Research Internship (Unpaid): Data integrity checking at the FrameNet Project
Spring, 2017

FrameNet (`http://framenet.icsi.berkeley.edu`) is a linguistic research project that produces a"super dictionary", a lexical database that is both human-readable and useful in Natural Language Processing (NLP). The FrameNet team develops descriptions of semantic frames based on the late Prof. Charles J. Fillmore's theory of Frame Semantics and manually annotates example text from corpora to document how the lexical units (word senses) in those frames are used in sentences: what semantic roles are represented by the other parts of the clause, and what syntactic patterns they can appear in. This ongoing project produces and freely distributes the FrameNet lexical database, which is used in thousands of NLP research projects and commercial applications around the world. FrameNet is hosted at the International Computer Science Institute (`http://icsi.berkeley.edu`), along with other CS research projects on Internet traffic analysis and security, bioinformatics, computer vision, etc. ICSI is less than 5 minutes' walk from BCC. This is an excellent opportunity to learn about databases in an NLP project that is both academic and practical.

The FN data is stored in a MySQL database containing several dozen tables; these can be divided conceptually into a frame-and-lexicon db and an annotation db. The former contains on the order of 10,000 records, latter on the order of 1 million records. Like all large databases, these contain errors, in this case, mainly due to errors in manual annotation. Previous work on data integrity has created two types of scripts, an elegant expert system, written in Haskell, to report on errors in the frame-lexical db, and the other, written in a mixture of Python and Java, to find errors in the annotation db.

The intern will be expected to (1) analyze the existing code bases, (2) reorganize them as needed to produce runnable systems, (3) test them by running them over the database, and (4) extend them to find other types of problems, and detect outliers in the data which may not be errors, but are interesting to researchers.

Skills needed: basic knowledge of database operations (coursework and/or experience), resonable fluency in Python (knowledge of Java also helpful).

Please apply by email to the following address, attaching (1) resume, (2) cover letter, and (3) current (unofficial) transcript. Incomplete applications will not be reviewed. Last date to apply: Feb. 2, 2017.

`Collin Baker`
`Project Manger, FrameNet`
`collinb@icsi.berkeley.edu`

1947 Center Street, Suite 600
Berkeley, CA 94704-1198 USA
(510) 666-2900, (510) 666-2956 fax
www.icsi.berkeley.edu