Urząd Statystyczny
w Poznaniu

Wdrożenie NACE Rev. 2.1 w krajowym rejestrze przedsiębiorstw przy użyciu algorytmów ML

wyzwania i wyniki

Ośrodek Inżynierii Danych

Grzegorz Bujwid
kierownik

19.11.2024 Poznań

Agenda

1. Początki
2. NACE Rev. 2 vs NACE Rev. 2.1
3. Prace projektowe
4. Podsumowanie

Początki

- 30.12.2022 - zaproszenie Departamentu Standardów i Rejestrów do udziału w planowanym grancie pn. *Przygotowanie do wdrożenia NACE Rev. 2.1 w rejestrach przedsiębiorstw*
- marzec 2023 - wysłanie wniosku do Eurostatu
- 1 października 2023 - start projektu
- Zadanie 8. *Wsparcie krajowych instytutów statystycznych w opracowywaniu i wdrożeniu oprogramowania lub innych metod (narzędzia sztucznej inteligencji i uczenia maszynowego) do automatycznego przeklasyfikowania krajowego rejestru przedsiębiorstw zgodnie z klasyfikacją NACE Rev. 2.1*

Agenda

1. Początki
2. NACE Rev. 2 vs NACE Rev. 2.1
3. Prace projektowe
4. Podsumowanie

NACE Rev. 2

- Data: 20 grudnia 2006

- Sekcje: 21
- Działy: 88
- Grupy: 272
- Klasy: 615

NACE Rev. 2.1

- Data: 10 października 2022 (+ 5773)

- Sekcje: 22 (+ 1)
- Działy: 87 (- 1)
- Grupy: 287 (+ 15)
- Klasy: 651 (+ 36)

NACE Rev. 2.1 – skąd ta aktualizacja?

- sposób wytwarzania towarów i usług w wielu rodzajach działalności uległ zmianie,
- pewne nowe rodzaje działalności nabrały znaczenia,
- inne straciły na znaczeniu,
- szybko zachodzące zmiany w środowisku informatycznym,
- powstania specjalistycznych rodzajów działalności służących ochronie środowiska,
- następstwo przyjęcia przez Komisję Statystyczną ONZ 5. rewizji Międzynarodowej Standardowej Klasyfikacji Rodzajów Działalności (ISIC Rev. 5),
- nowe rodzaje działalności, powstałe dzięki rozwojowi strukturalnemu, naukowemu i technologicznemu.

Agenda

1. Początki
2. NACE Rev. 2 vs NACE Rev. 2.1
3. Prace projektowe
4. Podsumowanie

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

PRACE PROJEKTOWE

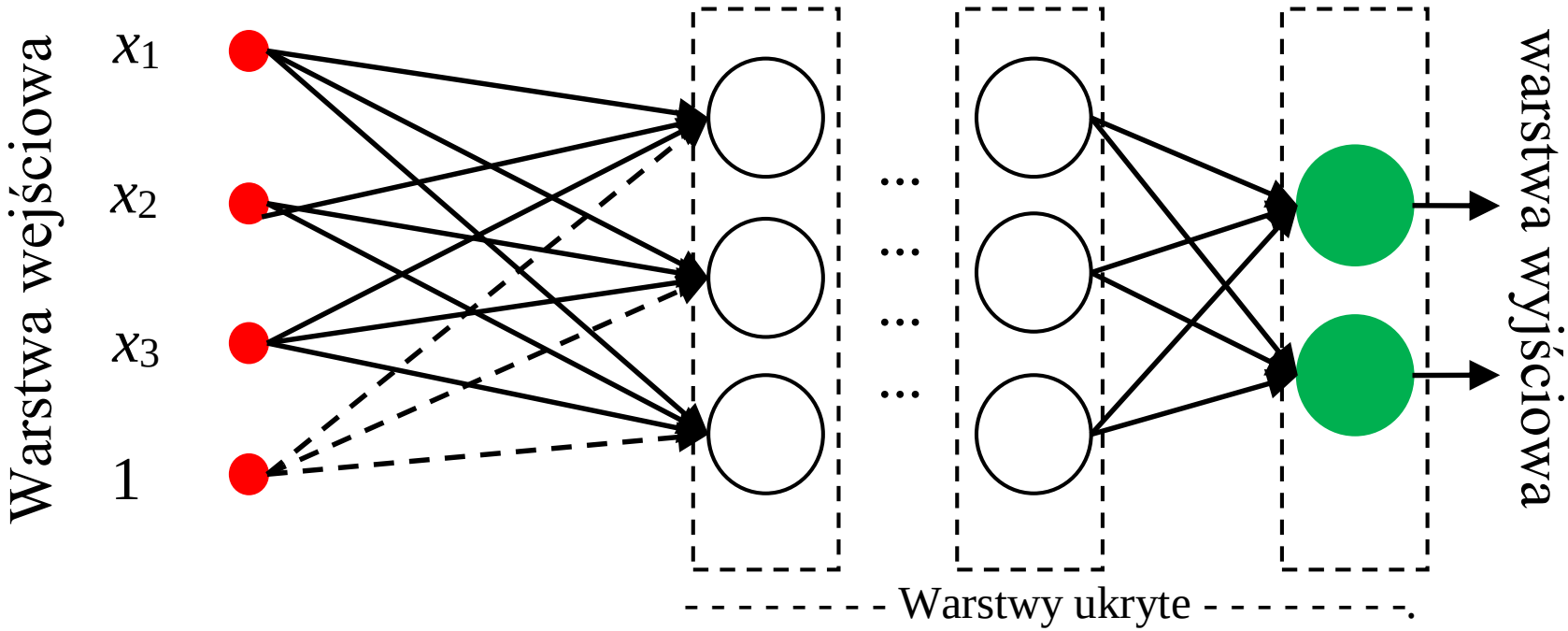
Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

