Haejin Cho

(438) 722-0368 | haejin2909@gmail.com |https://www.linkedin.com/in/haejin-cho-220a3a1b2/ |
https://jinjin-tonic.github.io/ | Piano & Video game & Travel | Montreal, QC

EDUCATION

University of British Columbia Vancouver, BC
Master of Data Science - Computational Linguistics EXPECTED 07/2021
Relevant Coursework: NLP; ML; Neural Network; Data Visualization; Corpus Linguistics; Algorithms; Statistics

Yonsei University Seoul, South Korea
Bachelor of Arts - Korean Language and Literature GRADUATED 08/2019
GPA: 4.22/4.3; Highest Honours 2018, 2019;
Relevant Coursework: Syntax; Semantics; Phonology; Morphology; Corpus Linguistics

EXPERIENCE

Yonsei Institute of Language and Information Studies Seoul, South Korea
Research Associate 03/2016 – 09/2017

- Participated in publishing the 8th edition of Dong-a Yonsei's Elementary Korean Dictionary
- Analyzed vocabulary frequency used in elementary textbooks to select lexical items for the dictionary
  Advanced skills: Lexicology, Corpus linguistics, MS excel

PERSONAL PROJECT

COVID-19 Vaccine Sentiment Analysis [Project Link] [GitHub] 12/2020-02/2021

- Achieved an accuracy of 78% with the best model among the trials of different ML models such as LGBM, Logistic Regression, XGBoost and deep learning architectures.
- Deployed an LSTM model to predict sentiments of COVID-19 vaccines related tweets using Heroku and Flask
- Scraped tweets using Tweepy API, manually annotated sentiments of tweets, and preprocessed raw text data using tools in NLTK to perform a supervised learning task
  Advanced skills: FNN, NLTK, LSTM, PyTorch, Tweepy, Sentiment Analysis, Text Annotation, Flask, Heroku

ACADEMIC PROJECTS

Fine-tuning Wav2Vec2 on Non-native English Speech Corpus 04/2021

- Improved the pre-trained wav2vec2-large-960h-lv60-self model from 12.5% WER to 9.7% WER on the entire L2-ARCTIC corpus and from 23.3% WER to 12.7% WER on the L2-ARCTIC Vietnamese corpus.
- Researched the current state of ASR technology on L2 English speakers, fine-tuned the base model with two different sizes of datasets, and experimented with different configurations to achieve better scores.
  Advanced skills: ASR, pyTorch, Huggingface, Wav2Vec 2.0, transformers, Phonology, Git, Python

Building a Multilingual Parallel Corpus using TED transcripts [Blog Link] 03/2021

- Scraped transcripts from TED talks to create a parallel multilingual corpus using BeautifulSoup
- Using spaCy, automatically extracted named entities in English, Chinese corpora and improved annotation quality by using Amazon Mechanical Turk
- Deployed the parallel corpora with a user interface using HTML, JS, CSS, and FastAPI
  Advanced skills: Web Scraping, Python, Corpus Linguistics, spaCy, NER, AMT, FastAPI, Docker, Git

ADDITIONAL INFORMATION

- **Technical Skills**: **Python**, Git, R, **PyTorch**, Pandas, Altair, **NLTK**, **spaCy**, BERT, MS Excel, Jupyter Notebook
- **Soft Skills**: Adaptability, Listening Skills, Self-motivation, Teamwork, Time Management, Problem Solving
- **Languages**: Korean (Native); English (Fluent); Japanese (Conversational); French (Elementary)
- **Extra Accredited Courses**: Linear Algebra, Introduction to Calculus at Athabasca University
- **Certificates:** Japanese Language Proficiency Test N1