**Intel® AI DLA MŁODZIEŻY 2019**

PROGRAM PILOTAŻOWY

PODRĘCZNIK MODERATORA

**Moduł 15 - Pozyskiwanie [Koncepcje typowe dla poszczególnych domen] (Przetwarzanie języków naturalnych)**

CAŁKOWITY CZAS TRWANIA SESJI: 480 MINUT

LICZBA INSTRUKTORÓW: 2

LICZBA UCZESTNIKÓW: 40

**UWAGA:** ŻADNA OSOBA SPOZA ORGANIZACJI INTEL CORPORATION NIE MOŻE KOPIOWAĆ, ROZPROWADZAĆ, REPRODUKOWAĆ ANI W ŻADEN INNY SPOSÓB UDOSTĘPNIAĆ INFORMACJI ZAWARTYCH W NINIEJSZYM DOKUMENCIE OSOBOM, INSTYTUCJOM I/LUB ORGANIZACJOM NIEPOWIĄZANYM Z NINIEJSZYM PROGRAMEM BEZ ZGODY WYŻEJ WYMIENIONEJ ORGANIZACJI.

NIEPRZESTRZEGANIE POWYŻSZYCH ZASAD BĘDZIE SKUTKOWAĆ ZAKOŃCZENIEM WSPÓŁPRACY PRZEZ WYŻEJ WYMIENIONĄ ORGANIZACJĘ.

NINIEJSZY DOKUMENT MUSI **ZOSTAĆ ZWRÓCONY** UPOWAŻNIONEMU PERSONELOWI FIRMY INTEL CORPORATION PO ZAKOŃCZENIU SESJI.

**Podręcznik moderatora**

|  |  |
| --- | --- |
| **Temat lekcji:** Moduł 15 - Pozyskiwanie [Koncepcje typowe dla poszczególnych domen] (Przetwarzanie języków naturalnych) | **Tryb:** Sesje interaktywne prowadzona przez instruktora |
| **Podsumowanie:** Ten moduł prezentuje podstawy teoretyczne kluczowych koncepcji i technik w domenie NLP (przetwarzanie języków naturalnych) przed ich szczegółowym omówieniem na etapie Doświadczenie. Omówione koncepcje/techniki obejmują różnicę między językiem naturalnym a formalnym, przetwarzanie danych w NLP, worek słów, tf-idf i miarę kosinusową. | |
| **Cele edukacyjne:**   1. Zrozumienie różnicy między językiem naturalnym a formalnym 2. Przegląd procesu AI w domenie NLP 3. Rozwój intuicyjnego rozumienia typowych technik AI w domenie NLP 4. Prezentacja potencjalnych zastosowań rozwiązań NLP wykorzystujących AI | |
| **Rezultaty edukacyjne:**   1. Opis różnicy między językiem naturalnym a formalnym 2. Opis etapów w przetwarzaniu danych NLP 3. Wyjaśnienie algorytmów i technik używanych w NLP, np. worek słów, tf-idf 4. Wskazanie aktualnych zastosowań NLP | |
| **Wymagania wstępne:**   1. Brak | |
| **Kluczowe pojęcia:**   1. Język naturalny a język formalny 2. Tokenizacja 3. Worek słów (bag of words) 4. tf-idf 5. Miara kosinusowa | |
| **Kluczowe umiejętności:**   1. Umiejętności typowe dla domeny NLP: Zrozumienie koncepcji i pojęć typowych dla domeny | |
| **Wykorzystywane materiały:**   1. [Slajdy] Moduł 15 - Pozyskiwanie [Koncepcje typowe dla poszczególnych domen] (Przetwarzanie języków naturalnych) 2. [Arkusz] Moduł 15 - Przetwarzanie tekstu, worek słów i tf-idf    1. Łącze: <https://docs.google.com/document/d/1SoSVPA8JLQ-kFEXKd9a5uvc5PlZ7u-aVQxC3utufB0Y/> | |
| **Kwestie etyczne dotyczące AI:**   1. Kodeks postępowania dotyczący AI | |
| **Zastosowanie w scenariuszach z prawdziwego życia [np. kwestie społeczne]:**   1. Zastosowanie NLP, np. chatboty, analiza sentymentu | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nie** | **Czas** | **Ćwiczenie** | **Opis** | **Cel** |
| 1. Wprowadzenie do przetwarzania języków naturalnych (NLP) | 30 minut | - Wykład: Wprowadzenie do NLP | - Wprowadzenie do przetwarzania języków naturalnych  - Zastosowanie NLP  - Proces AI w zastosowaniach NLP  - Uwaga: Użyj szablonu do ustalania zakresu problemu w odniesieniu do przykładu zastosowania w sektorze medycznym.   1. Ramy problemu 2. Pozyskiwanie danych 3. Eksploracja danych 4. Modelowanie 5. Ocena 6. Wdrożenie | Wprowadzenie do technologii przetwarzania języków naturalnych i jej zastosowań |
| 30 minut | - Zadanie: Testowanie chatbotów | - Jakie są funkcje chatbotów?  - Analiza różnych chatbotów |
| 20 minut | - Wykład: Język człowieka a język komputera | - Wprowadzenie do cech języka naturalnego, które stanowią wyzwanie dla przetwarzania maszynowego |
| 2. Podstawowe pojęcia NLP: Przetwarzanie danych | 60 minut | - Wykład: Przetwarzanie danych dla NLP  - Zadanie: Przetwarzanie danych dla NLP | - Uczestnicy zapoznają się wstępnym przetwarzaniem danych tekstowych | Zrozumienie podstawowych pojęć przetwarzania języków naturalnych |
| 3. Znajomość algorytmów uczenia maszynowego (ML): Worek słów (bag of words) | 40 minut | - Wykład: Worek słów  - Zadanie: Worek słów | - Tworzenie słownika modelu i worka słów na podstawie zbioru danych. | Intuicyjne zrozumienie algorytmu worka słów |
| 4. Zrozumienie algorytmów uczenia maszynowego (ML): częstość terminu - odwrotna częstość w dokumentach | 40 minut | - Wykład: tf-idf  - Zadanie: tf-idf | - Przegląd procesu tf-idf | Intuicyjne zrozumienie procesu częstość terminu - odwrotna częstość w dokumentach |
| 5. Wprowadzenie do biblioteki NLTK | 90 minut | Notatnik Jupyter | - Wprowadzenie do biblioteki NLTK | Zapoznanie uczestników z biblioteką NLTK i jej podstawowymi funkcjami |
| 6. Wizualizacja wektorów słów | 30 minut | - Wykład: Wizualizacja wektorów słów | - Redukcja wymiaru w celu wizualizacji wektorów | Wprowadzenie do wizualizacji wektorów |
| 7. Znajomość algorytmów uczenia maszynowego (ML): miara kosinusowa | 30 minut | - Wykład: Miara kosinusowa | - Podobieństwo 2 wektorów wskazuje na podobieństwo znaczenia | Intuicyjne zrozumienie miary kosinusowej |
|  |  |  |  |  |
| 8. Znajomość algorytmów uczenia maszynowego (ML): Modelowanie dla NLP | 40 minut | - Czynność: Proces trenowania modelu | - Uczestnicy użyją aplikacji internetowej do zapoznania się z metodą trenowania modelu NLP. | Zrozumienie jednego ze sposobów trenowania modelu |
| 8. Etyka a sztuczna inteligencja | 30 minut | - Dyskusja o etyce w sztucznej inteligencji | - Uczestnicy wypowiedzą się na temat osobistego asystenta zbudowanego w oparciu o sztuczną inteligencję, który działa podobnie do człowieka. | Zachęcenie uczestników do krytycznego myślenia na temat problemów związanych z zastosowaniem przetwarzania języków naturalnych. |
| 40 minut | - Opracowanie kodeksu etycznego | - Uczestnicy wspólnie tworzą kodeks etyczny dotyczący zastosowania NLP w dziedzinie sztucznej inteligencji  - od 5 do 10 punktów oraz wyjaśnienie, dlaczego etyka jest konieczna |

**3. Przewodnik po czynnościach**

***[Slajdy 1–29] Wprowadzenie do przetwarzania języków naturalnych [80 min]***

**Celem tej części jest zaprezentowanie uczestnikom przetwarzania języków naturalnych i jego potencjalnych zastosowań.**

***[Slajd 2]***

Przetwarzanie języków naturalnych jest dziedziną badań obejmującą interakcje między językiem komputerów a językami naturalnymi, a w szczególności sposoby programowania komputerów do celów przetwarzania i analizowania dużych ilości danych językowych.

Można gromadzić i przetwarzać bardzo wiele danych i wniosków!

***[Slajd 3]***

Na pewno znacie już niektóre zastosowania technologii przetwarzania języków naturalnych!

Gdzie zauważyliście przykłady zastosowania, w których komputery sprawiają wrażenie, że rozumieją ludzką mowę?

***[Slajd 4]***

* Tłumaczenie
* Wyniki wyszukiwania
* Siri/Alexa
* Chatboty
* Podpisy zdjęć
* Podpisy filmów
* Analiza sentymentu <https://tone-analyzer-demo.ng.bluemix.net/>
* Klasyfikacja tekstu
* Streszczenie

Przetwarzanie słów (autokorekta)

***[Slajd 5]***

Jak wygląda proces AI w przypadku przetwarzania języków naturalnych? Zobaczmy przykład.

***[Slajd 6]***

Załóżmy, że w trakcie waszego projektu zaobserwujecie wzrost stresu i będziecie chcieli stworzyć rozwiązanie terapeutyczne do obniżenia poziomu stresu.

Odkryliście, że terapia poznawczo-behawioralna (TPB) jest jedną z najlepszych metod przeciwdziałania stresowi, ponieważ łatwo mogą ją stosować osoby, które nie są ekspertami.

Chcecie więc stworzyć chatbot, który działa jak terapeuta TPB. Jaki będzie wasz pierwszy krok?

***[Slajd 7]***

Najpierw musicie dostarczyć kontekst dla chatbota. Na jakie zdania napotyka często terapeuta TPB? Jakie oczekiwania ma pacjent korzystający z terapii? Jakie są sposoby, na które można coś powiedzieć? Dane pozyskacie z Internetu, z wywiadów, ankiet i obserwacji sesji terapeutycznych TPB.

***[Slajd 8]***

Gdy będziecie mieli już wystarczająco dużo danych, musicie sprawdzić, czy nadają się do waszego algorytmu uczenia maszynowego. Dlatego przetworzycie dane tekstowe na format, który może dalej przetwarzać i analizować komputer.

***[Slajd 9]***

W zależności od typu chatbota i poziomu złożoności, jaki chatbot ma obsługiwać, możecie wybrać spośród różnych dostępnych modeli budowania chatbota.

***[Slajd 10]***

Po wybraniu i wytrenowaniu modelu ocenicie go i upewnicie się, że jest na tyle dokładny, by go zastosować do waszych celów.

***[Slajd 11]***

Ocena jest gotowa! Macie ukończone rozwiązanie i teraz musicie je przetestować i wdroży!

***[Slajd 12]***

Brawo! Zobaczyliśmy, jak wygląda proces AI w przypadku NLP. Dla przypomnienia pokazuję wam ten proces raz jeszcze.

***[Slajd 13] Zadanie: Testowanie chatbotów***

Istnieje już wiele gotowych chatbotów i niektóre z nich używają procesu, który wykorzystaliśmy do utworzenia chatbota. Wypróbujmy chatboty, które zostały już wdrożone.

Zastanawialiście się, po nam są chatboty? Mają sprawiać wrażenie, że rozmawiamy z prawdziwą osobą, ponieważ wtedy czujemy się komfortowo. Mamy tutaj kilka chatbotów. Podzielę was teraz na grupy Teraz wypróbujcie chatboty, porozmawiajcie z nimi, a później spróbujemy podsumować wasze doświadczenia.

***[Slajd 14]***

Okej, czas minął! Mam nadzieję, że wszyscy skończyli. Porozmawiajmy o naszych doświadczeniach. Zadaj następujące pytania:

- Które chatboty wypróbowaliście?

- Jaki jest cel chatbota?

- Jak wyglądała interakcja z chatbotem?

- Czy czuliście, że rozmawiacie z człowiekiem, czy z robotem? Dlaczego tak sądzicie?

- Czy uważacie, że chatbot ma osobowość?

Gdy jesteśmy w stanie sprawić, że komputery rozumieją nasz język, możemy również zaprojektować interakcje z maszyną w taki sposób, żeby przebiegały płynnie i bez zakłóceń.

Czy to oznacza, że chatbot, który wykazuje cechy robota, jest gorszy? Wcale nie. Jeśli tylko chatbot realizuje swój zamierzony cel, to możemy powiedzieć, że jest udany. Pamiętajcie o tym, przygotowując wasze projekty. Poznamy kilka metod tworzenia rozwiązań NLP. Będą one użyteczne w zależności od okoliczności, a wy musicie zdecydować, które metody wykorzystacie do swojego rozwiązania.

***[Slajd 15]***

Czy zauważyliście, że niektóre chatboty są prostsze od innych? Niektóre dają zasadniczo dają możliwość wyboru między opcjami, a inne pozwalają wpisywać tekst i rozpoznają to, co napisaliście.

- Które z nich to tradycyjne chatboty? Które wykorzystują sztuczną inteligencję?

- Jak działa chatbot tradycyjny i chatbot AI?

- Który jest oparty na regułach, a który na trenowaniu?

***[Slajd 16]***

Oto kilka różnic między chatbotami tradycyjnymi a inteligentnymi botami zazwyczaj zasilanymi sztuczną inteligencją.

***[Slajd 17]***

Krótko powiedzieliśmy, czym jest NLP i jakie ma ciekawe zastosowania. Teraz przejdziemy do szczegółów. Jakie są podstawowe pojęcia związane ze skutecznym przetwarzaniem języków naturalnych? Wchodzimy na terytorium, gdzie język spotyka maszynę. Porównajmy język ludzi i język komputerów!

***[Slajd 18] Wykład: Język człowieka a język komputera***

Cel przetwarzania języków naturalnych to nauczyć maszyny ludzkiej mowy.

***[Slajd 19]***

Dziś zajmiemy się słowami. Oto podstawowe reguły, którymi będziemy się kierować.

***[Slajd 20]***

Ludzie porozumiewają się za pomocą języka. Nieustannie przetwarzamy język. Nawet teraz, gdy do was mówię, wasze umysły przetwarzają moje słowa i rozumieją je.

Fale dźwiękowe z moich ust trafiają do bębenka w uchu i są przekształcane na impuls nerwowy przenoszony do mózgu, który go przetwarza. W ten sposób możecie rozumieć, co mówię. Jeśli macie wątpliwości, to zadajecie mi pytania.

Ludzie przetwarzają języki naturalne. W ten sposób się ze sobą komunikujemy.

***[Slajd 21]***

Jak rozmawiamy/komunikujemy się z komputerami? Zazwyczaj używamy kodów!

Musicie bardzo uważać na to, co wpisujecie do komputera. W przeciwnym razie komputer zarejestruje błąd. Zobaczcie, jak komputer potrafi interpretować informacje!

Obejrzymy teraz film, który nam to przedstawi! Łącze: <https://www.youtube.com/watch?v=xnyFYiK2rSY>

***[Slajd 22]***

Widzicie? Język komputera bardzo się różni od języka ludzi! NLP ma za zadanie nauczyć komputer naszego języka. Czy myślicie, że komputerowi łatwo przychodzi ta nauka?

***[Slajd 23]***

Nasz język rządzi się swoimi prawami i regułami. Są w nim rzeczowniki, czasowniki, przysłówki i przymiotniki. Słowa mogą być raz rzeczownikami, a innym razem przymiotnikami. Są też reguły, które określają strukturę języka.

Mówimy tutaj o **składni** języka. Składnia to gramatyczna struktura zdania. Gdy struktura jest obecna, możemy zacząć interpretować komunikat. Chcemy, żeby komputer postępował tak samo. Jednym ze sposobów jest oznakowanie części mowy. Dzięki temu komputer może zidentyfikować poszczególne części mowy.

Ale poza strukturą jest przecież znaczenie. Komunikacja międzyludzka to złożony proces. Jest wiele elementów ludzkiego języka, które nam jest łatwo zrozumieć, a komputer ma z nimi ogromne trudności.

Analogia z językiem programowania:

inna składnia, ta sama semantyka: 2+3 = 3+2

ta sama składnia, inna semantyka (1 i 1,5) 3 / 2 (Python 2.7) =/= 3/2 (Python 3)

***[Slajd 24]***

Wasza kolej: Czy znacie przykłady zdań z różną składnią (układem elementów), ale taką samą semantyką (znaczeniem)?

***[Slajd 25]***

Spójrzmy na te trzy zdania:

- Zrobił się czerwony na twarzy, gdy odkrył, że wziął niewłaściwą torbę

Co to oznacza? Czy było mu wstyd, bo wziął torbę kogoś innego zamiast swojej? Czy był zły, bo nie udało mu się ukraść tej torby, na którą czyhał?

- Czerwony samochód mignął mu przed nosem.

Chodzi tu prawdopodobnie o kolor samochodu

- Zrobił się czerwony na twarzy po zażyciu lekarstwa

Czy to reakcja alergiczna? Czy było mu wstyd, bo przegrał zakład („Nie mam zamiaru się przez to rozchorować”)? Czy zażył lekarstwo, które rozszerza tętnice?

Widzicie? Kontekst jest istotny. Rozumiemy zdanie niemal intuicyjnie w zależności od naszej historii używania języka i wspomnień, które nosimy w pamięci.

***[Slajd 26]***

Wasza kolej: Czy znacie inne słowa o wielu znaczeniach?

***[Slajd 27] - slajd obrazujący scenę***

Zdanie może mieć poprawną składnię, a jednocześnie nic nie oznaczać. Zobaczcie: Kurczaki płyną ekstrawagancko, gdy księżyc pije herbatę.

Co to oznacza?