Materiał językowy

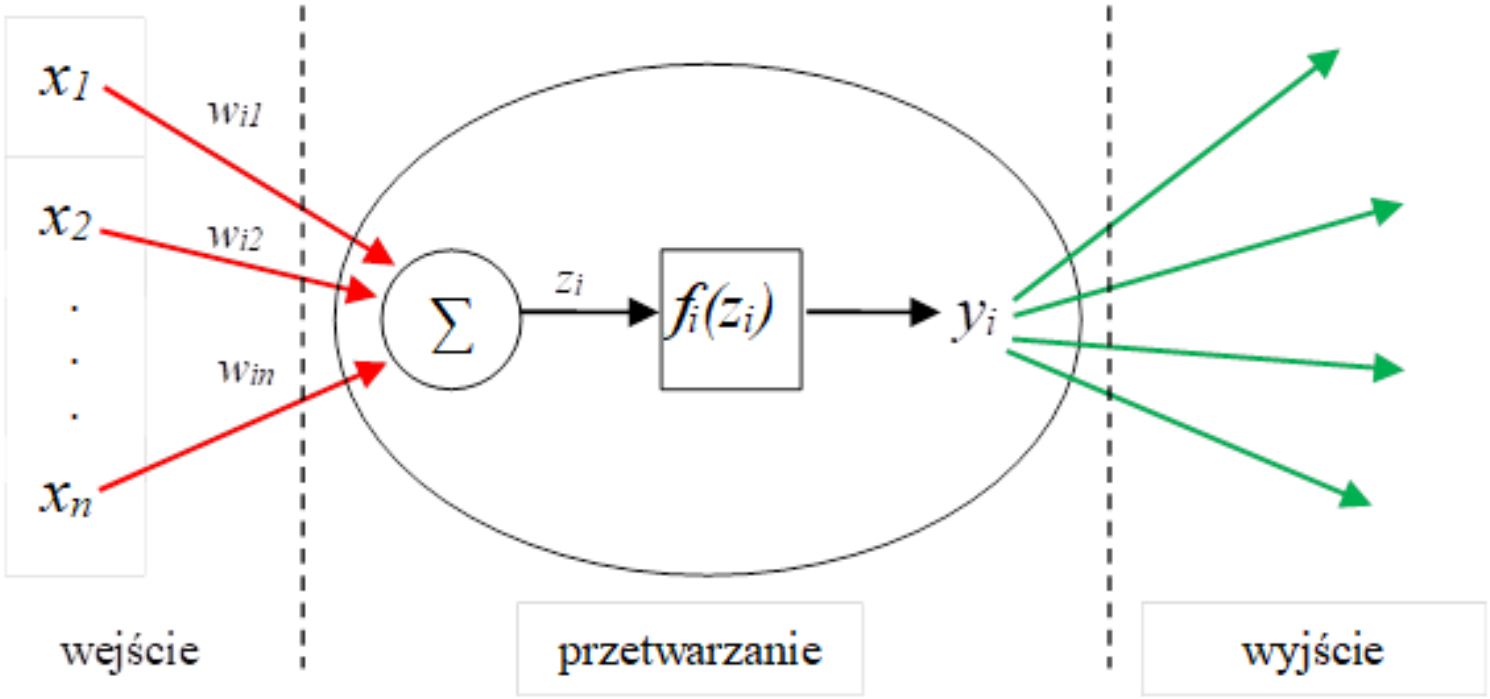
- nazwy i opisy klas PKD2007,
- nazwy i opisy klas PKD2025,
- opisy zawarte w nazwie podmiotu gospodarczego,
- zakres grupowania klas zawarty w kluczu powiazań.

