

WENJIA ZHAI

(402) 890-4973 | zhai.we@husky.neu.edu | <https://www.linkedin.com/in/wenjia-zhai/>

SKILLS

- **Programming Language:** Python, Scala, Java, R, Perl
- **Programming Frameworks & Tools:** TensorFlow, PyTorch, scikit-learn, NLTK, spaCy, Gensim, tm, tidytext, Regular Expression, Spark, SQLite, MongoDB, Bash Scripting, Git, Google Cloud Platform
- **Data Science Skills:** Data Mining, Feature Engineering, Data Visualization, Natural Language Processing
- **Machine Learning Algorithms:** Neural Network, Clustering, K-nearest Neighbors, Naïve Bayes, Linear & Logistic Regression, Support Vector Machine, Random Forest
- **Operating System:** Windows, Linux/Unix (MacOS, Ubuntu)
- **Report Writing:** Microsoft Office (Word, Excel, PowerPoint), LaTeX, Markdown, Jupyter Notebook
- **Language:** Chinese, English

WORK EXPERIENCE

- Internship: Python Artificial Intelligence Engineer** Beijing, China
Huike Group Sep – Dec 2019
- Rationale:** The company wanted to design a program to automatically segment the videos and labeled each segment by its topic
- Segmented videos based on the length of silence period
 - Built auto speech recognition system (ASR) using **CTC**, **CNN** and **Attention** algorithms, converted audios into texts from videos using **TensorFlow**
 - Conducted **TextRank** model to extract topic of each piece of videos
- Achievement:** The model achieved 62.6% voice recognition accuracy compared to transcript on 20 videos, 89.2% segmentation accuracy compared to manual segmentation on same videos

ACADEMIC PROJECT

- LendingClub Loan Records Analysis in Spark** Boston, MA
Feb 2020
- Constructed a **Spark** data pipeline in **Scala**
 - Deployed the pipeline on **Google Cloud Platform**, managed by **Airflow**
 - Loaded 2 million of loan and rejection records using **Dataproc** service
 - Aggregated rejections and collections records, output a json file containing result

- News with Similar Opinions in News in Python** Boston, MA
July 2019
- Parsed ~1000 news articles from mainstream news website
 - Built a **Word2Vec** model to find the similar words of “say” to locate opinions in news.
 - Captured and extracted opinions based on the location of “say”
 - Conducted a **Doc2Vec** model for **Latent Semantic Analysis** using **Gensim**, obtained the articles with most similar opinions

- Article summarization Generation in Python** Boston, MA
April 2019
- Team Leader*
- Preprocessed article contents and optional titles (tokenizing, converting to lower case, stemming, removing stop words) using **NLTK**
 - Counted appearances of each word as word weights
 - Adjusted word weights by multiplying frequency of words in content and title, if title is supplied
 - Calculated the weights of each sentences based on frequency of words
 - Selected the top 1/3 sentences with heaviest weights as article summarization
 - created an API for this application

Pet Adoption Speed Prediction in R

Team Leader

Boston, MA

Feb 2019

- Performed **Standardization** for continuous features, **Ordinal encoding** for ordinal features, **One-hot encoding** for categorical features, **Sentiment Analysis** for text features
- Generated new features for possibly related features by concatenating them
- Applied **K-nearest neighbors** algorithm to build a baseline model
- Implemented **decision tree**, **support vector machine**, **random forest** algorithms
- Evaluated model performance by accuracy, archived 0.50 on a fine-tuned **XGBoost** model

EDUCATION

Northeastern University, GPA: 4.0/4.0

Boston, MA

M.S. in Bioinformatics

May 2020

Coursework: Bioinformatics, Statistics, Algorithms, Data Mining, Data Visualization, Machine Learning, Neural Network, Natural Language Processing, Linguistics

University of Nebraska – Lincoln

Lincoln, NE

M.S. in Chemical Biology

May 2017

Coursework: Organic Chemistry, Organic Reaction, Chemical Biology, Biochemistry, Analytical Chemistry

China Pharmaceutical University

Nanjing, China

B.S. in Pharmaceutical Chemistry

July 2011

Coursework: Inorganic Chemistry, Organic Chemistry, Analytical Chemistry, Physical Chemistry, Anatomy, Physiology, Pharmacology, pharmacognosy, Pharmaceutical Chemistry, Computer Aimed Drug Design, Spectrum