Wstępne przetwarzanie danych: ekstrakcja cech i analiza tekstu

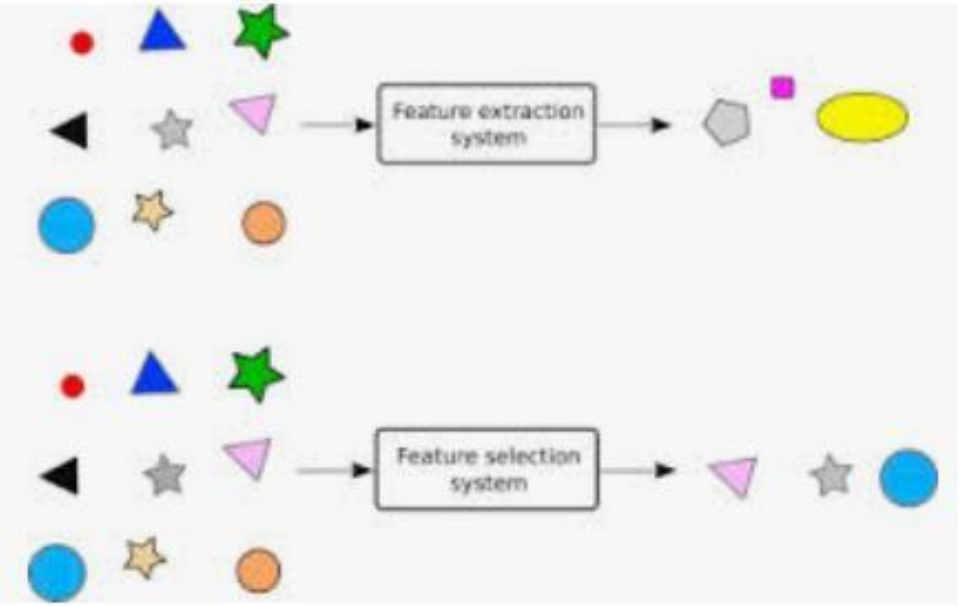
iBigWorld:
Innovations for Big Data in a Real World

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wybór cech(Feature selection , FS)

próbuje wybrać odpowiednie podzbiory istotnych cech bez znaczącej utraty informacji



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wybór cech

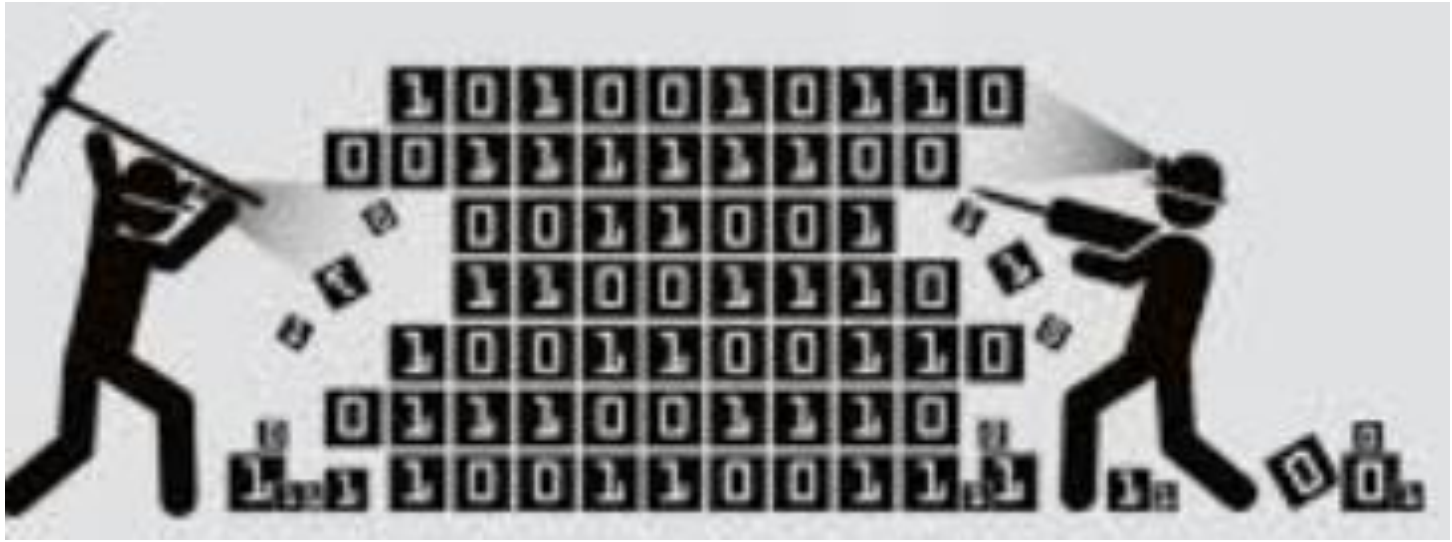
- **VectorSlicer**: użytkownik ręcznie wybiera podzbiór funkcji
- **RFormula**: wybiera funkcje zdefiniowane wzorem modelu R
- **Chi-Squared selector**: szereguje cechy kategoryczne za pomocą chi-kwadrat od klasy. Następnie wybiera najbardziej zależne funkcje.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Przetwarzanie tekstu