***[Slajd 28]***

Mamy teraz świadomość wyzwań, jakie przed nami staną, gdy będziemy próbowali nauczyć komputery naszego języka. Na tych warsztatach poznamy techniki, które pozwolą nam tego dokonać!

Po pierwsze, dowiemy się, jak przetwarzać dane tekstowe na formę, którą komputery rozumieją: Liczby. Po przekształceniu danych tekstowych na liczby jesteśmy w domu! Możemy użyć technik uczenia maszynowego, aby przetwarzać dane, znajdywać podobieństwa, grupować teksty, przeprowadzać analizę sentymentu itp.

***[Slajd 29]***

Czego się nauczyliśmy do tej pory?

- Czym jest NLP i jakie ma zastosowania

- Jak działają chatboty

- Analiza chatbotów

- Różnica między językiem naturalnym a formalnym

- Składnia i semantyka w NLP

***[Slajdy 30–41] Podstawowe pojęcia NLP: Przetwarzanie danych [60 min]***

**Celem tej części jest prezentacja podstawowych technik przetwarzania danych w ramach przetwarzania języków naturalnych.**

***[Slajd 31]***

Załóżmy, że mamy nasze dane tekstowe. Chcemy, żeby komputer je przeanalizował i coś nam o nich powiedział, np. słowa kluczowe, ton artykułu, w jakim stopniu ten artykuł jest podobny do innego. Od czego zacząć? Ludzie i komputery mówią bardzo różnymi językami.

Co się dzieje, gdy spotykacie kogoś z innego kraju i nie możecie się porozumieć? Tak! Potrzebujecie tłumacza. Ktoś musi przetłumaczyć to, co mówicie, na język, który rozumie wasz rozmówca.

Jaki język rozumie komputer? Język liczb! Dlatego będziemy tłumaczyć słowa na liczby!

***[Slajd 32]***

Jak wygląda krok po kroku proces przekształcania tekstu w liczby? Na tym dziś się skupimy. Przejdziemy przez metodę przetwarzania tekstu krok po kroku.

***[Slajd 33]***

W tej części zapoznacie się z procesem przetwarzania tekstu, korzystając z arkuszy, które wam rozdamy.

***[Slajd 34]***

Najpierw dzielimy tekst na zdania. Zakładamy, że każde zdanie zawiera istotną ideę. Możemy też założyć, że każde zdanie kończy się kropką. Prosty algorytm segmentacji użyje tej informacji do podziału tekstu na zdania na podstawie kropek. Nie zawsze jednak wykonujemy ten krok. Ronimy to tylko wtedy, gdy skupiamy się na tekście zdanie po zdaniu. To znaczy, chcemy sklasyfikować pojedyncze zdania. Albo jeśli chcemy wykonać segmentację zdania, czyli podzielić zdanie na jego gramatyczne części, aby (na przykład) zidentyfikować wszystkie rzeczowniki.

Zobaczcie, jak to wygląda z użyciem tego narzędzia: <http://textanalysisonline.com/spacy-sentence-segmentation>

[Weź przykładowe zdanie z <https://woebot.io/what-is-cbt/> lub <https://woebot.io/blog/why-we-need-mental-health-chatbots/> ]

***[Slajd 35]***

Po podziale tekstu na zdania możemy zająć się każdym zdaniem z osobna. Podzielimy zdanie na poszczególne słowa (tokeny). Spacje wyznaczają miejsca podziału. Znaki interpunkcyjne też są tokenami, ponieważ niosą ze sobą znaczenie. Zobaczcie, jak wygląda internetowe narzędzie do tokenizacji: <https://text-processing.com/demo/tokenize/>

***[Slajd 36]***

Następnie usuniemy stop-listy, znaki specjalne i cyfry.

Stop-listy to słowa, które są pomijalne, ponieważ nie mają żadnego znaczenia lub to znaczenie jest mało istotne. Są to słowa, które sprawiają, że zdanie jest gramatycznie poprawne, ale nie dodają one zbyt wiele do jego znaczenia. Takie słowa są w każdym tekście. Możecie podać ich przykłady?

[„jest”, „ten”, „i”]

Znaki specjalne. To jest bardziej oczywista sprawa. Jakie znacie znaki specjalne?

[@#%^]

Jeśli klasyfikujecie tekst z witryny internetowej, znaki specjalne takie jak znaczniki mogą znajdować się w tekście. Należy je usunąć. Na przykład: <p>, <h1> itp.

Usunięcie cyfr pozwoli komputerowi skupić się na słowach! Ale w zależności od konkretnych celów można zachować cyfry!

***[Slajd 37]***

Cześć, CZeść, CZEŚĆ i czEŚĆ

Co oznacza tekst na ekranie?

Wszyscy wiemy, że oznacza zwyczajne powitanie – „cześć” – i każde z tych słów mówi dokładnie to samo.

Komputer nie jest w stanie rozpoznać, że Cześć, CZeść, CZEŚĆ i czEŚĆ są jednym i tym samym słowem, bo nie wyglądają tak samo. Konwertujemy wszystkie słowa na pisane małą literą, aby komputer zrozumiał, że to jest to samo słowo.

Tekst może mieć wiele form, między którymi dostrzegamy różnice, ale komputer już tego nie potrafi.

Tó jěst těkst z akcěntami -> To jest tekst z akcentami

***[Slajd 38]***

Stemming (sprowadzanie do rdzenia). Zanim powiemy, czym jest stemming, musimy ustalić, czym jest rdzeń (ang. „stem”). Rdzeń to „początek albo przyczyna”. Rdzeń słowa to ta część słowa, która zostaje, gdy usuniemy przyrostki i przedrostki. Stemming to proces sprowadzenia słowa do jego rdzenia.

Na przykład, „uzdrowiciel”, „uzdrowiony” i „uzdrowienie” sprowadzają się do rdzenia „zdrowie”. Ale nie zawsze jest tak łatwo. Na przykład, „student” sprowadza się do „studiów”, a „studiowanie” do „studiować”.

***[Slajd 39]***

Lematyzacja jest podobna do stemmingu, bo też polega na odcięciu przyrostków i przedrostków słowa. Jednak forma bazowa w tym przypadku nosi nazwę „lemat”, a nie „rdzeń”.

