Podstawowe pojęcia- trochę stare metody ale potem przejdziemy do nowszych rzeczy.

Z NLP wiąże się test Turniga którego głównym celem miało być określenie kiedy komputery uzyskają inteligencje, a była ona utożsamiania z posługiwaniem się językiem naturalnym jak ludzie.

Podstawowa wersja to wybierana jest sędzia, informacje będą podawane przez kogoś jedną osobą jest człowiek a druga maszyna, sędzia nie wie kto jest kim i rozmowa może potoczyć się w każdym kierunku. Na koniec sędzia wskazuje maszynę.

Metody lingwistyczne:

Lingwistyka obliczeniowa bada tworzy i utrzymuje modele które pomagają technologii lepiej przetwarzać ludzki język. Zajmują się w szczególności analizą wyszukiwarek.

Modelowane statystyczne w NLP to proces przewidywania następnego słowa w sekwencji na podstawie słów poprzedzających.

NLP to pojęcie najogólniejsze dotyczy rozumienia i przetwarzania tekstu

Przykłady użycia w pdf

Podstawowe pojęcia:  
Podstawą w NLP jest przetwarzanie tekstu nawet mowę musimy przekształcić na tekst ale to już powoli się dezaktualizuje bo już są modele które działają na mowie ale są narzędzie do przetwarzania mowy na tekst.

Podstawowym obiektem będzie litera albo słowo

Popularne podejście to analiza słów, nie można wprowadzić do modelu czegoś czego nie uwzględnia słownik.

Słownik to zbiór wyrazów ułożonych i opracowanych według jakiejś zasady ale nas bardziej dotyczy definicja że jest to indywidualny zasób wyrazów.

Przystępując do NLP musimy ustalić na początku słownik który będziemy używać i będą tam wszystkie dostępne słowa ale w niektóre słowa wychodzą z użytku albo już nie są używane i trzeba. Generalnie słowo na temat podejścia analizy słów a analizy znaków w sumie to ładnie opisane w pdf

W przypadku słownika złożonego z pełnych słów każde z nich będzie osobnymi tokenami.

Na początku mamy zestaw dokumentów-> zbiór wszystkich dokumentów nazywamy korpusem. Dokumentami mogą być inne treści nie tylko język naturalny.

Pre-procesin-> wczytujemy dokumenty, dzielimy je na zdania potem na wyrazy i będą one naszymi tokenami- tokenizacja. Tych tokenów będzie bardzo dużo. Tokenizacja to podział na mniejsze jednostki definicja w pdf.

Przykład

Mamy zadanie „Data mining is the process of discovering patterns in large data sets”

Dla tego tekstu tokenami będą :

{data, mining, is, the, proces, of, discovering, patterns, in, large, data, sets}

Drugie wystąpienie słowa data jest pomijane bo już zostało wcześniej ujęte w słowniku.

Wydaje się że są to słowa podzielone względem spacji ale w rzeczywistości może być trudniej. Jeżeli będą znaki interpunkcyjne to mogą przykleić się do słów zrobił i zrobił? to dwa oddzielne tokeny. Historycznie po utworzeniu słownika następowała jego redukcja dzisiaj używa się tekstu jaki on jest naprawdę bo nie musimy przejmować się złożeniowością obliczeniową.

Aby odkleić te słowo stosuje się najczęściej wyrażenia regularne a następnie jak mamy już oddzielone przeprowadza się dalszą analizę- normalizacja słownika np. rzutowanie znaków na małe litery, ale tracimy informacje bo tracimy nazwę własną czy nazwiska obecnie jest to kwestia balansu miedzy wielkością słownika a mocą obliczeniową. Przeprowadza się bardziej złożoną analizę i tworzy się osobne tokeny dla ważnych słów.

Już się tego nie robi ale można sprowadzić słowo to formy podstawowej i aby to wykonać są dwa podejścia lematyzacja i stemming

Stemming-> ingeruje bardziej w słownik, pozwala na większą redukcje i obcina słowa do uzyskania rdzenia bazując na regułach może to skutkować utworzeniem rdzenia nie występującego w języku naturalnym.

Przykład Grać Grający Gram Grasz Gra

Obetniemy do formy gra w języku angielskim możemy ucinać ed ing s to co określimy w cechach danego języka.

Chodzić chodzę chodzisz chodzi chodzą chodźmy

W dobrym przypadku będzie zrzutowane do słowa chodz albo do słowa chod i ten rdzeń nie występuje w języku.

Lematyzacja-> podejście zależy od słownika grupa słów zostanie zniesiona do jednego tokenu ale zawsze występujące ze słownika

Iść idę idziemy idąć -> zostanie zniesione do słowa iść

Dobry lepszy najlpeszy -> dobry więc lematyzacja potrafi wychwycić kontekst i zależności

Dzisiaj raczej słownika się nie ogranicza ale wykonuje się w inny sposób tokenizację nowoczesne sposoby.

Usuwamy słowa które występują małą ilość razy

Stop words-> słowa występujące często w tekście nie noszące za sobą żadnego znaczenia. Każdy język zawiera swoje stop wordsy ale sami musimy je ustalić wiadomo że są gotowe biblioteki ale nie zawsze są wystarczające.

Znaki interpunkcyjne raczej teraz są uwzględniane

Tokenizacja to dzielenie zdań, używanie wyrażeni regularnych, rzutowanie terminów do tokenów jeżeli trzeba , stopwords, znaki interpunkcyjne.

W taki sposób uzyskaliśmy słownik. Teraz będziemy chcieli określić zadanie i musimy zamienić słowo na liczbę. Na razie zostajemy przy najprostszym One-hot encoding -> każde słowo, każdy token dostaje wektor który ma długość wielkości słownika i jedną 1 w nim