spróbuj uporządkować tekst wejściowy, uzyskując uporządkowane wzorce informacji



Co-funded by the
Erasmus+ Programme
of the European Union

Metody analizy tekstu

TF-IDF: To narzędzie ma na celu określenie znaczenia każdego terminu dla dokumentu, biorąc pod uwagę pełny zestaw dokumentów.

Term Frequency (TF) mierzy, ile dany termin pojawia się w dokumentach

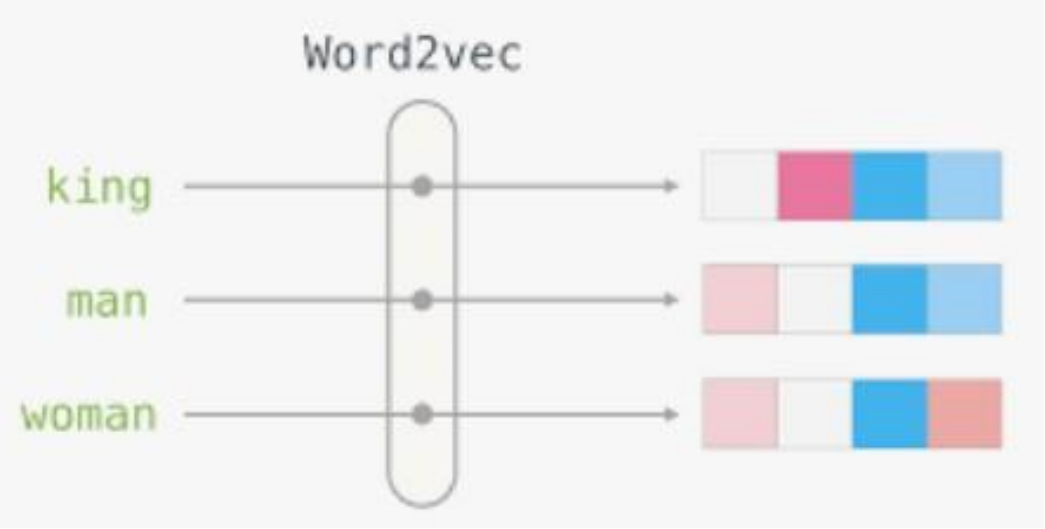
Inverse Document Frequency (**IDF**) mierzy, ile informacji dostarcza termin według jego częstotliwości w dokumencie.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody analizy tekstu

Word2Vec: pobiera korpus tekstu jako dane wejściowe i wyprowadza wektory słów jako dane wyjściowe. Najpierw buduje słownictwo z tekstu, a następnie uczy się wektorowej reprezentacji słów