Różnica jest taka, że lemat musi być słowem, które występuje w słowniku. Na przykład, rdzeniem słowa „studia” może być „studi”, ale jego lematem będzie „studium”.

Jak myślicie? Czy lematyzacja nie jest bardziej zaawansowana od stemmingu? To prawda, trudniej jest zbudować narzędzie do lematyzacji niż do stemmingu. Narzędzie do lematyzacji wymaga wykorzystania znacznie głębszej wiedzy lingwistycznej. Kompromis polega na tym, że lematyzacja zabiera znacznie więcej czasu niż stemming.

Opracowano już wiele narzędzi do stemmingu i lematyzacji. Sprawdźcie tu <http://textanalysisonline.com/nltk-porter-stemmer> i tu <http://textanalysisonline.com/spacy-word-lemmatize>, aby zobaczyć je w działaniu! Jak widzicie, wynik końcowy nie zawsze się zgadza z tym, czego się nauczyliśmy, bo inżynierzy i naukowcy często stosują różne metody projektowania.

***[Slajd 40]***

Oto kolejny przykład porównujący stemming z lematyzacją.

***[Slajd 41]***

Podsumujmy to, czego się do tej pory dowiedzieliśmy. Zaczęliśmy od danych tekstowych. Było one w formie zrozumiałej dla ludzi, ale nie dla komputerów. Przeszliśmy krok po kroku przez proces przekształcenia danych na tokeny, prostszą formę zbliżoną do naszego celu.

Czy to już koniec? Nie!

Gdy już mamy tokeny, czyli najprostsze formy słów obecnych w korpusie tekstowym, czas je zamienić na liczby. W tym celu użyjemy algorytmu zwanego workiem słów.

***[Slajdy 42–52] Podstawowe pojęcia NLP: Worek słów [40 min]***

**Celem tej części jest prezentacja algorytmu worka słów i ćwiczenia praktyczne z nim związane dla lepszego zrozumienia.**

***[Slajd 43]***

Co teraz? Sprowadziliśmy nasz tekst do formy tokenów. Bliżej nam do tego, co mogą zrozumieć komputery (liczby), ale to jeszcze nie koniec. Na tej sesji poznamy model worka słów, aby nauczyć się zamienić teksty na liczby.

***[Slajd 44]***

Komputer z łatwością odczytuje tabele liczbowe. Mają one ustalone rozmiary i zawierają wyłącznie liczby. Tekst w języku naturalnym jest trudny do zrozumienia dla komputera, bo każdy dokument może mieć inną długość i nie składa się wyłącznie z liczb.

Z tego powodu chcemy dokonać konwersji tekstu na liczby.

***[Slajd 45]***

Czym jest worek słów? Worek słów liczy, ile słów występuje w każdym dokumencie, i wyraża to w postaci wektora (rzędu liczb). W tej formie komputer może się uczyć tych liczb i generować predykcje!

Worek słów obejmuje dwa elementy:

1. Słownik słów
2. Liczbę słów

Nazywamy go workiem słów, bo kolejność słów w zdaniu nie jest istotna.

***[Slajd 46]***

Jak wygląda krok po kroku proces przekształcania tekstu w liczby?

1. Pozyskajcie dane i przetwórzcie je zgodnie z opisem z poprzedniej sesji
2. Utwórzcie słownik: Zróbcie listę wszystkich unikatowych słów zawartych w danych
3. Utwórzcie wektory dokumentu: Dla jednego dokumentu: sprawdźcie listę unikatowych słów i policzcie, ile razy każde słowo występuje w dokumencie
4. Powtórzcie tę czynność dla wszystkich dokumentów

***[Slajd 47]***

W tym momencie wasz tekst jest już przetworzony ze zdań na lematy. W tym przykładzie nie będziemy jednak wykonywać przetwarzania, aby łatwiej dostrzec proces tworzenia worka słów.

***[Slajd 48]***

Następnie utworzycie słownik. Z waszych tekstów wyciągnięcie wszystkie unikatowe słowa i tokeny.

***[Slajd 49]***

Następnie utworzycie wektory dokumentu.

Zapiszcie wasz słownik w jednym wierszu. Przejdźcie przez dokument. Dla każdego słowa w dokumencie zapiszcie 1 nad tym słowem w słowniku. Jeśli natraficie na słowo drugi raz, zmieńcie 1 na 2. To samo powtórzcie w następnym wierszu dla kolejnego dokumentu.

***[Slajd 50]***

Ilekroć natraficie na słowo, którego nie ma na liście, dodajcie je do listy. Po zakończeniu wasz słownik będzie gotowy. Kolejność słów nie ma w tym momencie znaczenia.

***[Slajd 51]***

Powtórzcie tę czynność dla wszystkich dokumentów.

***[Slajd 52]***

Podsumujmy to, czego się do tej pory dowiedzieliśmy. Przekształciliśmy tokeny na liczby. Znajdują się w naszym słowniku wraz z informacją o częstości występowania.

***[Slajdy 53–67] Podstawowe pojęcia NLP tf-idf [40 min]***

**Celem tej części jest zaprezentowanie uczestnikom koncepcji tf-idf.**

***[Slajd 54]***

Załóżmy, że macie przed sobą książkę. Jakie słowa/znaki będą w niej najczęściej występować? Czy słowa częściej występujące są przez to ważniejsze?

***[Slajd 55]***

Utworzyliśmy już worek słów, który liczy, jak często każde słowo występuje w dokumencie. Worek słowa zakłada, że im częściej dane słowo występuje w dokumencie, tym większe ma znaczenie.

Ale nie zawsze tak jest. Czy znacie słowa, które często występują w wielu dokumentach, ale nie mają (istotnego) znaczenia? Właśnie, słowa takie jak „jest”, „również”, „i” mogą się wielokrotnie pojawiać w dokumencie, ale nie mówią nam zbyt wiele. Można również na tej samej zasadzie usunąć słowa pomijane (stop-listy). Z drugiej strony, jeśli wiemy, że dokumencie występują słowa „komputer” i „gry”, to już nam coś mówi o treści tego dokumentu.

