## Lépések:

Létrehoztam a github repot, majd a virtuális környezetet a projektnek venv-el, majd a poetry segítségével fogom kezelni a függőségeket, hogy meglegyenek egy helyen.

A megfelelő függőségeket leszedtem, illetve az ASR rendszernek a az openAI Whisperét választottem, amelyből van több is, de én a tinyt használom, a gyorsaság miatt, nem lesz probléma ha apróbb hibák lesznek a leiratozásban. Építettem egy gradio web appot, amelybe képesek vagyunk feltölteni hanganyagot, vagy a mikrofonon keresztül beszélni, és ezeket az előbb említett whisper modell leiratozza és elmentjük az eredményt a további lépésekhez.

Fontos lépés a megkapott leiratozás eredményének a vektorizációja, amelyet egy embedding modell végez el. Ezt először az Ollamán elérhező nomic embeddert használtam, ami azonban minden embedeléshez egy HTTP kérést küldene és emiatt nagyon lelassulna a teljes folyamat. Ezek miatt átálltam a Huggingfacen elérhető offline beágyazó modellt a sentence transformersen belül a "all-MiniLM-L6-v2"-t, amelynél használtam a batchelésnek a lehetőségét, hogy egyszerre 32 chunkot képes embeddelni így ha nagy dokumentumokat kapunk akkor sem lassul le nagyon a rendszer.

Evaluation:

Elkezdtem foglalkozni a TTS modul kiértékelésével itt több érdekes dologgal találkoztam vannak a szubjektív értékelések ilyen a MOS értékelés (Mean opinion score), ahol emberek pontozzák egy 5-s skálán, hogy mennyire természetes az adott hang (ezt megcsinálom a szobában lévőkkel), és vannak az objektív pontozók, ahol pedig valamilyen algoritmus szerint vizsgáljuk a szintetizált hangot. Ilyen metrika a WER (word error rate), amelynek a kiértékelési folyamat a következő:

* írtam egy hosszabb és komplex témáról szóló szöveget (max 250 karakter hosszú az XTTS miatt), majd ezt megkapta a két darab TTS modulom (pytts, xtts), amelyek előállítottak ebből egy hangfájlt a felolvasott szöveggel.
* Ezt a hangfájlt odaadtam az ASR rendszeremnek (whisper small), amely előállított egy leiratot és a leirat és az eredeti szövegen mértem ezt a WER metrikát
* A WER metrika azt vizsgálja, hogy a kapott szöveg, mennyire tér el bármilyen módon is a referencia szöveegtől, szóval ha rossz hanganyagot állított volna elő bármelyik TTS modul, akkor a leiratozó hibásan leiratozna és magas lenne ez az error rate.

A math equations and formulas

AI-generated content may be incorrect.

https://learnius.com/slp/9+Speech+Synthesis/1+Fundamental+Concepts/3+Evaluation/word+error+rate+(WER)

Másik objektív metrika, amit számolni fogok, az a MCD és a STOI.