Speech-to-Text with Named Entity Recognition (STT + NER)

(Ushbu hujjatda asosan loyiha maqsadi, qiyinchiliklar va kelajakdagi rejalar yoritilgan Loyihaning texnik kodlari, jarayonlari va natijalari bilan readme file orqali tanishishingiz mumkin)

Hujjat qismlari:

Loyiha Haqida

Loyiha O'zbek tilida ovozni matnga aylantiruvchi (STT) va matndagi nomlangan entitetlarni aniqlovchi (NER) tizimini yaratish uchun amalga oshirildi. STT modeli audio faylni matnga o'zgartiradi, so'ngra NER modeli bu matndagi shaxs, joy, tashkilot kabi nomlangan obyektlarni ajratib beradi. Loyihada maxsus o'rgatish jarayonlari va O'zbek tiliga moslashtirilgan modellar yaratildi.

Loyihaning Asosiy Qismlari:

1. **Speech-to-Text**: O'zbek tilidagi ovozli ma'lumotlarni matn shaklida olish.
2. **Named Entity Recognition**: Matndan joylar, odamlar, tashkilotlar kabi nomlangan obyektlarni aniqlash.
3. **Pipeline**: Ikki modelni birlashtirish va natijada audiodan entitelarni olish

Ilk Qadam:

Loyihani boshlashdan avval:

- Bu bo’yicha avval yaratilgan loyihalar, ularning ishlash sifati va ochiq ma’lumotlar to’liq o’rganib chiqildi.
  (uzbekvoice.ai, aisha.group)

- Kerakli modelni va datasetni tanlash bo’yicha ko’plab maqolalar o’rganib chiqildi
- Umumiy reja tuzib olindi. Foydalanish uchun resurslar va datasetlar ko’zdan kechirildi.

Ushbu jarayon 4-5 kun vaqt oldi

Model, Muammo va Yechim:

STT Modeli:

Speech to text modelini yaratishda fine tuning uchun **Whisper-base** modeli, Dataset uchun **Common Voice 17.0** dataseti tanlandi

Nega whisper-base va Common Voice 17.0?

Model:

- O’zbek tili uchun ham o’qitilgan.
- Nisbatan kichik model va ishlashga oson.
- Model va u bilan ishlash bo’yicha manba va qo’llanmalar ko’p.

Dataset:

- O’zbek tilida va ancha katta bo’lgan dataset
- 17.0 uning so’ngi taqdim etilgan to’plami hisoblanadi

Yuzaga kelgan qiyinchiliklar:

Loyiha davomida uchragan asosiy qiyinchilik bu resurslarning (GPU, RAM, DISK) ning yetishmasligi bo’ldi. Birinchi urunishda diskdagi xotira yetishmasligi tufayli jarayonlar bekor qilindi

Qo’llangan Yechim:

O’qitilishi reja qilingan dataset ikki qismga bo’lib yuborildi. Va ikki marotaba o’qitishga qaror qilindi. Birinchi olingan model katta xatoliklar bilan ishladi va bizga kerakli natijani bermadi.

Ikkinchi bosqichda, to’liq tayyor bo’lmagan model ikkinchi marotaba (datasetning ikkinchi qismi uchun) o’qitildi. Olingan modelning natijasi qoniqarli darajada

Natijada bizda ikkita model paydo bo’ldi.

1. Xatoliklar bilan ishlovchi, ikkinchi modelni train qilish uchun yaratilgan - **whisper-uz**
2. To’liq tayyor bo’lmagan model asosida qayta o’qitilgan - **whisper-uz-v2**

NER Modeli:

Named Entity Recognition modelini o'zbek tili uchun moslash nisbatan qiyin va ko'p vaqt oldi.
Fine tuning uchun **Roberta-base,** Dataset uchun **uzbek_ner** dataseti tanlandi

Nega Roberta-base va uzbek_ner?

Model:

- O'zbek tilida NER uchun o'qitilgan sanoqli modellardan biri .
- Nisbatan kichik model va ishlashga oson.

Dataset:

- Katta dataset (19k qator)
- Dataset JSON formatida va foydalanishga qulay
- Keraksiz ustunlar mavjud emas, yuklash va qayta ishlash ko'p vaqt olmaydi

Yuzaga kelgan qiyinchiliklar:

Asosiy muammo dataset va modelning tokenizerida kuzatildi, Datasetda xatoliklar mavjudligi va undagi qiymatlarda og'ish model sifatiga salbiiy tasir qildi.

Qo'llangan Yechim:

Base modelning tokenizeridan voz kechildi va 130.000 qatordan iborat dataset yordamida **Custom Tokenizer** yaratildi: jamshidahmadov/uz_tokenizer

Base Tokenizer va Custom Tokenizer o'rtasidagi farqni notebook orqali ko'rishingiz mumkin

Olgan modelimiz katta aniqlik bermadi ammo asosiy modelni va Tokenizer yaratib oldik,
kelajakda yangi va sifatli datasetlarni topish orqali pre-train jarayonini o'tkazish yaxshi natija beradi.

Yakuniy Natijalar

- **STT Modeli:** Word Error Rate (wer) ~ 30**.**
- **NER Modeli:** Precision ~ 97%

Kelajakdagi Rejalar

Venchur fond yoki startup tanlovlarda ishtirok etish

STT model uchun:

- Modelni shevalar uchun ham tayyorlash
- Katta resurslar bilan katta datasetlarni o’qitish

NER model uchun:

- Model uchun sifatli dataset topish yoki **sun’iy yaratish,** uni model uchun o’qitish:
    - GPT api
    - Translator
    - Data Augmentation

Contacts: Jamshid Ahmadov

Linkedin - linkedin.com/in/jamshid-ds

Telegram - @jamshidds

Gmail - ahmadovv54@gmail.com