Uwaga: Najczęstsze z tych słów są usuwane jako słowa pomijalne (stop-listy) podczas przetwarzania wstępnego. Jednak można to zastosować również do konkretnej dziedziny. Na przykład, kiedy analizujemy zbiór tweetów i trafiamy na słowo „film”, może nam ono pomóc w ustaleniu tematyki. Kiedy jednak analizujemy zbiór recenzji filmowych, słowo „film” pojawi się wielokrotnie w każdej z nich, lecz niewiele nam to powie.

***[Slajd 56]***

Czym jest tf-idf? Jest to angielski skrót oznaczający „częstość terminu - odwrotna częstość w dokumentach”. Ta nazwa mówi wszystko. Co się tutaj odbywa? Porównujemy częstość określonego terminu w jednym dokumencie z częstością w całym korpusie tekstów. Zasada jest następująca: coś, co występuje wszędzie, może nie być aż tak ważne jak coś, co występuje tylko w kilku dokumentach.

***[Slajd 57]***

Przejdźmy przez proces analizy tf-idf!

Po pierwsze, tf oznacza całkowitą częstość (total frequency). Jest to liczba wystąpień każdego słowa w przeliczeniu na dokument. Robiliśmy to już wcześniej – częstość terminu w worku słów.

***[Slajd 58]***

idf oznacza odwrotną częstość w dokumentach (inverse document frequency) i oblicza się ją logarytmicznie (liczba dokumentów/liczba dokumentów zawierających słowo).

***[Slajdy 59–65]***

Aby obliczyć częstość w dokumentach, zacznijcie od pierwszego słowa w waszym słowniku.

1. Przejdźcie przez wszystkie dokumenty i policzcie, w ilu dokumentach to słowo występuje przynajmniej jeden raz.
2. Zapiszcie tę liczbę obok słowa w waszym słowniku.
3. Powtórzcie tę czynność dla każdego słowa z waszego słownika. Powstanie sekwencja liczb, po jednej dla każdego słowa w słowniku. To jest częstość w dokumentach.
4. A teraz odwrotna częstość w dokumentach.
5. Pomnóżcie każdy wiersz w waszym worku słów według elementów przez logarytm odwrotnej częstości w dokumentach. W ten sposób uzyskacie wartość tf-idf.

***[Slajd 65]***

Spójrzcie na wartość! Słowo,które występuje częściej, ma mniejszą wartość tf-idf. Aby słowo miało wysoką wartość tf-idf w dokumencie, musi występować wiele razy w jednym dokumencie i być nieobecne w innych dokumentach. Jest to słowo identyfikujące dokument.

***[Slajd 66]***

Jakie są zastosowania tf-idf?

* Uzyskanie istotnych wektorów cech do dalszego przetwarzania (pamiętacie znaczenie „cechy”?)
* Pozyskiwanie informacji, np. wyszukiwanie
* Filtrowanie słów pomijalnych (stop-listy)

***[Slajd 67]***

Podsumowanie: czego się dowiedzieliśmy o tf-idf?

***[Slajdy 68–79] Podstawowe pojęcia NLP: Wektory słów [30 min]***

**Celem tej części jest prezentacją wektorów słów i sposobów ich wizualizacji dla łatwiejszego zrozumienia.**

***[Slajd 69]***

Ostatnio utworzyliśmy worek słów. Teraz mamy tablicę pełną liczb. Choć są to wektory, to nie możemy ich zobaczyć, bo wektory mają więcej wymiarów niż 3. Jednak na ostatniej sesji widzieliśmy miarę kosunisową z przedstawieniem wektorów. Jak to możliwe?

Łatwiej jest zrozumieć coś, co możemy zobaczyć. Naukowcy i inżynierowie opracowali techniki redukcji wymiaru, aby stworzyć wizualizację tego, co ma więcej wymiary niż 3. Nie będziemy się zagłębiać w te algorytmy i techniki, ale zasadnicza idea jest następująca: bierzemy coś, czego nie da się wizualnie przedstawić w 3 wymiarach, i redukujemy wymiar do 3D, aby umożliwić wizualizację.

***[Slajd 70]***

Wiemy, że wektory mają kierunek i wielkość. Te wykresy przedstawiają wektory w 2D i 3D. Można je zilustrować na rysunku i możemy je zrozumieć, bo żyjemy w przestrzeni trójwymiarowej.

Ale jak stworzyć wizualizację wektorów wykraczając poza 3D?

***[Slajd 71]***

Wyobraźmy sobie, że żyjemy w innym wymiarze. Jak by to było?

Zauważcie, jak trudno wyobrazić sobie przestrzeń poza trzema wymiarami.

Chcemy więc zredukować złożoność do postaci, którą możemy zrozumieć.

Dlatego tworzymy wizualizację wektorów 4D, 5D itd. poprzez redukcję ich wymiarowości.

Co to oznacza?

***[Slajd 72]***

To jest przykład, który pomoże nam to zrozumieć. Załóżmy, że chcemy przedstawić obiekt trójwymiarowy (np. piłkę lekarską) na kartce papieru. Co zrobimy? Narysujemy okrąg. Aby stworzyć efekt trzeciego wymiaru, dodamy cienie. Nie będzie to obiekt 3D. Utracimy niektóre informacje na temat piłki. Jakie? Być może piłka ma z tyłu inny kolor? Może to nie piłka, a jedynie połowa piłki? Rysunek w 2D nie obejmie całego przedmiotu 3D, ale zrobiliśmy, co w naszej mocy, aby wiernie przedstawić piłkę.

***[Slajd 73]***

To jest graficzne przedstawienie redukcji wymiaru. Czy ktoś może objaśnić nam ten wykres?

***[Slajd 74]***

Ta aplikacja internetowa (<https://ronxin.github.io/wevi/>) jet przykładem redukcji wymiaru wektorów słów. Wykonajcie poniższe czynności:

- Spójrzcie na wektory i zauważcie, jak są położone względem siebie

- Ponownie wytrenujcie model i sprawdźcie, jak wektory są mapowane

W ten sposób możemy zobaczyć wektory słów!

