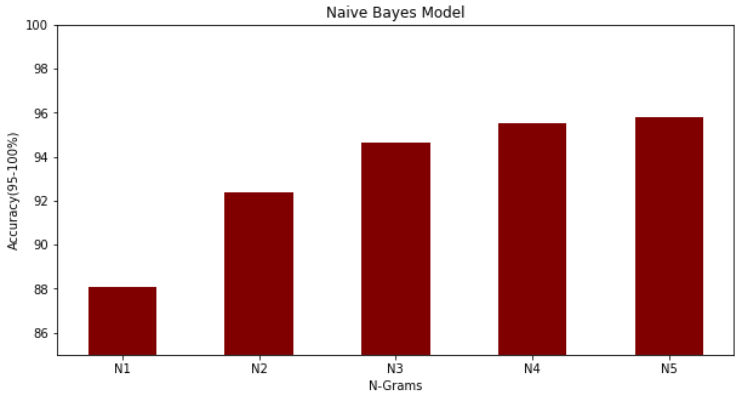
For the 100-language trial we have not been able to produce results for several models other than the Naïve Bayes because of its speed and the logistic regression for its ram optimization, and the naïve bayes was the highest as it reached the accuracy of 95.8% in the n-gram = 5.

Our best model in the 100-language scope was the Naïve Bayes at the N-Gram of 5 on the 100 Languages that are as follows.

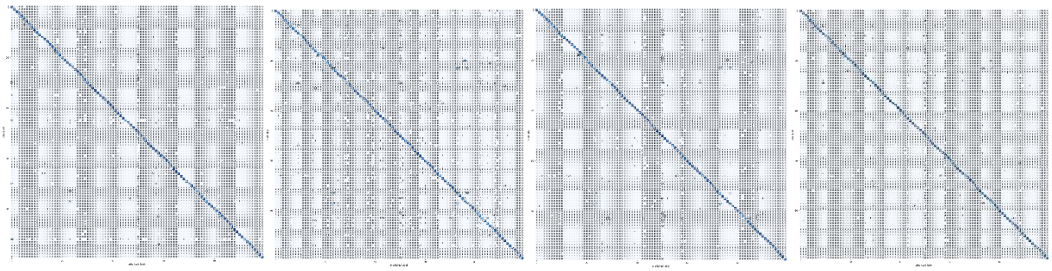
	Strings	N-1	N-2	N-3	N-4	N-5
Naïve Bayes	94.844%	88.066%	92.372%	94.672%	95.526%	**95.8%**
Logistic Regression	OOM	89.28%	95.052%	OOM	OOM	OOM

(Out of memory (OOM) is an often-undesired state of computer operation where no additional memory can be allocated for use by programs or the operating system.)

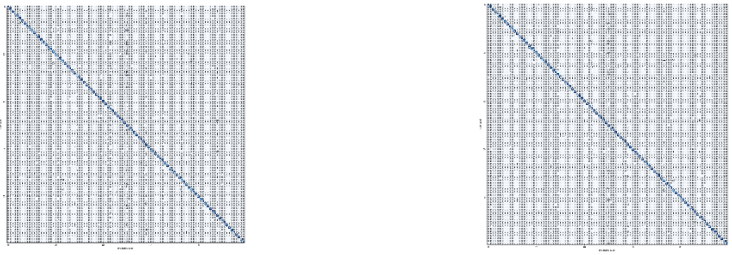
We were only able to run Naïve Bayes and 2 instances of Logistic Regression only as all the others will just crash when trying to allocate that much memory, this is a diagram to show the comparison:



Confusion Matrices:



Naïve Bayes String *Naïve Bayes N1* *Naïve Bayes N2* *Naïve Bayes N3*



Naïve Bayes N4 *Naïve Bayes N5*

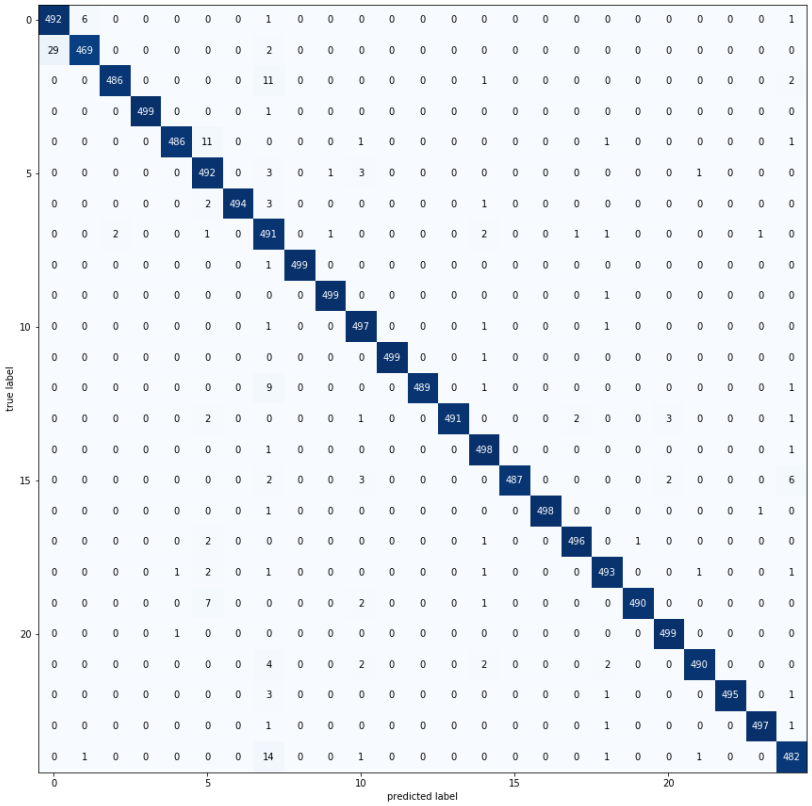
Model performance:

1-For the 25-Language Scope:

Our best model in the 25-language scope was the Logistic Regression with the accuracy of **98.464%** at the N-Gram of 3 on the 25 Languages that are as follows.

	precision	recall	f1-score	support
ara	0.94	0.98	0.96	500
arz	0.99	0.94	0.96	500
asm	1.00	0.97	0.98	500
azb	1.00	1.00	1.00	500
bul	1.00	0.97	0.98	500
deu	0.95	0.98	0.97	500
ell	1.00	0.99	0.99	500
eng	0.89	0.98	0.94	500
fas	1.00	1.00	1.00	500
fin	1.00	1.00	1.00	500
fra	0.97	0.99	0.98	500
heb	1.00	1.00	1.00	500
hin	1.00	0.98	0.99	500
hye	1.00	0.98	0.99	500
ita	0.98	1.00	0.99	500
jpn	1.00	0.97	0.99	500
kur	1.00	1.00	1.00	500
nld	0.99	0.99	0.99	500
por	0.98	0.99	0.98	500
roh	1.00	0.98	0.99	500
rus	0.99	1.00	0.99	500
spa	0.99	0.98	0.99	500
tha	1.00	0.99	0.99	500
tur	1.00	0.99	0.99	500
wuu	0.97	0.96	0.97	500
accuracy			0.98	12500
macro avg	0.99	0.98	0.98	12500
weighted avg	0.99	0.98	0.98	12500

With the Confusion Matrix as follows:



2-For the 100-Language Scope:

Our best model in the 100-language scope was the Naïve Bayes with the accuracy of **95.8%** at the N-Gram of 5 on the 100 Languages.

With the Confusion Matrix as follows: