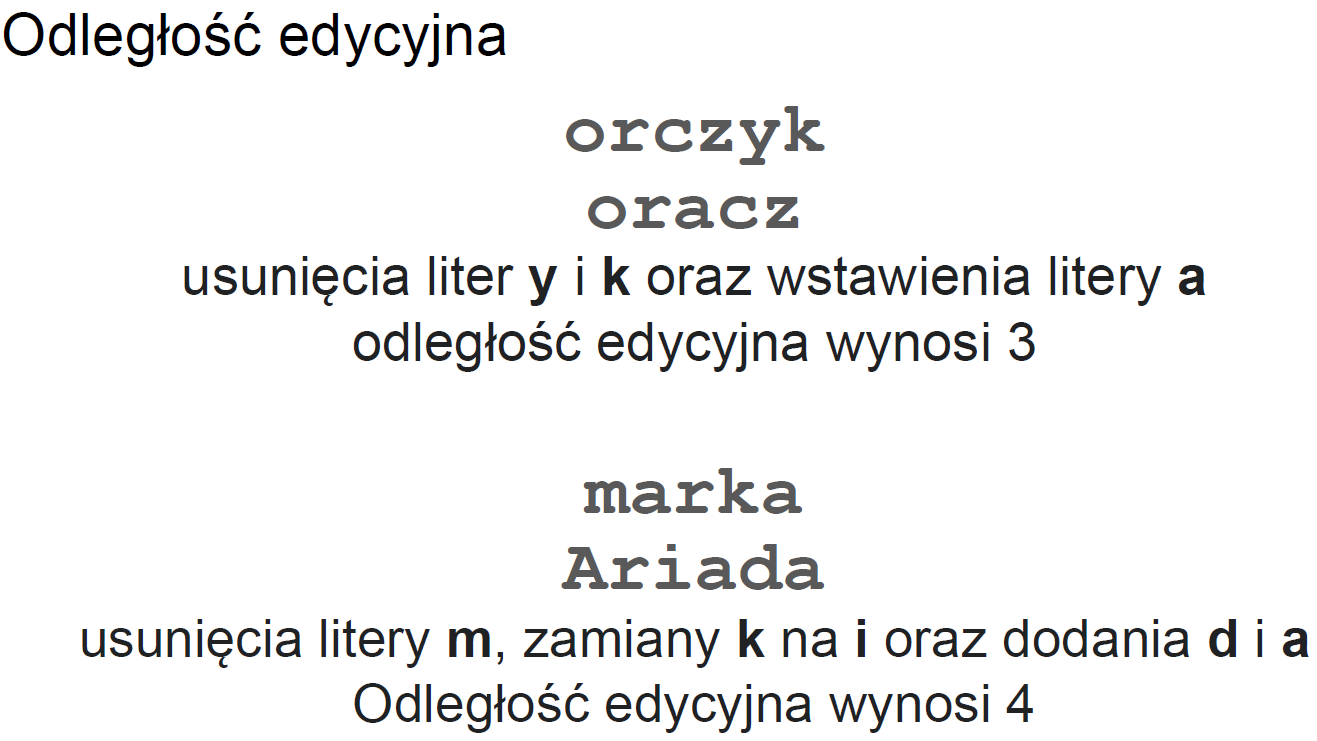
**Stopwords** – w terminologii Wikipedii najczęściej występujące słowa języka. Na ogół nie niosą ze sobą żadnych istotnych treści, więc Wikipedia powinna ignorować próbę ich wyszukiwania, m.in. aby nie obciążać serwera.

**Stemming** – jest to proces usunięcia ze słowa końcówki fleksyjnej pozostawiając tylko temat wyrazu. Np. „connection”, „connections”, „connective”, „connected”, „connecting” poddane stemmingowi dadzą ten sam wynik, czyli słowo „connect”

**Lematyzacja** - sprowadzenie słowa do jego podstawowej postaci. W przypadku czasownika będzie to bezokolicznik, w przypadku rzeczownika – mianownik liczby pojedynczej.

Ala ma kota. ---lematyzacja---> Ala mieć kot.

**Stemming i lematyzacja** generują podstawową formę odmienionych słów.  
Różnica polega na tym, że stem może nie być rzeczywistym słowem, podczas gdy lemma jest prawdziwym słowem w języku.



**Dokument tekstowy** - Typ "kontenera", który jest zbudowany z ciągu linii tekstu.

**Klasyfikacja dokumentów tekstowych** - Segregowanie dokumentów według określonego porządku

**Wektorowa reprezentacja dokumentu tekstowego:**

**Bag-of-words** - Dokument reprezentowany jest przez wektor, w którym poszczególne pozycje

Odpowiadają określonemu słowu ze słownika, a wartości na tych pozycjach liczbie wystąpień słowa w dokumencie.

- prosty w zrozumieniu i wdrożeniu

- potencjalnie duży rozmiar słownika

- rzadka reprezentacja danych

- nie uwzględnianie znaczenia (kontekstu) słów

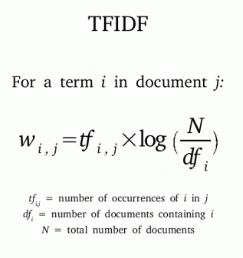
**TF-IDF -** ang. TF – term frequency, IDF – inverse document frequency **-** Ważenie częstością termów – Odwrotna częstość w dokumentach

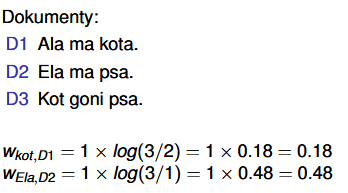
- mały koszt obliczeń

- pozwala wskazać najbardziej charakterystyczne terminy w dokumencie

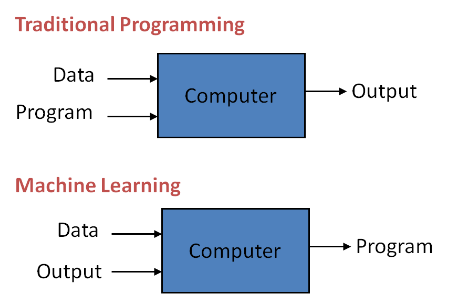
pozwala łatwo obliczyć podobieństwo między dwoma dokumentami

- opiera się na bag-of-words i cierpi na podobne wady, tj. duży wymiar wektora, nieuwzględnianie kontekstu słów



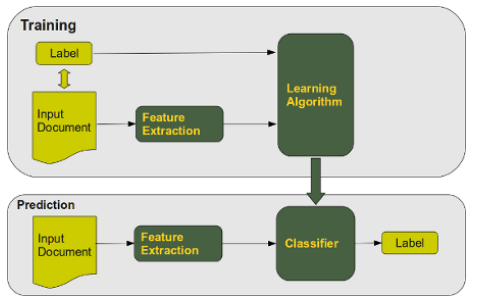


**Uczenie maszynowe -** samouczenie się maszyn albo systemy uczące się (ang. machine learning) – obszar sztucznej inteligencji poświęcony algorytmom które poprawiają się automatycznie poprzez doświadczenie. Algorytmy uczenia maszynowego budują model matematyczny na podstawie przykładowych danych, zwanych danymi treningowymi, w celu prognozowania lub podejmowania decyzji bez bycia bezpośrednio przez człowieka zaprogramowanym do tego.



**Klasyfikacja** - Problemem identyfikacji, do której z kategorii (subpopulacji) należy nowa obserwacja, na podstawie zestawu treningowego zawierającego obserwacje (lub przypadki), których przynależność do kategorii jest znana.

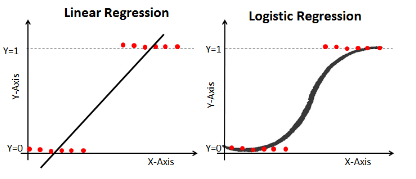
**Budowa klasyfikatora**



Zastosowania klasyfikatorów: - filtr spamu, kategoryzacja artykułów (np. sport, kulinarny itd.)

**Algorytmy klasyfikacji:**

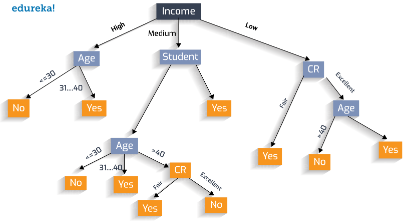
**Linear Regression, Logistic Regression:**

****

- klasyfikator binarny

- model statystyczny

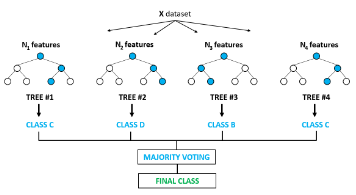
**DecisionTree**



- klasyfikator wieloklasowy

- pozyskiwanie wiedzy na podstawie przykładów

**RandomForest**

****

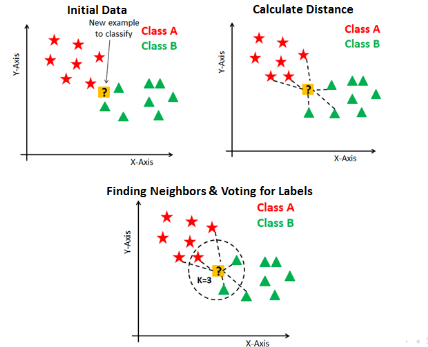
- konstruowanie wielu drzew decyzyjnych

- generowanie klasy, która jest dominantą klas pojedynczych drzew

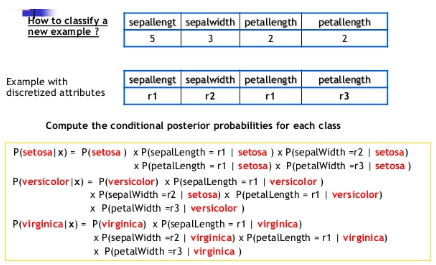
- losowanie danych do budowy drzew przy użyciu techniki bootstrap

- zmniejsza wariancję modelu, bez zwiększania obciążenia

**kNN**

****

**NaiveBayes**

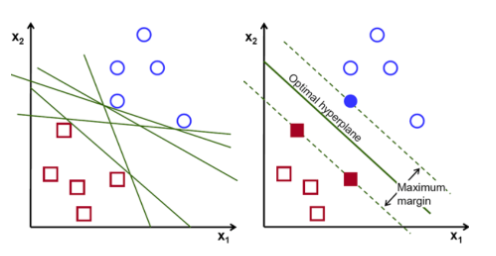
****

- prosty klasyfikator probabilistyczny

- oparty na założeniu o wzajemnej niezależności zmiennych niezależnych

- korzysta z twierdzenia Bayesa

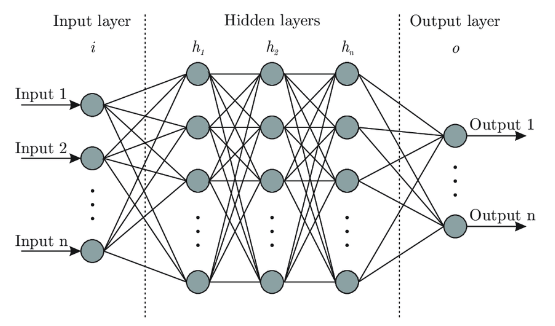
**SVM**



- maszyna wektorów nośnych

- nauka ma na celu wyznaczenie hiperpłaszczyzny rozdzielającej z maksymalnym marginesem przykłady należące do dwóch klas

**Sieci neuronowe**

****

**Klasteryzacja** - Grupowanie, analiza skupień, klasteryzacja (ang. data clustering) - metoda tzw. klasyfikacji bez nadzoru (ang. unsupervised learning) dokonująca grupowania elementów we względnie jednorodne klasy. Podstawą grupowania w większości algorytmów jest podobieństwo pomiędzy elementami – wyrażone przy pomocy funkcji (metryki) podobieństwa.

Metody klasteryzacji:

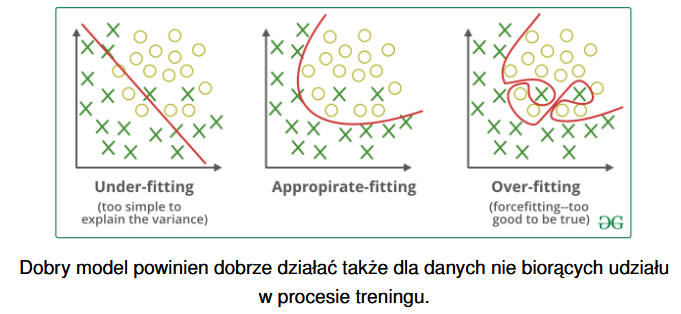
- metody hierarchiczne

- grupa metod k-średnich

- samouczące się mapy

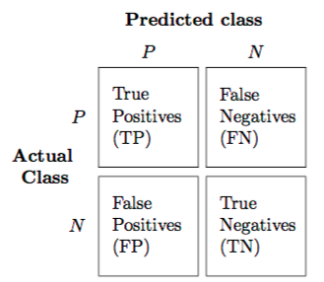
- inne

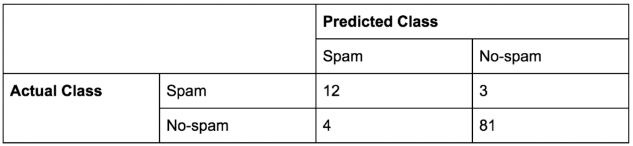
**Nadmierne dopasowanie do danych**

****

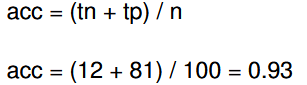
**Miary oceny jakości modelu:**

- macierz pomyłek

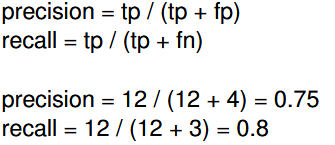




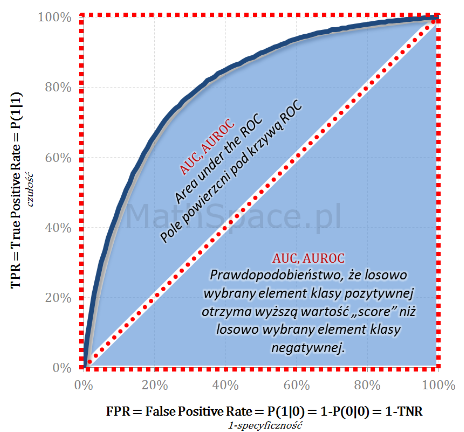
- trafność



- precision, recall



- roc, roc-auc

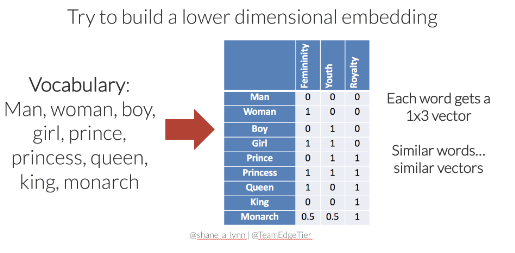


**Word embedding –** osadzanie słów - Wspólna nazwa zbioru technik modelowania języka i uczenia się cech w przetwarzaniu języka naturalnego, w których słowa lub wyrażenia są mapowane na wektory liczb rzeczywistych. Matematyczne osadzanie z przestrzeni cech o wysokim wymiarze, w której reprezentowane ą słowa, do ciągłej przestrzeni wektorowej o znacznie mniejszym wymiarze.

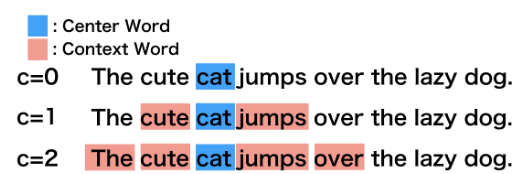
- jest jednym z najbardziej popularnych sposobów przedstawiania słownictwa z dokumentów tekstowych

- jest w stanie uchwycić kontekst słowa w dokumencie, podobieństwo semantyczne i syntaktyczne, relację z innymi słowami itp.

- jest wektorową reprezentacją słów z dokumentów tekstowych



**Word2Vec** - algorytm pozwalający na wygenerowanie osadzeń słów na podstawie korpusu tekstowego. Do generowania wektorów słów stosuje jedną z dwóch metod: Skip Gram lub Common Bag Of Words (CBOW), gdzie w obu metodach wykorzystywane są sieci neuronowe



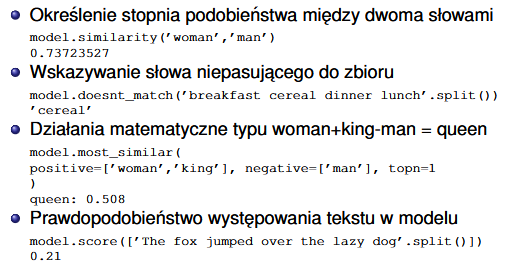
Parametry:

