Створення тренувальних датасетів для корекційних великих мовних моделей

Нестеренко Володимир

Кафедра радіофізики та комп'ютерних технологій

Львівський національний університет імені Івана Франка

Львів, Україна

Вступ

- Вивчення іноземної мови
- Чат-бот
- Великі мовні моделі

Постановка завдання

- Fine-tuning
- Генерація тренувального датасету
- Дослідження способів генерації

Способи генерації помилок

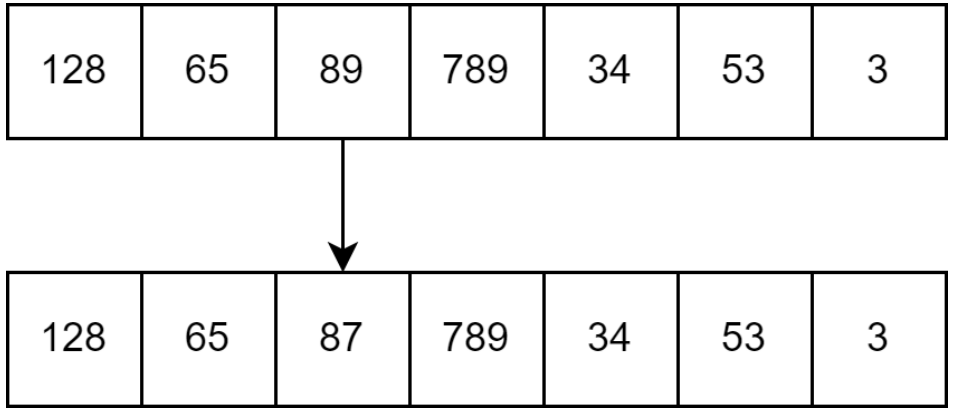
- Лематизація
- Зміна векторного представлення
- Зміна порядку
- Заміна
- Видалення

Лематизація

- Інструменти: spaCy
- Приклади:
  - strahlender → strahlend
  - befindet → befinden
  - Personen → Person
  - ganzen → ganz

Зміна векторного представлення

- Інструмент: huggingface.AutoTokenizer
- Приклади:
  - Ja, wir haben eine große Terrasse mit direktem Blick auf den See.
  - Ja, wir haben eine große terrkon mit direktem blick auf den see.



Зміна порядку

- Інструменти: spaCy
- Приклади:
  - Michael reserviert einen Tisch zum Abendessen.
  - Reserviert Michael einen Tisch zum Abendessen.

  - So, möchten Sie drinnen oder draußen sitzen?
  - So, Sie möchten drinnen oder draußen sitzen?

Заміна

- Інструмент: словники замін
- Приклади:
  - Oh, Sie haben auch eine Terrasse?
  - Oh, euch haben auch eine Terrasse?

  - Um welche Uhrzeit möchten Sie bei uns sein?
  - Um welche Uhrzeit möchten Sie über uns sein?

Видалення

- Інструменти: модуль random
- Приклади:
  - 18 Uhr wäre gut.
  - 18 Uhr gut.

  - Da heute strahlender Sonnenschein ist, würden wir gerne draußen sitzen.
  - Da heute strahlender Sonnenschein ist, würden gerne draußen sitzen.

Результати

Спотворення	Схожість
Лематизація	0,994
Спотворення ВП	0,842
Перестановка	0,99
Заміна	0,977
Видалення	0,945