Eines de codi obert per a la creació de corpus

Pol Pastells

29 d'abril de 2025

1 Introducció

Aquest treball descriu un repositori de codi obert dissenyat per facilitar la creació de corpus lingüístics a partir de fonts web. El projecte conté eines per a l'scraping de dades textuals i audiovisuals, i està pensat per ser utilitzat tant en recerca com en desenvolupaments aplicats, com ara sistemes de reconeixement automàtic de la parla (ASR). El codi del projecte es troba disponible al repositori: https://github.com/Pastells/corpus_scraping/.

2 Estructura del repositori

El repositori es divideix en dues seccions principals:

- **Twitter**: Inclou un quadern amb instruccions i tres projectes d'exemple per extreure contingut textual de la plataforma.
- **audio_subs**: Conté scripts per generar corpus a partir de parells àudio–transcripció extrets de vídeos amb subtítols.

3 Creació de corpus a partir de Twitter

Per extreure dades de Twitter fem servir la biblioteca twscrape, una API no oficial que permet fer cerques avançades i gestionar múltiples comptes. Aquesta eina resulta especialment útil per descarregar grans volums de tuits, ja que permet alternar entre diversos comptes per evitar límits de consulta.

3.1 Funcionament general

El procés comença amb la configuració dels comptes a la base de dades interna de `twscrape`. Un cop configurats, es poden llançar cerques mitjançant l'ordre `api.search()`, que accepta consultes com ara hashtags, mencions, paraules clau o noms d'usuari. També és possible restringir les cerques per dates.

3.2 Exemples de projectes

El repositori conté tres exemples pràctics:

- **RAE:** Recollida de tuits amb `#dudaRAE` en diferents franges temporals.
- **La Revuelta:** Recollida mitjançant mencions com `#LaRevuelta` o `@LaRevueltaTVE`.
- **Users:** Recollida massiva de tuits a partir d'una llista d'usuaris.

Els resultats es poden guardar en fitxers `pickle` o transformar en `pandas.DataFrame` per facilitar-ne l'anàlisi. Finalment, s'exporten a CSV o format Excel per al seu ús posterior.

4 Corpus d'àudio amb transcripció a partir de TV3

El subprojecte `audio_subs` incorpora un exemple pràctic basat en contingut de la plataforma 3Cat. L'objectiu és obtenir un corpus ASR a partir de vídeos amb subtítols.

El procés de creació del corpus segueix els següents passos:

4.1 Obtenció d'enllaços de vídeos de TV3

Per construir el corpus de TV3, es parteix de la identificació dels enllaços dels vídeos rellevants a la plataforma `3cat.cat`. Aquesta tasca es realitza mitjançant un script que permet recuperar automàticament les URL de vídeos associats a una sèrie concreta o a partir dels resultats d'una cerca per paraules clau.

L'script utilitza navegació programàtica per simular l'scroll de la pàgina i assegurar que es carreguin tots els resultats disponibles. Les URL extretes es guarden en fitxers de text per a ser processades posteriorment. Aquest procés constitueix la primera fase de construcció del corpus: la recopilació i indexació del material audiovisual d'interès.

4.2 Descàrrega d'àudio i subtítols

A partir de la llista d'enllaços prèviament recollida, es procedeix a la descàrrega del material audiovisual. Utilitzant `yt-dlp`, es descarrega l'àudio dels vídeos en format `m4a` així com els subtítols en català, si estan disponibles. L'àudio es converteix posteriorment a `mp3` amb una sola pista i una freqüència de mostreig de 16 kHz, per garantir compatibilitat amb models d'ASR.

Els subtítols es processen per netejar-ne el format i es converteixen a `srt`. Finalment, tant l'àudio com els subtítols es reanomenen per eliminar caràcters especials i es mouen a la carpeta corresponent dins l'estructura del corpus.

4.3 Segmentació del corpus

Un cop descarregats i preprocessats, els fitxers d'àudio es divideixen en segments a partir dels temps dels subtítols. Per defecte, els segments tenen una durada màxima de 30 segons. El funcionament intern es pot consultar al script `utils/srt-audio-split.py`. Aquest pas facilita l'entrenament o avaluació de models ASR en unitats petites i ben sincronitzades amb la transcripció.