Sieci neuronowe



Schemat przykładowej sieci neuronowej (na podst. Sikora W. (red.) Badania Operacyjne, PWE, 2008)

Sieci neuronowe



Schemat neuronu (na podst. Sikora W. (red.) Badania Operacyjne, PWE, 2008)

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

Klucz przejścia

- 1 do 1 – 491 klas
- n do 1
- 1 do n

1 DO 1 PRZYKŁAD

- NACE REV. 2 *A 01.12* Uprawa ryżu
- **NACE REV. 2.1** *A 01.12* Uprawa ryżu

- NACE REV. 2 *C 14.12* Produkcja odzieży roboczej
- **NACE REV. 2.1** *C 14.23* Produkcja odzieży roboczej

- NACE REV. 2 *C 25.50* Kucie, prasowanie, wytłaczanie i walcowanie metali; metalurgia proszków
- **NACE REV. 2.1** *C 25.40* Kucie i formowanie metali oraz metalurgia proszków

- NACE REV. 2 *J 63.91* Działalność agencji informacyjnych
- **NACE REV. 2.1** *J 60.31* Działalność agencji informacyjnych

- NACE REV. 2 *T 97.00* Gospodarstwa domowe zatrudniające pracowników
- **NACE REV. 2.1** *U 97.00* Gospodarstwa domowe zatrudniające pracowników

- NACE REV. 2 *S 96.09* Pozostała działalność usługowa, gdzie indziej niesklasyfikowana
- **NACE REV. 2.1** *T 96.99* Pozostała działalność usługowa, gdzie indziej niesklasyfikowana

Klucz przejścia

- 1 do 1 – 491 klas
- n do 1 – 242 klas -> 161 klas
- 1 do n

N DO 1 PRZYKŁAD

- NACE REV. 2 *A 01.13* UPRAWA WARZYW, WŁĄCZAJĄC MELONY, ORAZ UPRAWA ROŚLIN KORZENIOWYCH I ROŚLIN BULWIASTYCH
- NACE REV. 2 *A 01.19* Uprawy rolne inne niż wieloletnie, pozostałe
  - UPRAWA KORZENI I BULW, NP.: BRUKWI I BURAKÓW PASTEWNYCH,
  - uprawa koniczyny, lucerny, esparcety, kukurydzy pastewnej, łubinu pastewnego i pozostałych traw, kapusty pastewnej i podobnych roślin pastewnych,
  - uprawa nasion buraka i nasion roślin pastewnych,
  - uprawa kwiatów,
  - cięcie kwiatów oraz pąków kwiatowych,
  - uprawa nasion kwiatów.
- NACE REV. 2 *A 01.28* Uprawa roślin przyprawowych i aromatycznych oraz roślin do produkcji leków i wyrobów farmaceutycznych
  - UPRAWA CHILLI I PAPRYKI SŁODKIEJ (CAPSICUM)
  - uprawa roślin przyprawowych i aromatycznych, wieloletnich i innych niż wieloletnie:
    - pieprzu (z rodzaju Piper),
    - gałki muszkatołowej, kwiatu muszkatołowego i kardamonu,
    - anyżu, badianu i kopru włoskiego, przyprawowych lub aromatycznych,
    - cynamonu (canella),
    - goździków,
    - imbiru,
    - wanilii,
    - chmielu,
    - kminku i gorczycy,
  - uprawa roślin narkotykowych i odurzających.
- **NACE REV. 2.1** *A 01.13* UPRAWA WARZYW, WŁĄCZAJĄC MELONY, ORAZ UPRAWA ROŚLIN KORZENIOWYCH I ROŚLIN BULWIASTYCH

N DO 1 PRZYKŁAD

- NACE REV. 2 A 01.63 DZIAŁALNOŚĆ USŁUGOWA NASTĘPUJĄCA PO ZBIORACH
  - PODKLASA PKD 2007 W PEŁNI ODZWIERCIEDLA PODKLASĘ PKD 2025
- NACE REV. 2 A 01.64 OBRÓBKA NASION DLA CELÓW ROZMNAŻANIA
  - PODKLASA PKD 2007 W PEŁNI ODZWIERCIEDLA PODKLASĘ PKD 2025

- **NACE REV. 2.1** *A 01.63 DZIAŁALNOŚĆ USŁUGOWA NASTĘPUJĄCA PO ZBIORACH ORAZ OBRÓBKA NASION DLA CELÓW ROZMNAŻANIA ROŚLIN*

Klucz przejścia

- 1 do 1 – 491 klas
- n do 1 – 242 klas -> 161 klas
- 1 do n – 126 klas -> 277 klas

