SIJIA GE

+1(720) 530-7370 ◇ Boulder, CO
sijiage007@gmail.com ◇ konic-nlp.github.io

EDUCATION

MS in Computational Linguistics, University of Colorado-Boulder **Aug 2021-Dec 2023**
GPA: 3.957/4
MA in Computational Linguistics, Nanjing Normal University **Sep 2016-Jun 2019**
GPA: 86.3/100
BA in Chinese Literature and Linguistics, Shanxi University **Aug 2021-Dec 2023**
GPA: 85.8/100

SKILLS

Technical Python, Keras, Pytorch, Java, NLTK, Scikit-learn, Transformers, Tensorflow, SQL, Django, Flask,HTML Axure RP JavaScript, JQuery, Linux, Shell, SPSS, LaTex, Git, R, PlantUML, Neo4j, Tableau, Heroku, Numpy, Pandas, Scipy, Google Cloud
Languages Madarin Chinese, English

PROJECTS

Bike shared count Kaggle competition *CSCI 5622 Machine Learning Coursework*

- leverage Exploratory Data Analysis, feature engineering for handling data format features, feature transformationnormalization.
- Utilize XGBoost with Parameter Grid search after the feature engineering.
- Rank 3/52 in Public leaderboard with $R^2 = 0.935$ and 9/52 in Private Leaderboard with $R^2 = 0.939$.

Course Project: The bigger, the better? *CSCI 5622 Machine Learning Project*

- compare the performance of the image classification task on the CIFAR-10 dataset between a vanilla CNN and a CNN including ResNet-50 pre-trained models.
- The experiments show that after preprocessing like Deep-Image-Prior(DIP) on the noisy data, the performance on a vanilla CNN can perform at a similar level to that on a complicated CNN leveraging ResNet-50 without denoising. The vanilla CNN trains faster, with fewer weights.
- Also find variables such as optimizer, and upsampling impact differently on different CNN.

UMR-writer: An efficient annotation tool for Uniform Meaning Representation(UMR)

- Update the interface from clicking to typing, supporting frame lexicon display on the fly.
- Add multiple events listener and DOM operation that increased the annotation efficiency by 60%.
- Efficiently optimize the UX and UI, cover multiple languages such as Arapahoe, Navajo, Kukama, Arabic, etc.

E-commerce online shopping website

- Simulate the fundamental functions of an online shopping website, such as user account management, products display, edit cart and checkout.
- Built on Django 2.0, adopted JQueryJavaScript to interact with the users at the front end.
- Adopted SQLite as the model and built-in admin module as the backstage management system.

Viz-Wiz Visual Question Answer Challenge: Answer Visual Questions from People Who Are Blind

- Adopted the feature map on the next to last layer of VGG-16 as the image features.
- Adopted the hidden state of the last layer of BERT as question features.
- classify the concatenated feature into the first 50000 frequent answers and got a score of 0.47.

Named Entity Recognition for Gene text

- Implemented gene named entity recognition with crf++,crfsuite, Bi-LSTM-CRF, and pre-trained language model(BioBERT), and reached entity-based F-scores of 62%, 62%, 77%, and 85% respectively.
- utilized comet_ml as a tool for training visualization

Music Store Simulation

- **Simulate the operation of a music store, including ordering new items, selling items, buying items, and so on.**
- **Apply design patterns to the project, such as strategy, decorator, observer, singleton, etc.**
- **Based on Java and utilizing Junit for unit test**

EXPERIENCE

Student Research Assistant **May 2022 - present**
CLEAR lab, University of Colorado-Boulder ***Boulder, CO***

- **compute the similarity of pair of events whose arguments and events mapped to the nodes in the wikidata using the rule-based and machine-learning algorithms.**
- **update the online annotation tool with a more efficient and better user experience interface with front-end skills and Flask.**

Product Manager Intern **Mar 2021 - Jun 2021**
Beijing Lingosail Tech Co., Ltd. ***Beijing, China***

- **Mainly focus on a term extraction tool whose target users are the students and teachers who use CAT software.**
- **Design the prototype of the new products including basic UI, function, and interaction; Schedule the overall progress of development as well as promotion model;**
- **Test the function of new products with black box testing.**

PUBLICATIONS

- **Sijia Ge. 2022. Integration of Named Entity Recognition and Sentence Segmentation on Ancient Chinese based on Siku-BERT. In Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pages 167–173, Taipei, Taiwan. Association for Computational Linguistics.**
- **Ning Cheng, Bin Li, Liming Xiao, Changwei Xu, Sijia Ge, Xingyue Hao, and Minxuan Feng. 2020. Integration of Automatic Sentence Segmentation and Lexical Analysis of Ancient Chinese based on BiLSTM-CRF Model. In Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 52–58, Marseille, France. European Language Resources Association (ELRA).**
- **Song, L., Wen, Y., Ge, S., Li, B., Qu, W. (2020). An Easier and Efficient Framework to Annotate Semantic Roles: Evidence from the Chinese AMR Corpus. In: Hong, JF., Zhang, Y., Liu, P. (eds) Chinese Lexical Semantics. CLSW 2019. Lecture Notes in Computer Science(), vol 11831. Springer, Cham.**

EXTRA-CURRICULAR ACTIVITIES

- **Peer mentor for international students and linguistic first-year master students during 2022-2023 academic year.**
- **Course Manager for CSCI 5622 Machine Learning, mainly focus on help writing autograding scripts, sanity check and canvas page management.**