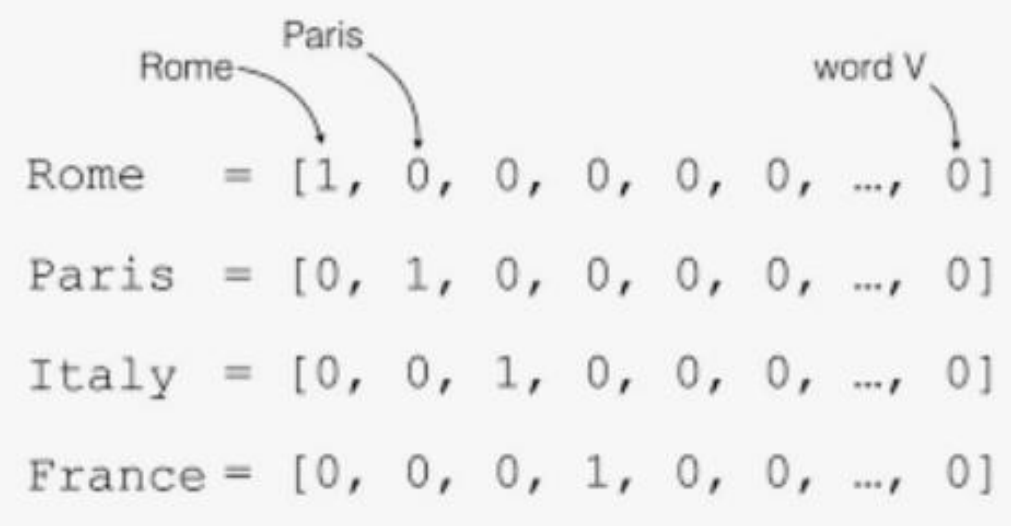


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Metody analizy tekstu

CountVectorizer: przekształca korpus w zbiór wektorów o liczbie tokenów. Wyodrębnia słownik za pomocą estymatora i zlicza liczbę wystąpień dla każdego terminu



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody analizy tekstu

Tokenizer: dzieli tekst na poszczególne terminy za pomocą wyrażeń regularnych lub wyrażeń regularnych

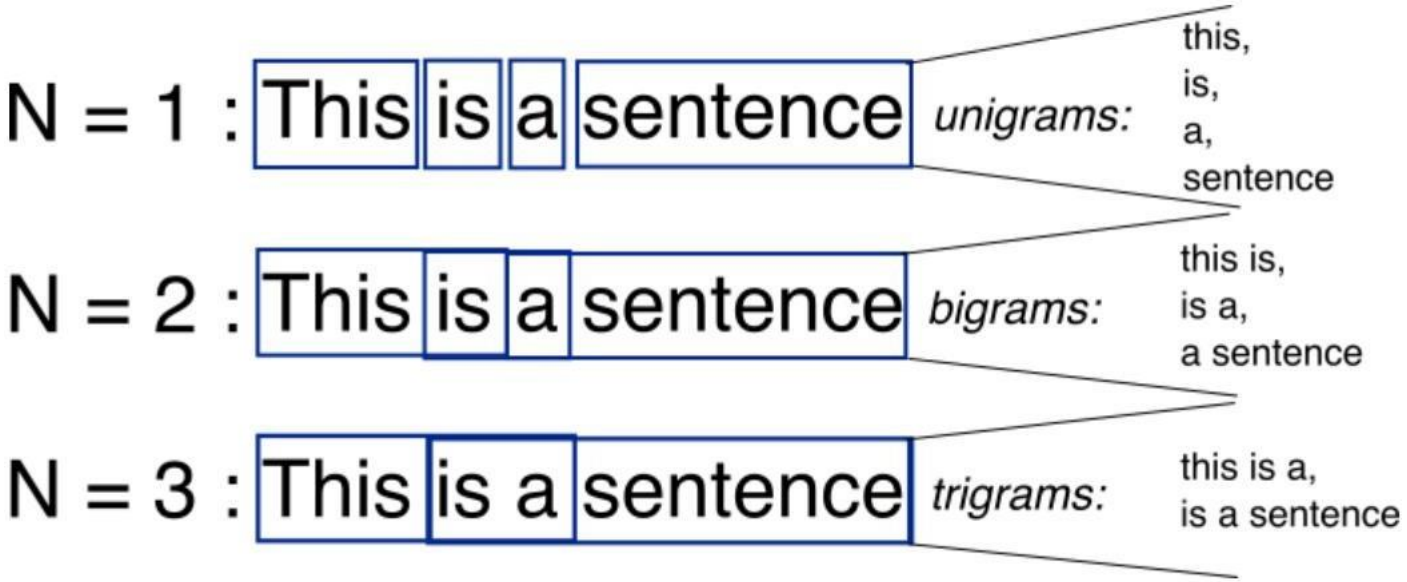
StopWordsRemover: usuwa nieodpowiednie słowa z wprowadzonego tekstu. Jak parametr, podana jest lista słów zatrzymania (stop words)

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody analizy tekstu

n-gram: generuje ciągi terminów z n gramów, z których każdy składa się z ciągu n kolejnych słów oddzielonych spacjami



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

FS methods аналізу тексту

Ant Colony Optimization (ACO) w celu znalezienia optymalnego podzbioru cech.

Działa równolegle w Hadoop MapReduce niektóre części algorytmu, takie jak: tokenizacja, obliczanie stopni asocjacji i ocena decyzji.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

FS methods text mining

Evolutionary feature weighting model- Ewolucyjny model ważenia obiektów do nauki wag obiektów na mapie. Wprowadzili fazę Zmniejsz, dodając wagi i używając progu, aby wybrać najbardziej odpowiednią instancję

ANOVA, Kruskal–Wallis, and Friedman test) na podstawie testu statystycznego. Wszystkie zostały zrównoleglone w Hadoop MapReduce, więc każda funkcja jest oceniana niezależnie

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

FS methody analizy tekstu

Metoda selekcji cech oparta na prywatności różnicowej (szum Laplace'a) i pomiarze wskaźnika Giniego została zaimplementowana przy użyciu ogólnego modelu MapReduce.

