Language difficulty project – Alpina

Table des matières

Results of models without cleaning data 2
Methods 2
Results of models with cleaning data 2
Conclusion 3

Results of models without cleaning data

We use directly TF-IDF method on our models according to the recommendations of the assistants. And for each model, we do a duplicate because we just test it with other parameters.

File : code > models_without_cleaning

	Model selected	Accuracy	Precision	Recall	F1_Score
0	Logistic Regression 1	0.466667	0.466667	0.466667	0.466667
1	Logistic Regression 2	0.516667	0.516667	0.516667	0.516667
2	KNN	0.318750	0.318750	0.318750	0.318750
3	KNN 2	0.318750	0.366667	0.366667	0.366667
4	Decision Tree Classifier 1	0.310417	0.310417	0.310417	0.310417
5	Decision Tree Classifier 2	0.317708	0.317708	0.317708	0.317708
6	Random Forest Classifier 1	0.404167	0.404167	0.404167	0.404167
7	Random Forest Classifier 2	0.427083	0.427083	0.427083	0.427083

Here the max result is the Logistic Regression 2 with this parameter:

- TfidfVectorizer(ngram_range=(1, 12), min_df=1, norm='l2', analyzer="char",sublinear_tf=True)
- LogisticRegression(solver ='lbfgs', penalty='l2', C=5)

Methods

First, we create a function “spacy_token(sentence)”, that tokenizes sentence, by using conversion into lowercase, removing punctuation, stop words and anonymous dates & people .

Then, a method evaluate (y_test, y_pred) is also created to check some accuracy measures of a model.

Results of models with cleaning data

We normalize first all sentences, then we create 2 methods to use later in the chapiter 4 training models. To train models, we train with the previous models. Because each model has a duplicate, we choose one which has the best score.

File : code > models_with_cleaning

	Model selected	Accuracy	Precision	Recall	F1_Score
0	Logistic Regression	0.514583	0.514583	0.514583	0.514583
1	KNN	0.390625	0.390625	0.390625	0.390625
2	Decision Tree Classifier	0.318750	0.318750	0.318750	0.318750
3	Random Forest Classifier	0.368750	0.368750	0.368750	0.368750

And the best one is the Logistic Regression with 0.51 of accuracy.

Conclusion

Finally, the model Logistic Regression without cleaning data is the best in our analyze.

However, we can suppose with another cleaning method, we can achieve a better result.