Widzieliście, jak słowa mogą być przedstawiane za pomocą wektorów. Co możemy z nimi zrobić? Możemy zastosować metody matematyczne do analizy i wyciągnięcia wniosków na ich podstawie!

***[Slajd 75]***

To jest narzędzie do wizualizacji word2vec, typowe narzędzie używające sieci neuronowych w celu konwersji słów na wektory.

***[Slajd 76]***

Spójrzcie na kontrolkę.

***[Slajd 77]***

Wczytajcie zestaw „Król i Królowa”, wytrenujcie 1000 powtórzeń (kliknijcie przycisk [500], poczekajcie do końca, kliknijcie jeszcze raz). Następnie kliknijcie przycisk [PCA]. Spójrzcie na wykres punktowy w dole po prawej, sprawdźcie niebieskie znaczniki (wektory wejściowe) i zauważcie analogię: „król - królowa = mężczyzna - kobieta”.

***[Slajd 78]***

Czy potraficie znaleźć inne relacje?

***[Slajd 79]***

Do zapamiętania:

Zmieniliśmy słowa na liczby w formie wektorów dokumentu.

Te wektory mają zazwyczaj więcej niż 3 wymiary, nie możemy ich przedstawić wizualnie.

Przeprowadzamy redukcję wymiaru, aby łatwiej nam było zrozumieć wektor i pokazać go wizualnie.

Nie wszystkie relacje między słowami można łatwo zrozumieć, bo tracimy dane podczas redukcji wymiaru.

***[Slajdy 80–87] Podstawowe pojęcia NLP: Miara kosinusowa [30 min]***

**Celem tej części jest zaprezentowanie uczestnikom koncepcji miary kosinusowej.**

***[Slajd 81]***

Poprzednio używaliśmy zestawu NLTK to analizy korpusu. Wyszukiwaliśmy słowa, drukowaliśmy je w kontekście/z sąsiednimi słowami i porównywaliśmy, jak słowa są używane w różnych dokumentach.

Spójrzcie na ekran. Mamy tu pomieszane słowa. Jak byście je pogrupowali?

Jak to robiliśmy?

***[Slajd 82]***

Jak mierzymy, czy idee są ze sobą powiązane? [poczekaj na odpowiedzi uczniów]

Sprawdzamy, czy występują razem. Sprawdzamy, czy pojawiają się obok siebie w podobnym kontekście. Przez lata zgromadziliśmy tyle kontekstu, że możemy utworzyć 2 grupy idei.

W wielu zastosowaniach NLP porównujemy zdania i sprawdzamy, w jakim stopniu są podobne do siebie. Na przykład, chatbot może wziąć pytanie i przeszukać swoją bazę wiedzy, znajdując informację, która jest najbardziej podobna do pytania, i podać tę informację jako odpowiedź.

Jak to podobieństwo jest mierzone? Za pomocą miary kosinusowej.

***[Slajd 83]***

Pamiętacie, że komputery używają liczb do przetwarzania danych.

Na początku utworzyliśmy wektor dokumentu, który jest numerycznym przedstawieniem naszego dokumentu.

Wcześniej wspomnieliśmy, że dokumenty można przedstawiać jako wektory dzięki redukcji wymiaru.

Jak zatem odkryć podobieństwo między 2 wektorami?

***[Slajd 84]***

Miara kosinusowa mierzy różnicę między dwoma wektorami. Im bliżej siebie są dwa wektory, tym większa wartość miary kosinusowej!

***[Slajd 85]***

Miara kosinusowa mierzy różnicę między dwoma wektorami. W tabeli TFIDF każdy wiersz to wektor i możemy obliczyć odległość kosinusową między każdym wierszem, aby zmierzyć ich podobieństwo. Użyjemy tej metody do zbudowania prostego chatbota na etapie Doświadczenie!

***[Slajd 86]***

Teraz wy spróbujcie!

***[Slajd 87]***

Powtórka z kluczowych pojęć

* Ludzie porównują idee, używając pamięci i kontekstu.
* Komputer porównuje idee za pomocą liczb.
* Zmieniliśmy słowa na liczby w formie wektorów dokumentu.
* Używamy odległości między wektorami dokumentu, aby zmierzyć podobieństwo między ideami.
* Miara kosinusowa mierzy tę odległość.
* Bliższe wektory = większa miara kosinusowa = większe prawdopodobieństwo, że idee są ze sobą powiązane

***[Slajdy 88–93] Trenowanie modelu w NLP [40 min]***

**Celem tej części jest zaprezentowanie uczestnikom podstawowej koncepcji trenowania modelu w NLP.**

***[Slajd 89]***

Mówiliśmy o trenowaniu modelu przy okazji danych i komputerowego rozpoznawania obrazów. Czy ktoś nam przypomni, jak się przeprowadza takie trenowanie?

[Zwróć uwagę na odpowiedzi wskazujące „dane do trenowania”, „dane wejściowe i wyjściowe”, „redukcję błędów”]

***[Slajd 90]***

Brawo! Możemy też na wiele sposobów uczyć komputery naszego języka. Wypróbujemy teraz prostą aplikację internetową do trenowania NLP. Najpierw jej użyjemy, a później przeanalizujemy jej działanie. Oto łącze: <http://shrdlurn.sidaw.xyz/acl16/>

[Uczestnicy będą trenować prostą maszynę, która ma za zadanie układać bloki w różnych kolorach. Następnie komputer nauczy się wykonywać nowe polecenia na podstawie wpisywanych poleceń i informacji zwrotnych użytkownika.

Uczestnicy mogą ze sobą rywalizować o to, kto zajdzie najdalej w wyznaczonym czasie.]

Instrukcja:

Gra zawiera dwie kombinacje bloków: początkową kombinację i końcową kombinację. Gra polega na tym, aby wprowadzać polecenia na klawiaturze (np. „usuń pomarańczowe bloki”) i przejść od kombinacji początkowej do końcowej.

Na początku model będzie losowo przesuwał i usuwał bloki. Użyjcie strzałek „góra” i dół”, aby przejść przez cykl posunięć modelu, aż natrafi on na prawidłowe posunięcie. Wtedy naciśnijcie przycisk zaznaczenia, aby model wiedział, że dobrze trafił. Wyświetli się wtedy nowy zestaw kombinacji początkowej i końcowej.

