Modele językowe
Pracownia 1
Zajęcia 1 i 2
Uwaga: 2 zadania można bez straty punktowej przenieść na pracownię 2

Na skosie znajdzie się link do materiałów potrzebnych do rozwiązywania zadań z tej listy. W każdym zadaniu powinieneś korzystać z modelu językowego dla języka polskiego, ze strony huggingface. Sugerowane są dwa stosunkowo niewielkie modele: PapuGaPT oraz polka, ale nie musisz się do nich ograniczać. Na stronie wykładu będą przykładowe programy, które możesz wykorzystywać w tym zadaniu.

Zadanie 1. (**4p**) Wykorzystując wybrany model jęzkowy stwórz czatbota (masz pełną dowolność odnośnie tego, o czym ten czatbot ma rozmawiać, nie musi wykonywać żadnej pożytecznej funkcji). Twój czatbot powinien:

a) Generować niezbyt długie, w miarę sensowne odpowiedzi.

b) Jakoś zarządzać historią dialogu, umieszczając ją (częściowo?) w promptcie.

c) Generować więcej niż jedną odpowiedź, wybierać „optymalną" odpowiedź wg jakiegoś kryterium, które sam wybierzesz.

Pewnie do tego zadania będziemy wracać, nie przejmuj się zatem, jeżeli wynik nie będzie na razie w pełni satysfakcjonujący (niektóre modele trenowane po prostu na tekstach trudno zmusić promptem do dialogu).

Zadanie 2. Rozważmy trzy zdania:

> Babuleńka miała dwa rogate koziołki.
> Wiewiórki w parku zaczepiają przechodniów.
> Wczoraj wieczorem spotkałem pewną wspaniałą kobietę, która z pasją opowiadała o modelach językowych.

Zwróć uwagę, że w języku polskim, dla każdego z tych zdań bardzo wiele z wszystkich możliwych permutacji wyrazów to zdania mniej lub bardziej poprawne, choć czasem mocno nacechowane stylistycznie, jak przykładowo:

> Dwa miała rogate koziołki babuleńka. Przechodniów w parku zaczepiają wiewiórki.

Twoim zadaniem jest napisanie programu, który dla danego (multi)zbioru wyrazów wypisuje kilka przykładowych uszeregowań tych wyrazów, w kolejności od najbardziej naturalnego. Przyjmujemy dodatkowo, że kropka jest zawsze na końcu, a jedyne słowo pisane wielką literą to pierwsze słowo zdania. Przy rozwiązywaniu zadania **jedynym** źródłem wiedzy o języku polskim powinien być model językowy, który da się uruchomić na dostępnym Ci komputerze (PapuGaPT, polka, ...).
Zadanie ma dwa, osobno punktowane warianty:

a) Liczba słów jest na tyle mała, że można przejrzeć wszystkie permutacje. (**3p**)

b) Liczba słów jest na tyle mała, że można przejrzeć ich wszystkie pary, znaleźć słowa, które do siebie wyraźnie pasują, utworzyć prawdopodobne zbitki wyrazów i tylko je permutować.(**4p**)

Przetestuj program(y) na przykładowych zdaniach i na kilku własnych.

Zadanie 3. (**4p**) Wykorzystując model językowy i funkcje oceniania prawdopodobieństwa tekstu napisz program, który znajduje wydźwięk opinii (czy jest ona pozytywna, czy negatywna). Przykładowe opinie pozytywne:

> Parking monitorowany w cenie.
> Hotel czysty, pokoje były sprzątane bardzo dokładnie.
> Generalnie mogę go polecić, kierował mnie na potrzebne badania, analizował ich wyniki,

cierpliwie odpowiadał na pytania.
Fajny klimat pofabrykanckich kamienic.
Sala zabaw dla dzieci, plac zabaw na zewnątrz, kominek, tenis stołowy.

Przykładowe opinie negatywne:

W wielu pokojach niedziałająca klimatyzacja.
Jedzenie mimo rzekomych dni europejskich monotonne.
Drożej niż u konkurencji w podobnym standardzie.
Może za szybko zrezygnowałam, ale szkoda mi było wydawać pieniędzy na spotkania, które nie przynosiły efektu.
Omijaj to miejsce!

W tym zadaniu **nie** masz stosować techniki few-shots learning, jedynie operować na prawdopodobieństwach pewnych napisów. Sprawdź więcej niż jeden sposób konstrukcji tych napisów. Podaj skuteczność Twojego programu (jak często znajdujesz poprawne odpowiedzi) dla danych ze strony wykładu (wystarczy reprezentatywna próbka losowa).

Zadanie 4. (**6p**) Napisz program odpowiadający na proste pytania o fakty. Pytanie są różnorodne, ale łatwo zauważysz w nich pewne grupy (pomaga posortowanie). Są one wzięte z konkursu https://2021.poleval.pl/, ale nie powinieneś tam (na razie) zaglądać. Podstawową techniką będzie tu zero-shot/one-shot/few-shots learning i model polka, ale dodatkowo powinieneś:

1. Przynajmniej jedną grupę pytań obsłużyć jakąś heurystyką
2. Przynajmniej w jednej grupie pytań skorzystać z modelu językowego działającego w trybie obliczania prawdopodobieństw.

Nie wolno korzystać z żadnej formy wyszukiwania informacji, jedynym źródłem wiedzy ma być model językowy i pytania w części uczącej zbioru danych. Podaj (szacowany) procent poprawnych odpowiedzi. We wspomnianym konkursie rozwiązanie bazowe zapewnione przez organizatorów miało kilka procent skuteczności – nie powinno być to trudne do pobicia.