YI WEI

wei256@usc.edu
(213) 2466261
github.com/watsonwei
linkedin.com/in/watsonwei

EDUCATION

University of Southern California *Los Angeles, US*
- Master of Science in Data Informatics. GPA: 3.77/ 4.0 Expected December 2018
- Courses: Machine Learning, Information Visualization, Building Knowledge Graphs, Data Mining

Hebei University of Technology *Tianjin, China*
- Bachelor of Engineering in Software Engineering September 2012- June 2016
- Courses: Data Structure, Operating System, Database Theory and Application, Computer Network

EXPERIENCE

Software Engineer Intern *Sohu.Com Inc [Java, Python, Spark, xLearn]* *May 2018 – August 2018*
- Implemented bucket testing on backend of related news recommendation and remotely debugged bad cases.
- Calculated news similarity based on tag vector using cosine similarity and eliminated duplicates in enchache.
- Built a hadoop data pipeline using spark to generate data in libffm format, created cron jobs to update data per hour. Trained field-aware factorization model with xLearn using Screen on linux server.

Student Researcher *Keck School of Medicine of USC [Python,Bash]* *January 2018 – February 2018*
- Built data pipeline to generate sequences alignment, founded local alignment using Blast, outputted as csv file.
- Implemented global alignment using context dependent algorithm, saved as fasta files.
- Eliminated gaps-dominant part for both ends, calculated mismatches and percentage of identical matches.

PROJECTS

Knowledge Graph for S&P 500 Companies *[Scrapy, Tweepy, StanfordNERTagger, Elasticsearch, Kibana]*
- Built a visualization board, search bar for user-interested companies and companies knowledge graph.
- Batched loading data to elasticsearch from CNBC news, tweets, 10-k forms, and stock price from yahoo finance.
- Extracted company names from news and Entity linking between news and companies using Jaro-Winkler.

Visualization for Data on Income vs. University/Degree *[D3.js,javascript,Bootstrap]*
- Constructed an interactive website using bootstrap that visualizes the data of salaries by majors and schools.
- Built a stacked bar chart using D3's stack shape layout, a parallel coordinate plot with D3's brush on every axis and an interactive map using GeoJSON data.

Recommendation System with Spark Framework *[Spark, Spark MLlib, Scala]*
- Implemented user-based and a model-based collaborating filtering recommendation system.
- Generated similar products from Amazon review data using LSH algorithm based on Jaccard, cosine similarity.
- Implemented SON algorithm based on Apriori and Girvan-Newman algorithm using MapReduce on Spark.

Breast Cancer Prediction with Random Forest and SVM *[Python, scikit-learn, Pandas]*
- Eliminated dimensions of data based on heat map and principal component analysis.
- Generated predictions applying random forest and made classification using SVM.
- Outputted accuracy of each model and compared accuracy among different groups of data.

Data Cleaning of Cars Market Information from Craigslist *[Python, Scrapy, NLP, scikit-learn]*
- Built wrapper by Beautiful Soup to extract desired information from structured text and saved as JSON file.
- Constructed a CRF classifier by python-crfsuite to classify unstructured text with NLP toolkit and reported classifier's precision, recall and F-1 measure, cleaned data using OpenRefine and exported as JSON file.

SKILLS

- **Programming:** Python, Java, SQL, Spark, Scala, HTML, CSS, D3.js, Matlab, R
- **Tools:** Git, IntelliJ, Linux, PyCharm, Jupyter Notebook, vim