Poziom trudności stopniowo się zwiększa, więc dojście do końcowej kombinacji może wymagać wielu poleceń. Nauczy to uczestników dzielić złożone zadania na proste jednostki, w których mogą wytrenować model.

***[Slajd 91]***

Zatrzymujemy się! Zobaczmy, kto dotarł do najwyższego poziomu w wyznaczonym czasie!

* Jaką macie technikę?
* W jaki sposób komputer uczy się waszych poleceń?

***[Slajd 92]***

Brawo! I jeszcze jedno! Komputer nie jest ograniczony do jednego języka naturalnego. Nauczy się wszystkiego, co mu zapodacie. Możecie powiedzieć mu, że prawa strona jest z lewej i lewa strona jest u góry, a komputer posłusznie nauczy się tego waszego „języka”.

Wiedząc o tym, jaką technikę można wykorzystać?

***[Slajd 93]***

To jest jeden sposób trenowania maszyny. Podczas tego kursu używamy kilku technik służących do trenowania modeli. Więcej dowiemy się o nich na etapie Doświadczenie!

Na zakończenie quiz: <https://create.kahoot.it/share/duplicate-of-natural-language-processing/6c142a0d-d1cc-408b-8496-6c54594db4a1>

***[Slajdy 94–98] Dyskusja na temat etyki [70 min]***

**Celem tej części jest zachęcenie uczestników do krytycznego myślenia na temat problemów związanych z zastosowaniem przetwarzania języków naturalnych.**

***[Slajd 94]***

Dowiedzieliśmy się, jak stosować NLP do tworzenia zaawansowanych rozwiązań (rozpoznawanie głosu, klasyfikacja, chatbot). Czy to jest w porządku? Jakie potencjalne problemy wynikają z użycia tej technologii?

***[Slajd 95]***

Obejrzymy teraz film (<https://www.youtube.com/watch?v=JvbHu_bVa_g>)

***[Slajd 96]***

Czy to nie ciekawe? Technologia umożliwia komputerom rozumienie ludzi i komunikowanie się z nimi. Techniki, jakie poznaliście na tych warsztatach, są podstawowymi elementami składowymi takich technologii. Pamiętacie woebota? Chatbota medycznego? Wyobraźcie sobie, że któregoś dnia nie będziecie już nawet wiedzieli, czy rozmawiacie z robotem! Jak będzie wyglądać życie, gdy interakcje między ludźmi będą nieodróżnialne od interakcji między ludźmi a maszynami?

Zastanówmy się nad tym pytaniem:

1. W jaki sposób te naśladujące ludzi systemy sztucznej inteligencji są narażone na nadużycia?
2. Jak twórcy takich systemów powinni postępować na etapie projektowania?
3. O czym powinni pamiętać? Czy ludzie będą bezgranicznie ufać tym robotom?

***[Slajd 97]***

Załóżmy, że jesteście w organizacji regulujące zastosowanie sztucznej inteligencji, w tym również zastosowanie osobistego asystenta, którego przed chwilą widzieliśmy. Jakie zasady należy wdrożyć i dlaczego? Utwórzcie własne zasady etyczne dotyczące projektowania robotów

[Pokieruj uczestnikami w taki sposób, by opracowali kodeks etyczny dotyczący pracy z AI. Około 4 lub 5 punktów podsumowujących odpowiedzialne projektowanie systemów AI. Poproś uczestników, by zaprezentowali swój kodeks etyczny. To dobry moment, aby zwrócić uwagę na wątpliwości i priorytety uczestników. Nie krytykuj ambiwalencji. Cały czas uczymy się, jak traktować coraz płynniejszą interakcję między ludźmi a maszynami.]

***[Slajd 98]***

W porządku! Czas się skończył. Czy ktoś chce się z nami podzielić swoimi pomysłami? Jakie zapisaliście zasady i dlaczego waszym zdaniem są one ważne.

[Zakończ sesję]

**4. Wskazówki dotyczące rozwiązywania problemów**

***Nie dot.***

**5. Rubryka oceny**

|  |  |  |  |
| --- | --- | --- | --- |
| Umiejętności | **Słabe** | **Średnie** | **Dobre** |
| Opis zastosowania NLP w codziennym życiu | Nie potrafią nazwać żadnych zastosowań NLP. | Potrafią nazwać zastosowania wymienione na wykładzie. | Potrafią nazwać kilka zastosowań i dać przykłady sposobu interakcji z tymi systemami. |
| Wiedzą, jak komputery rozumieją znaczenia, wykorzystując wektorowe przedstawienie słów. | Nie są w stanie wykazać żadnej intuicji dotyczącej reprezentacji właściwościowej słów. | Potrafią wyjaśnić, że słowa mogą być przedstawiane jako wektory (reprezentacje właściwościowe słów). | Potrafią wyjaśnić, dlaczego słowa znajdujące się blisko siebie w przestrzeni wektorowej mają często podobne znaczenie. |
| Potrafią sformułować kodeks etyczny w zakresie pracy ze sztuczną inteligencją, szczególnie w odniesieniu do przetwarzania języków naturalnych. | Nie potrafią uzasadnić potrzeby etyki w dziedzinie AI. | Potrafią opisać skutki nieetycznego postępowania w dziedzinie AI. | Potrafią uzasadnić potrzebę etyki i wskazać ich własne zasady dotyczące AI, a także opisać skutki nieetycznego postępowania w dziedzinie AI. |

**6. Polecane lektury**

Proste wprowadzenie do przetwarzania języków naturalnych <https://towardsdatascience.com/introduction-to-nlp-5bff2b2a7170>

Czym jest stemming i lematyzacja? <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

<https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>

Wykład na temat przetwarzania języków naturalnych

<https://www.youtube.com/watch?v=mhHfnhh-pB4>

Miara kosinusowa

<https://www.machinelearningplus.com/nlp/cosine-similarity/>

Etyka a sztuczna inteligencja <https://www.technologyreview.com/s/612318/establishing-an-ai-code-of-ethics-will-be-harder-than-people-think/>