



HLG-MOS Machine Learning Project: Sharing ML Techniques and Algorithms, how to tackle large data sets and build international capability

Alex Measure, U.S. Bureau of Labour Statistics
Krystyna Piątkowska, Statistics Poland
Marta Kruczek-Szepel, Statistics Poland

AGENDA

- Is machine learning useful for official statistics?
- If so, how should we do it?
- Coming this Fall...
 - Pilot projects
 - Research
 - Tutorials

TUTORIAL

Text Classification



ml_tutorial.ipynb ☆

File Edit View Insert Runtime Tools Help Last saved at 2:15 PM

+ Code + Text

We have to import all the python libraries that we will need in our classification.

```
[ ] import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.feature_extraction.text import CountVectorizer
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import classification_report
```

Now we read our excel file to the dataframe (a dataframe is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns). I have prepared the file with 60 sentences: 20 in polish, 20 in english and 20 in spanish. The task for the classifier is to learn how to classify the sentence to the correct language.

Source: https://colab.research.google.com/drive/1Epn2eFRuFC_XyXtQ4qezGVBA5aAzqlh

GITHUB

Knowledge Sharing

github.com/statisticspoland/ecoicop_classification

Linear_SVC	ecoicop classification - initial commit	2 months ago
Logistic_Regression	ecoicop classification - initial commit	2 months ago
Naive_Bayes	ecoicop classification - initial commit	2 months ago
Random_Forest	ecoicop classification - initial commit	2 months ago
.gitignore	authors	2 months ago
categories_distribution.py	ecoicop classification - initial commit	2 months ago
ecoicop_histogram.png	ecoicop classification - initial commit	2 months ago
polish_stopwords.txt	ecoicop classification - initial commit	2 months ago
products_allshops_dataset.xlsx	ecoicop classification - initial commit	2 months ago
readme.md	authors	2 months ago
requirements.txt	ecoicop classification - initial commit	2 months ago



Source: https://github.com/statisticspoland/ecoicop_classification

ECOICOP APPLICATION

Klasyfikator produktów według ECOICOP

Niniejsza aplikacja służy do klasyfikacji produktów zgodnie z ECOICOP (Europejską Klasyfikacją Spożycia Indywidualnego według Celu). Aktualnie możemy klasyfikować produkty z pierwszej grupy (żywność i napoje bezalkoholowe).
Dokładność naszych algorytmów na próbie testowej wyniosła około 90%.

Produkty można wpisywać pojedynczo za pomocą odpowiedniego formularza albo wczytać cały plik z nazwami produktów. Tabelę z zaklasyfikowanymi produktami można wyeksportować do pliku Excela lub csv.

1. Wybierz sposób podawania produktów 2. Wybierz metodę klasyfikacji 3. Podaj nazwę produktu

Pojedynczo Logistic Regression Np. chleb pełnoziarnisty [Klasyfikuj](#) [Dodaj kolejny](#)

Tabela klasyfikacji







OSTATNIO DODANO:

PRODUKT: sos pomidorowy

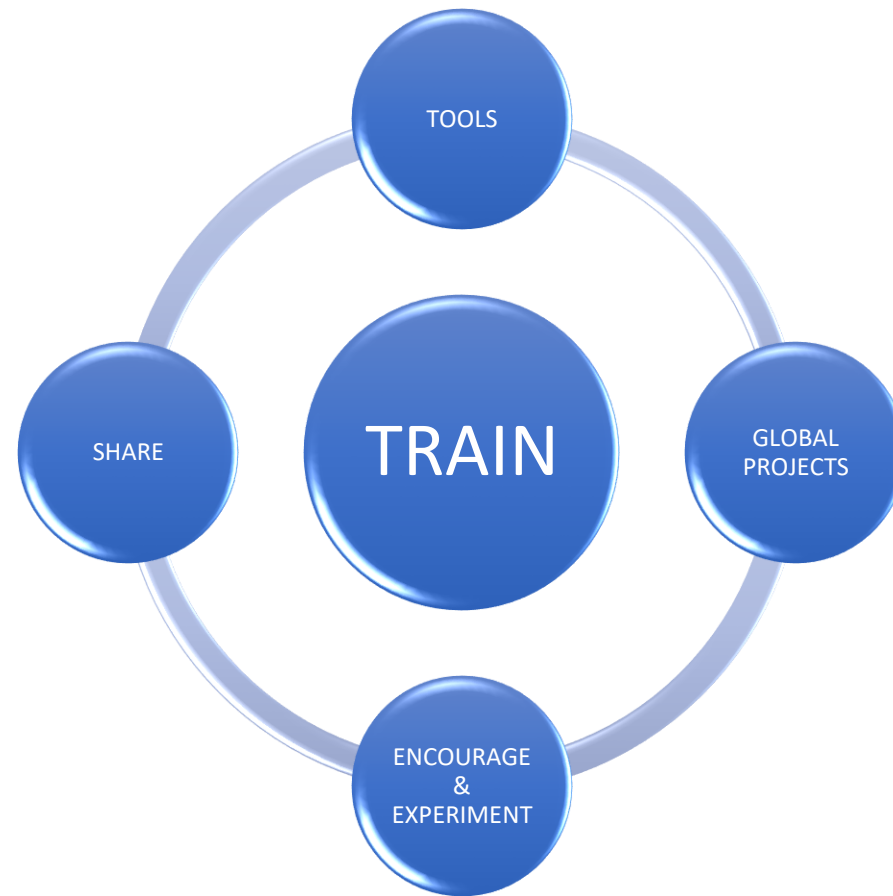
PROPONOWANA KATEGORIA: Sosy, przyprawy i jej prawdopodobieństwo wynosi 83.12 %

WYBRANA KATEGORIA: Sosy, przyprawy - prawdopodobieństwo: 83.12 %

[Eksportuj](#)

Lp.	Nazwa produktu	Kod ECOICOP	Kategoria	%
1	bułka kajzerka z sezamem	01.1.1.3	Pieczyno	85.33  
2	ptasie mleczko	01.1.8.4	Wyroby cukiernicze	80.72  
3	sos pomidorowy	01.1.9.1	Sosy, przyprawy	83.12  

SUMMARY



Q&A

