(https://github.com/HazyResearch/deepdive)

DeepDive
(http://deepdive.stanford.edu/)

Fork me on GitHub

- Quick Start (http://deepdive.stanford.edu/quickstart)
- Documentation (http://deepdive.stanford.edu/#documentation)
- Showcase (http://deepdive.stanford.edu/showcase/apps)
- Papers (http://deepdive.stanford.edu/papers)
- Data (http://deepdive.stanford.edu/opendata)
- Chat (https://gitter.im/HazyResearch/deepdive)
- Forum (https://groups.google.com/d/forum/deepdive-users)

Distant supervision

Most machine learning techniques require a set of **training data**. A traditional approach for collecting training data is to have humans label a set of documents. For example, for the marriage relation, human annotators may label the pair "Bill Clinton" and "Hillary Clinton" as a positive training example. This approach is expensive in terms of both time and money, and if our corpus is large, will not yield enough data for our algorithms to work with. And because humans make errors, the resulting training data will most likely be noisy.

An alternative approach to generating training data is **distant supervision**. In distant supervision, we make use of an already existing database, such as Freebase (http://www.freebase.com/) or a domain-specific database, to collect examples for the relation we want to extract. We then use these examples to automatically generate our training data. For example, Freebase contains the fact that Barack Obama and Michelle Obama are married. We take this fact, and then label each pair of "Barack Obama" and "Michelle Obama" that appear in the same sentence as a positive example for our marriage relation. This way we can easily generate a large amount of (possibly noisy) training data. Applying distant supervision to get positive examples for a particular relation is easy, but generating negative examples (generating_negative_examples) is more of an art than a science.

Questions? Chat with us (https://gitter.im/HazyResearch/deepdive) or ask deepdive-users (https://groups.google.com/d/forum/deepdive-users)!
Fork DeepDive on Github (https://github.com/HazyResearch/deepdive) and join deepdive-dev (https://groups.google.com/d/forum/deepdive-dev).