- training algorithm - hierarchical softmax and/or negative sampling

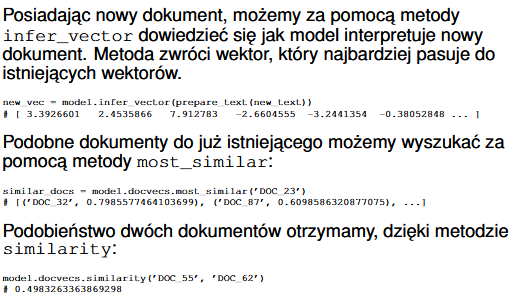
- Sub-sampling - podpróbkowanie często występujących słów

- Dimensionality - liczba neuronów w warstiwe ukrytej, a tym samym rozmiar wektorów reprezentujących słowa, zazwyczaj wartość z przedziału 100-1000

- Context window - rozmiar kontekstu, zalecane 10 dla Skip gram i 5 dla CBOW



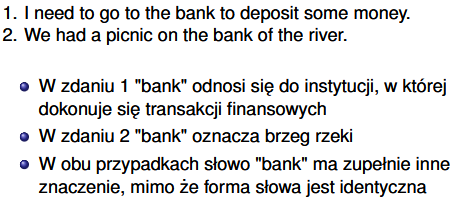
**Doc2Vec -** metoda reprezentacji wektorowej całych dokumentów (zdań, paragrafów lub całych książek), a nie pojedynczych słów. Zasada działania jest podobna jak w przypadku word2vec gdzie dodano dodatkowy unikalny (wśród wszystkich dokumentów) wektor, reprezentujący dokument



**GloVe -** Global Vectors for Word Representation - przeformułowanie optymalizacji word2vec jako specjalnego rodzaju faktoryzacji macierzy współwystępowania słów. Algorytmy word2vec i GloVe odróżnia sposób wyznaczania wektorów słów. Zasadniczą częścią algorytmu GloVe jest redukcja wymiarowości macierzy liczników współwystępowania słów w korpusie tekstowym

**fastText** - jest swego rodzaju rozszerzeniem Word2Vec. Traktuje ona każde słowo jako składające się z n-gramów znakowych. Wektor dla słowa składa się z sumy tworzących go n-gramów znakowych

**ELMo** (Embeddings from Language Models) - Generuje kontekstowe osadzenia, które dostosowują się do konkretnego kontekstu, w jakim dane słowo jest używane.



- Kontekstualne osadzenia (contextual embeddings)

- Wielowarstwowa architektura (deep contextualized representations)

- Dwa strumienie informacji (bidirectional LSTM)

- Dostosowanie wag na poziomie zadań (task-specific weighted combinations)

- Efektywne wykorzystanie danych przedtrenowanych (pretraining)

BERT (Bidirectional Encoder Representations from Transformers)

- Dwukierunkowość (bidirectionality) przetwarza dane w obu kierunkach jednocześnie – zarówno od lewej do prawej, jak i od prawej do lewej, analizuje wszystkie słowa w kontekście całego zdania

- Oparty na architekturze Transformer umożliwia efektywne modelowanie zależności między odległymi słowami w tekście, mechanizm self-attention

- Pretraining i fine-tuning - najpierw przechodzi przez etap pretrainingu na ogromnych zbiorach danych (np. Wikipedia, BookCorpus), a następnie może być dostosowany do specyficznych zadań poprzez fine-tuning

- Maskowanie słów (Masked Language Modeling, MLM) - w trakcie pretrainingu stosuje technikę masked language modeling, polega to na tym, że model losowo maskuje część słów w zdaniu i próbuje przewidzieć brakujące słowa na podstawie kontekstu

- Zadanie przewidywania następnego zdania (Next Sentence Prediction, NSP) model musi zdecydować, czy dwa zdania pojawiają się w logicznej sekwencji

Uniwersalność - skuteczny w wielu różnych zadaniach NLP, uznawany za model uniwersalny w NLP

Potencjalne pytania:

Co to jest kontekst w technice osadzania słów

Co to jest odległość edycyjna. Wyliczyć dla dwóch słów

Analiza gramatyczna.

Różnica między lematyzacją i stemingiem