1 DO N PRZYKŁAD

- WYDOBYCIE NADKŁADU Z TERENÓW GÓRNICZYCH
- PRZYGOTOWANIE TERENU POD WYDOBYCIE

- NACE REV. 2 *F 43.12* Przygotowanie terenu pod budowę

- **NACE REV. 2.1** *B 05.10* WYDOBYWANIE WĘGLA KAMIENNEGO
- **NACE REV. 2.1** *B 05.20* WYDOBYWANIE WĘGLA BRUNATNEGO (LIGNITU)
- **NACE REV. 2.1** *B 07.10* GÓRNICTWO RUD ŻELAZA
- **NACE REV. 2.1** *B 07.21* GÓRNICTWO RUD URANU I TORU
- **NACE REV. 2.1** *B 07.29* GÓRNICTWO POZOSTAŁYCH RUD METALI NIEŻELAZNYCH
- **NACE REV. 2.1** *B 08.11* WYDOBYWANIE KAMIENI OZDOBNYCH, WAPIENIA, GIPSU, ŁUPKÓW ORAZ POZOSTAŁYCH KAMIENI I SKAŁ
- **NACE REV. 2.1** *B 08.12* WYDOBYWANIE ŻWIRU, PIASKU, GLINY I KAOLINU
- **NACE REV. 2.1** *B 08.91* WYDOBYWANIE MINERAŁÓW DLA PRZEMYSŁU CHEMICZNEGO ORAZ DO PRODUKCJI NAWOZÓW
- **NACE REV. 2.1** *B 08.92* WYDOBYWANIE TORFU
- **NACE REV. 2.1** *B 08.93* WYDOBYWANIE SOLI
- **NACE REV. 2.1** *B 08.99* POZOSTAŁE GÓRNICTWO I WYDOBYWANIE, GDZIE INDZIEJ NIESKLASYFIKOWANE
- **NACE REV. 2.1** *F 43.12* Przygotowanie terenu pod budowę

1 DO N PRZYKŁAD

- wydobycie nadkładu z terenów górniczych
- przygotowanie terenu pod wydobycie

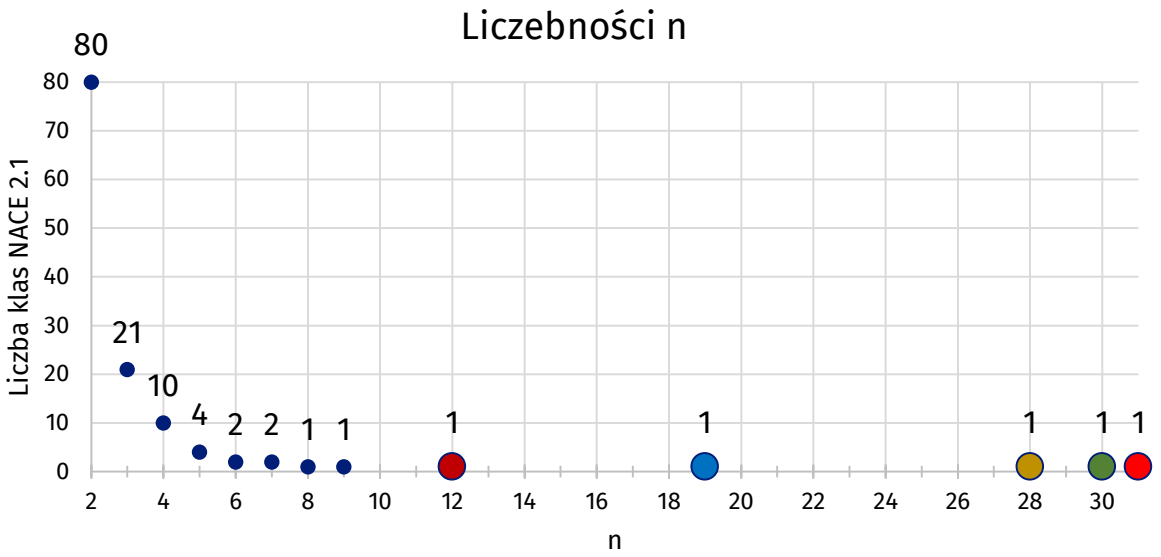
- NACE REV. 2 *F 43.12* Przygotowanie terenu pod budowę

- OCZYSZCZANIE PLACÓW BUDOWY
- PRZENOSZENIE ZIEMI, NP. WYKOPY, SKŁADOWANIE NIWELACJA, KOPANIE ROWÓW, ODSTRZAŁY
- PRZYGOTOWANIE TERENU POD WYDOBYCIE
- USUWANIE NADKŁADU I PROFILOWANIE TERENU NA PLACACH BUDOWY
- ODWODNIENIE PLACU BUDOWY
- ODWODNIENIE GRUNTÓW ROLNYCH LUB LEŚNYCH
- PRZYGOTOWANIE TERENU POD WYKOPALISKA ARCHEOLOGICZNE

