HW #9

Goals and Task

- Implementing Coherence Relation Sense Classification
  - Part of Shallow Discourse Parsing pipeline
- Goals:
  - Explore issues in shallow discourse parsing.
  - Gain familiarity with the Penn Discourse Treebank and CoNLL data.
  - Gain some further familiarity with vector-based word embeddings
  - Implement a relation sense classification system.

Components

- We provide:
  - Gold data in CoNLL16 format
    - Train and test split
  - 50-dimensional GloVe embeddings trained on Wikipedia and Gigaword

- You:
  - Read in the data
  - Build train/test classification vectors
    - For each of Arg1, Arg2: average word vectors together to build total vector
  - Train a classifier on test vectors, evaluate on test vectors

Data Example (One Line of JSON)

- Arg1:
  - RawText
  - …
- Arg2:
  - RawText
  - …
- Connective:
  - RawText
- **Sense**
- Type (Explicit or Implicit)
- …

Training a Classifier

- You can use any pre-implemented classifier that you'd like
- scikit-learn offers many, e.g.:
  - SVM
  - Nearest neighbors
- Usual API:
  - Instantiate model
  - model.fit(X, Y): train the model
  - model.predict(X): make predictions on new inputs