Ujednolicona struktura wykorzystująca pamięć binarnej macierzy korelacji (CMM) do przechowywania i pobierania wzorców przy użyciu arytmetyki macierzy. Proponują sekwencyjne obliczanie CMM, a następnie dystrybucję ich na Hadoop w celu uzyskania ostatecznych współczynników

Co-funded by the Erasmus+ Programme of the European Union

Wizualizacja big data

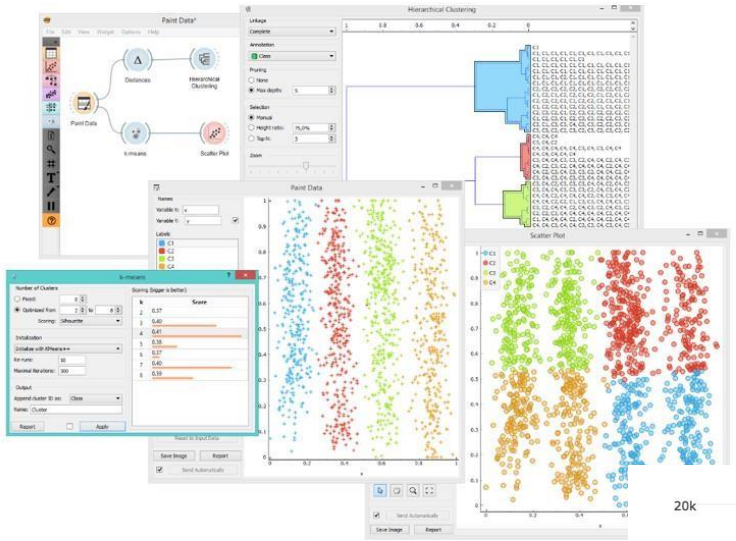
- jasno i skutecznie przekazywać informacje użytkownikom;
- pomoc użytkownikom w analizie i wyciąganiu wniosków na temat danych i wiedzy;
- uczynić złożone dane bardziej dostępnymi, zrozumiałymi i przyjaznymi dla użytkownika;
- pomoc w identyfikacji wzorców, zrozumieniu pomysłów, zbadaniu źródeł danych.

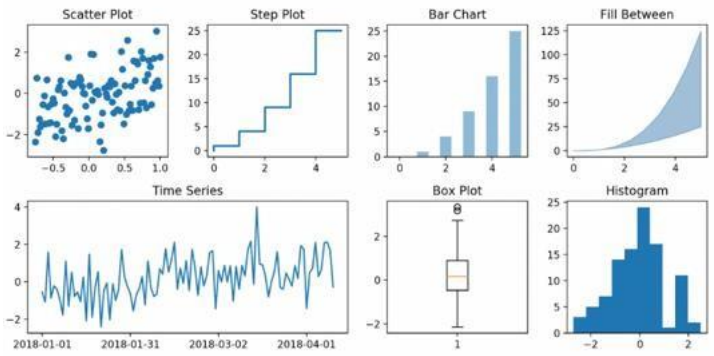
iBigWorld:2020-1-PL01-KA203-082197

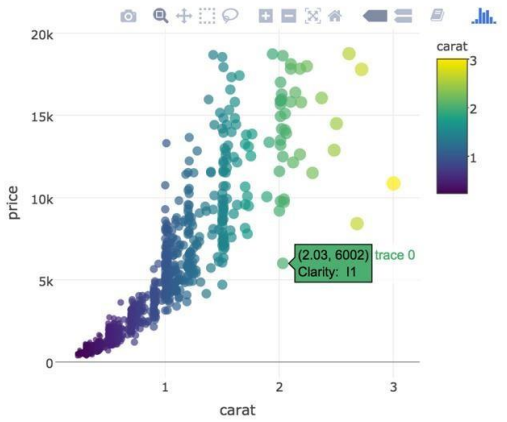
Co-funded by the Erasmus+ Programme of the European Union

Wizualizacja Big Data: narzędzia i platformy

- **Matplotlib**
- **Orange**
- **Tableau**
- **Jupyter**







iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dziękuję za uwagę!

Co-funded by the
Erasmus+ Programme
of the European Union