- **NACE REV. 2.1** *B 05.10* Wydobywanie węgla kamiennego
- **NACE REV. 2.1** *B 05.20* Wydobywanie węgla brunatnego (lignitu)
- **NACE REV. 2.1** *B 07.10* Górnictwo rud żelaza
- **NACE REV. 2.1** *B 07.21* Górnictwo rud uranu i toru
- **NACE REV. 2.1** *B 07.29* Górnictwo pozostałych rud metali nieżelaznych
- **NACE REV. 2.1** *B 08.11* Wydobywanie kamieni ozdobnych, wapienia, gipsu, łupków oraz pozostałych kamieni i skał
- **NACE REV. 2.1** *B 08.12* Wydobywanie żwiru, piasku, gliny i kaolinu
- **NACE REV. 2.1** *B 08.91* Wydobywanie minerałów dla przemysłu chemicznego oraz do produkcji nawozów
- **NACE REV. 2.1** *B 08.92* Wydobywanie torfu
- **NACE REV. 2.1** *B 08.93* Wydobywanie soli
- **NACE REV. 2.1** *B 08.99* Pozostałe górnictwo i wydobywanie, gdzie indziej niesklasyfikowane
- **NACE REV. 2.1** *F 43.12* PRZYGOTOWANIE TERENU POD BUDOWĘ

1 DO N

- n ∈ {2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 16, 17, 19, 28, 30, 31, 32},
- klasy, które zostały rozdzielone na najwięcej klas NACE Rev. 2.1
  - F 43.12 Przygotowanie terenu pod budowę
  - G 47.89 Sprzedaż detaliczna pozostałych wyrobów prowadzona na straganach i targowiskach
  - N 82.99 Pozostała działalność wspomagająca prowadzenie działalności gospodarczej, gdzie indziej niesklasyfikowana
  - G 47.99 Sprzedaż detaliczna pozostała prowadzona poza siecią sklepową, straganami i targowiskami
  - G 47.91 Sprzedaż detaliczna prowadzona przez domy sprzedaży wysyłkowej lub Internet
- > 91% podziałów wg kluczy:
  - 1 do 2,
  - 1 do 3,
  - 1 do 4,
  - 1 do 5.



Liczba podmiotów

- BJS - Baza Jednostek Statystycznych:
  - Bjsum – jednostki umowne 30.955
  - Bjslo – jednostki lokalne 431.986
    - działalności drugorzędne 182.234
  - Bjspr – jednostki prawne 1.764.255
    - działalności drugorzędne 4.514.785

