Boltzmann Brain Planning

[Structuralism is] “the belief that phenomena of human life are not intelligible except through their interrelation. These relations constitute a structure and behind local variations in the surface phenomena there are constant laws of abstract culture.”

T. Accessing Freebase

What it is: Getting Freebase into an easily accessible form
What it isn’t: Any learners / intelligence built on Freebase

- Access in a graph form
- Identify predicates (classes)
- Store multiple versions (remove training sets)
- Start back up versioned downloads
- Eval Long term storage overhead
- Document “Hello Freebase” access
- iPyNB / SciTalk / LunchTalk

T. Evaluation Metrics

What it is: an automated system for testing RDF extractions
What it isn’t: Turk work

- Define evaluation (# correct RDFs / # incorrect RDFs) under LCWA
- Ensure statistically rigid sampling (train/test split %, stratified sampling, …)
- Storage and versioning (both algo and freebase version) of metrics, extracted named entities per page, text features per page

T. DOM extraction

What it is: using relative DOM paths to extract RDF triples
What it isn’t: full blown text extraction

- Build NER on docs via hbase table
- Store NER-extracted terms in ES field
- Licensing terms on Stanford NER & POS
- Eval dom extractors (lxml / webstruct / custom) for speed and ease implementation
- Build infra to run extractor via hbase table (X_latest_Y)
- Train logistic regression (via L-BFGS) w/ Gauss regularization
- Eval accuracy + tune ML
- Entity-linked field in ES

T. Discuss Entity Bidding

What it is: evaluate biz side & tech side of whether entity extarction good enough for auction
What it isn’t: building an auction

- Discuss model stability & how it relates to marketplace stability

- Debate implications relating to LCWA for auctions
- ? are there enough entities to form a dynamic marketplace
- Implications with regard to SOX
- Discuss implications of record linkage
    - multiple linking
    - link confidence scores

T. Initial search boost
What it is: a new personality to boost on extracted entities and relations
What it isn't: perfect

- Discuss how to improve search relevancy using NER+DOM

T. Offline DOM classifier model-building
What it is: A system for automatically re-training classifiers based on extracted facts
What it isn't: a webserver. classifying queries. productionalized

- Decide which classifiers are worth pursuing
- Geo / Maturity / Date / Price / Page Topic labeling (mturk)
- Build ML models to learn above classifiers using extracted DOM features, entities and relations
- science debrief
- evaluations for classification accuracy

T. Productionalize classifier-building
What it is: Storage & automation around classifier-building.
What it isn't: webserver, online query classification

- model versioning
- algo versioning
- model build a-la spark
- model deployment (ZK)
- matrix transformation
- storage

T. Online DOM classifier service
What it is: ability to use classifiers online for incoming queries
What it isn't: model building

- Build search personality stuffs
- webservice for classifier
- circuit breaker on classifier
- evaluate query performance with classifiers: AB, mturk, SRM

T. I can haz more text
What it is: Natural language text extractor
What it isn't: fusion learning

- Eval technologies for lexical and syntactic (MINIPAR) extractors
- Use previous infra to robustly extract features for each page
- Train log reg
- Eval accuracy and tune ML (independently from other classifiers)
- science debrief

T. Scaling & Regularized Features

What it is: properly scaling features so fusion learners don't freak out
What it isn't: actual fusion

- Learn wtf is Platt scaling (...logistic regression)
- implement said scaling per feature
- Figure out how to represent feature scale parameters
- Figure out if we want to scale features (?Rand forest)
- science debrief

T. Fusion Learners

What it is: A voting "boosting" ensemble for determining bias and selection in feature extractors
What it isn't: Not a bagging ensemble, nor a cascade ensemble.

- Separate classifier per predicate
- Implement Adaboost w/decision stumps (sklearn)
- Scaling & regularization issues?
- Storage and annotating which learners produced which features
- Eval RDF performance
- New search personality?
- Retrain classifiers?
- science debrief ?

T. Fusion Learner Online usage

What it is: Test the ensemble vs not using it
What it isn't:

- A/B test / mturk / srm and all that jazz
- science debrief

T. HTML Tables

What it is: A system for extracting RDFs & K=V pairs from web tables
What it isn't: able to extract all information from all tables ever conceived

- Implement web table extractor and storage (what's a table?)
- ML (so prior, very Bayes (maybe association rules etc etc.)
- Eval accuracy and tune learners
- science debrief

T. Schema.org

What it is: using human labels to boost text extractors
What it isn't: Dense

- Define manual mappings
- Conjure up ontological mapping
- Feature description from JSON-LD schema.org -> RDF triples
- Consider automated mappings
- Reconsider and stick with manual mappings
- Extract *scores* along w Schema.org markup
- Eval accuracy and tune

T. Refusion
What it is: retuning the fusion learner w/4 new features (2 new extractors)
What it isn't: the full Boltzmann Brain

- Hopefully *easy* if T7 is done well.
- Retuning fusion learner with more features
- Eval performance on RDF extraction
- search personality
- classifiers

T. Path Ranking Algorithm (PRA)
What it is: A random walk over Freebase to build prior knowledge
What it isn't: complete. ever.

- Storage and computability of random walks
- Computation of pr(edge | walk)
- gibbs sampling
- Storage of associated rules and probabilities
- Tuning

T. Refusion
What it is: Fusion of PRA with fused text leaners
What it isn't: The full BB

- See above refusions

T. Neural Networks (Or faking low rank tensor decomposition)
What it is: another prior
What it isn't: fast. or efficient. or perfect.

- Read all the papers
- Experiment with different tensor decomp methods. Recommended is multilayer perceptron 'cause it's speedy
- ? SDA for sparse coding
- Find one or a fusion that works well

T. Nirvana

T. Future work? see paper; LCWA, mutual exclusion of facts (soft correlation), layers of abstraction, correlated source bias, temporal nature of facts, adding new entities / relations...