Wyzwania 1

- opóźnienia tłumaczenia,
- zmiany w tłumaczeniach,
- braki w tłumaczeniu

PRZYKŁAD 1. BRAKÓW W TŁUMACZENIU

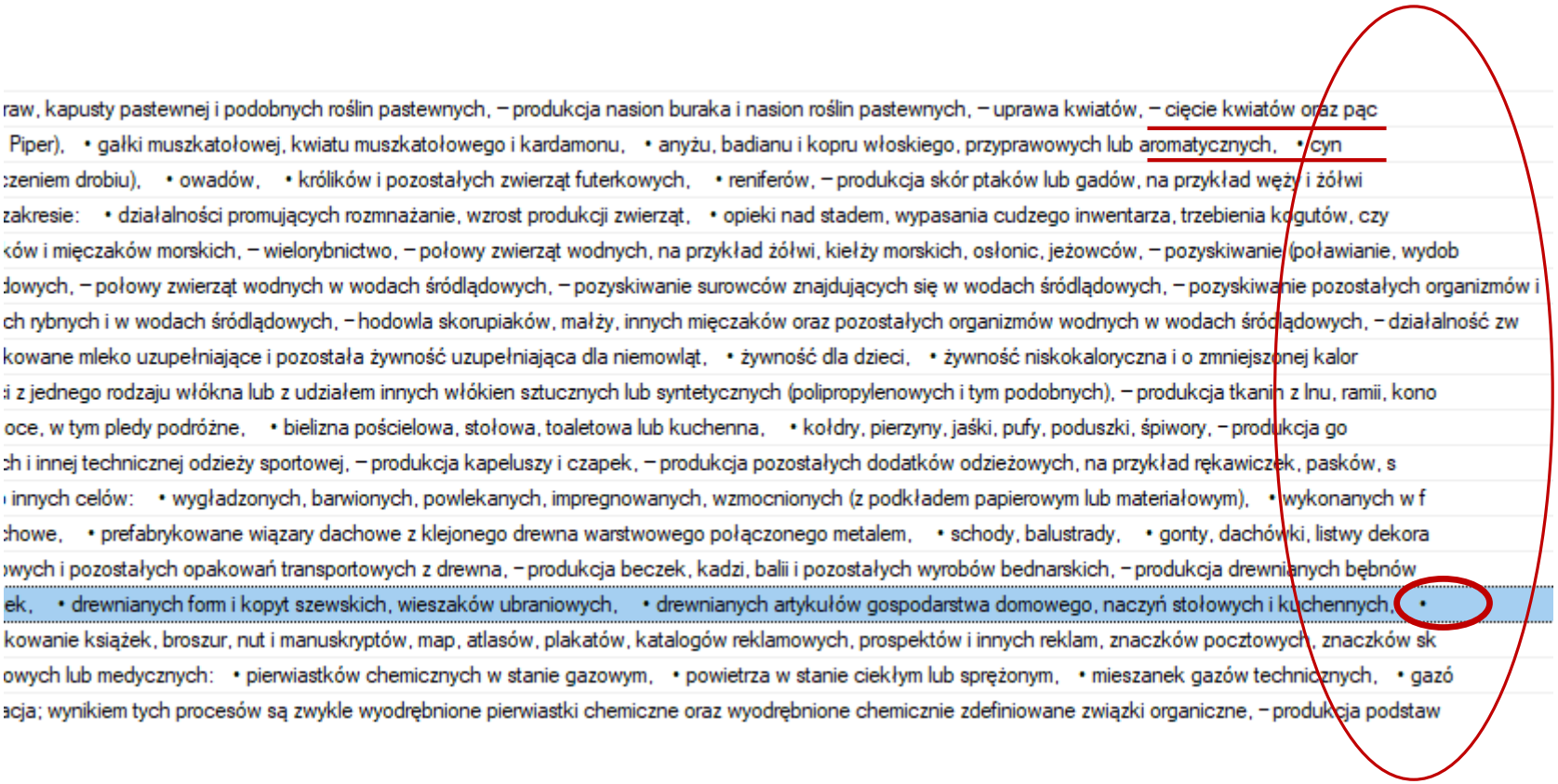
- NACE REV. 2 *A 01.13* Uprawa warzyw, włączając melony, oraz uprawa roślin korzeniowych i roślin bulwiastych
- NACE REV. 2 *A 01.19* Uprawy rolne inne niż wieloletnie, pozostałe
  - uprawa korzeni i bulw, np.: brukwi i buraków pastewnych,
  - uprawa koniczyny, lucerny, esparcety, kukurydzy pastewnej, łubinu pastewnego i pozostałych traw, kapusty pastewnej i podobnych roślin pastewnych,
  - uprawa nasion buraka i nasion roślin pastewnych,
  - uprawa kwiatów,
  - cięcie kwiatów oraz pąc

- NACE REV. 2 *A 01.28* Uprawa roślin przyprawowych i aromatycznych oraz roślin do produkcji leków i wyrobów farmaceutycznych
  - uprawa chilli i papryki słodkiej (Capsicum),
  - uprawa roślin przyprawowych i aromatycznych, wieloletnich i innych niż wieloletnie:
    - pieprzu (z rodzaju Piper),
    - gałki muszkatołowej, kwiatu muszkatołowego i kardamonu,
    - anyżu, badianu i kopru włoskiego, przyprawowych lub aromatycznych,
    - cyn

PRZYKŁAD 1. BRAKÓW W TŁUMACZENIU

- NACE REV. 2 *A 01.13* Uprawa warzyw, włączając melony, oraz uprawa roślin korzeniowych i roślin bulwiastych
- NACE REV. 2 *A 01.19* Uprawy rolne inne niż wieloletnie, pozostałe
  - uprawa korzeni i bulw, np.: brukwi i buraków pastewnych,
  - uprawa koniczyny, lucerny, esparcety, kukurydzy pastewnej, łubinu pastewnego i pozostałych traw, kapusty pastewnej i podobnych roślin pastewnych,
  - uprawa nasion buraka i nasion roślin pastewnych,
  - uprawa kwiatów,
  - cięcie kwiatów oraz pąc PĄKÓW KWIATOWYCH,
  - UPRAWA NASION KWIATÓW.
- NACE REV. 2 *A 01.28* Uprawa roślin przyprawowych i aromatycznych oraz roślin do produkcji leków i wyrobów farmaceutycznych
  - uprawa chilli i papryki słodkiej (Capsicum),
  - uprawa roślin przyprawowych i aromatycznych, wieloletnich i innych niż wieloletnie:
    - pieprzu (z rodzaju Piper),
    - gałki muszkatołowej, kwiatu muszkatołowego i kardamonu,
    - anyżu, badianu i kopru włoskiego, przyprawowych lub aromatycznych,
    - cyn CYNAMON (CANELLA),
    - GOŹDZIKI,
    - IMBIR,
    - WANILIA,
    - CHMIEL,
    - KMINEK I GORCZYCA,
  - UPRAWA ROŚLIN NARKOTYKOWYCH I ODURZAJĄCYCH.

PRZYKŁAD 2. BRAKÓW W TŁUMACZENIU



raw, kapusty pastewnej i podobnych roślin pastewnych, – produkcja nasion buraka i nasion roślin pastewnych, – uprawa kwiatów, – cięcie kwiatów oraz pąc
Piper), • gałki muszkatołowej, kwiatu muszkatołowego i kardamonu, • anyżu, badianu i kopru włoskiego, przyprawowych lub aromatycznych, • cyn
czeniem drobiu), • owadów, • królików i pozostałych zwierząt futerkowych, • reniferów, – produkcja skór ptaków lub gadów, na przykład węży i żółwi
zakresie: • działalności promujących rozmnażanie, wzrost produkcji zwierząt, • opieki nad stadem, wypasania cudzego inwentarza, trzebienia kogutów, czy
ków i mięczaków morskich, – wielorybnictwo, – połowy zwierząt wodnych, na przykład żółwi, kiełży morskich, osłonic, jeżowców, – pozyskiwanie (poławianie, wydob
dowych, – połowy zwierząt wodnych w wodach śródlądowych, – pozyskiwanie surowców znajdujących się w wodach śródlądowych, – pozyskiwanie pozostałych organizmów i
ch rybnych i w wodach śródlądowych, – hodowla skorupiaków, małży, innych mięczaków oraz pozostałych organizmów wodnych w wodach śródlądowych, – działalność zw
kowane mleko uzupełniające i pozostała żywność uzupełniająca dla niemowląt, • żywność dla dzieci, • żywność niskokaloryczna i o zmniejszonej kalor
i z jednego rodzaju włókna lub z udziałem innych włókien sztucznych lub syntetycznych (polipropylenowych i tym podobnych), – produkcja tkanin z lnu, ramii, kono
oce, w tym pledy podróżne, • bielizna pościelowa, stołowa, toaletowa lub kuchenna, • kołdry, pierzyny, jaśki, pufy, poduszki, śpiwory, – produkcja go
ch i innej technicznej odzieży sportowej, – produkcja kapeluszy i czapek, – produkcja pozostałych dodatków odzieżowych, na przykład rękawiczek, pasków, s
innych celów: • wygładzonych, barwionych, powlekanych, impregnowanych, wzmocnionych (z podkładem papierowym lub materiałowym), • wykonanych w f
howe, • prefabrykowane wiązary dachowe z klejonego drewna warstwowego połączonego metalem, • schody, balustrady, • gonty, dachówki, listwy dekora
wych i pozostałych opakowań transportowych z drewna, – produkcja beczek, kadzi, balii i pozostałych wyrobów bednarskich, – produkcja drewnianych bębnów
ek, • drewnianych form i kopyt szewskich, wieszaków ubraniowych, • drewnianych artykułów gospodarstwa domowego, naczyń stołowych i kuchennych, •
kowanie książek, broszur, nut i manuskryptów, map, atlasów, plakatów, katalogów reklamowych, prospektów i innych reklam, znaczków pocztowych, znaczków sk
owych lub medycznych: • pierwiastków chemicznych w stanie gazowym, • powietrza w stanie ciekłym lub sprężonym, • mieszanek gazów technicznych, • gazó
acja; wynikiem tych procesów są zwykle wyodrębnione pierwiastki chemiczne oraz wyodrębnione chemicznie zdefiniowane związki organiczne, – produkcja podstaw

PRZYKŁAD 3. BRAKÓW W TŁUMACZENIU

- NACE Rev. 2 *A 16.29* Produkcja pozostałych wyrobów z drewna; produkcja wyrobów z korka, słomy i materiałów używanych do wyplatania
    - produkcja różnych wyrobów z drewna:
        - drewnianych trzonków i korpusów do narzędzi, mioteł, szczotek,
        - drewnianych form i kopyt szewskich, wieszaków ubraniowych,
        - drewnianych artykułów gospodarstwa domowego, naczyń stołowych i kuchennych,
        -

PRZYKŁAD 3. BRAKÓW W TŁUMACZENIU

- NACE Rev. 2 *A 16.29* Produkcja pozostałych wyrobów z drewna; produkcja wyrobów z korka, słomy i materiałów używanych do wyplatania
  - produkcja różnych wyrobów z drewna:
    - drewnianych trzonków i korpusów do narzędzi, mioteł, szczotek,
    - drewnianych form i kopyt szewskich, wieszaków ubraniowych,
    - drewnianych artykułów gospodarstwa domowego, naczyń stołowych i kuchennych,
    - DREWNIANYCH FIGUREK I OZDÓB, DREWNIANYCH MARKIETERII, INTARSJI,
    - DREWNIANYCH PUDEŁEK NA BIŻUTERIĘ, SZTUĆCE I PODOBNE ARTYKUŁY,
    - DREWNIANYCH SZPUL, MOTKÓW, WRZECION, SZPULEK NA NICI DO SZYCIA I PODOBNYCH ARTYKUŁÓW Z TOCZONEGO DREWNA,
    - INNYCH ARTYKUŁÓW DREWNIANYCH,
  - PRODUKCJA DREWNIANYCH RAM DO PŁÓCIEN ARTYSTYCZNYCH,
  - OBRÓBKA NATURALNEGO KORKA, PRODUKCJA KORKA Z GRANULATU,
  - PRODUKCJA KORKA DO BUTELEK,
  - PRODUKCJA WYROBÓW Z NATURALNEGO LUB KORKA Z GRANULATU, W TYM WYKŁADZIN PODŁOGOWYCH,
  - PRODUKCJA PLECIONEK I WYROBÓW Z MATERIAŁÓW DO WYPLATANIA: MAT, MAT SŁOMIANYCH, EKRANÓW, ETUI ITP.,
  - PRODUKCJA KOSZYKÓW I WYROBÓW Z WIKLINY,
  - PRODUKCJA DREWNIANYCH RAM DO LUSTER I OBRAZÓW,
  - PRODUKCJA UCHWYTÓW DO PARASOLI, LASEK I PODOBNYCH,
  - PRODUKCJA BLOKÓW DO WYROBU FAJEK.

Wyzwania 2

- opóźnienia tłumaczenia,
- zmiany w tłumaczeniach,
- braki w tłumaczeniu (min. 166 opisów powiązania – 18%),
- opis działalności – BRAK,
- uniknięcie błędów typu false-positive w przypadku nazwisk odapelatywnych, których źródłem były np. nazwy zawodów,
- ogrom działalności jedynie z imieniem i nazwiskiem w firmie,
- moc obliczeniowa,
- dostarczenie poprawnych wyników.

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

Algorytm ML

- Model BERT (Bidirectional Encoder Representations from Transformers)
  - dwukierunkowość,
  - transformery,
  - pretrening,
  - fine-tuning,
  - zastosowanie,
  - brak komponentów rekurencyjnych – trening przebiega szybciej.

Algorytm ML

- Sentence-BERT (SBERT)
  - wielojęzykowy model
  - sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

- Miara podobieństwa - podobieństwo cosinusowe:
  - skalowalność,
  - niewrażliwość na długość,
  - uwzględnienie kierunku
  - dobra sprawność w NLP,
  - interpretowalność – wynik z przedziału od –1 do 1

$$cos(\theta) = \frac{a \times b}{|a||b|}.$$

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

PRACE PROJEKTOWE

Zadanie 8 zostało podzielone na mniejsze procesy:

1) prace testowe na próbnym zbiorze danych,
2) wytypowanie zbioru najbardziej optymalnych algorytmów uczenia maszynowego dla zadania klasyfikacyjnego,
3) zebranie, przygotowanie oraz oczyszczanie danych, które będą użyte do modelu,
4) analiza danych,
5) wybór modelu algorytmu uczenia maszynowego,
6) podział zbioru danych na zbiór uczący i testowy,
7) trenowanie modelu,
8) dostosowywanie parametrów w celu poprawy wydajności,
9) ocena poprawności działania modelu,
10) przeklasyfikowanie kodami NACE Rev.2.1 docelowego zbioru danych wytrenowanym modelem i przekazanie otrzymanych wyników do oceny eksperckiej.

Agenda

1. Początki
2. NACE Rev. 2 vs NACE Rev. 2.1
3. Prace projektowe
4. Podsumowanie

PODSUMOWANIE

- Usunięcie imion z firmy potrafiło dać dziwne wyniki, np.:
  - „Zoko Baby Aneta Kopeć” → „Zoko Kopeć”
  - „Józefa Fabian” → „”
- „Kraina Pupila”

0,7164	Sprzedaż detaliczna artykułów używanych
...	...
0,5725	Sprzedaż detaliczna kwiatów, roślin, nawozów, żywych zwierząt domowych i karmy dla zwierząt domowych

- „VIVIAN STUDIO FRYZJERSKIE Natalia Szpyrka”

0,7287	Działalność kosmetyczna
0,7062	Działalność fryzjerska

- „LUM VINTAGE STORE Ewa Koźlakiewicz”

0,6674	Sprzedaż detaliczna artykułów używanych
0,6171	Sprzedaż detaliczna mebli, sprzętu oświetleniowego, artykułów stołowych oraz pozostałych artykułów użytku domowego

Ośrodek Inżynierii Danych

Grzegorz Bujwid
kierownik