DATASCI W266: Natural Language Processing with Deep Learning

Course Overview

Understanding language is fundamental to human interaction. Our brains have evolved language-specific circuitry that helps us learn it very quickly; however, this also means that we have great difficulty explaining how exactly meaning arises from sounds and symbols. This course is a broad introduction to linguistic phenomena and our attempts to analyze them with machine learning. We will cover a wide range of concepts with a focus on practical applications such as information extraction, machine translation, sentiment analysis, and summarization.

Prerequisites:

- Language: All assignments will be in Python using Jupyter notebooks, NumPy, and TensorFlow.
- Time: There are 5-6 substantial assignments in this course as well as a term project. Make sure you give yourself enough time to be successful! In particular, you may be in for a rough semester if you have other significant commitments at work or home, or take both this and any of 210 (Capstone), 261, or 271 :)
- MIDS 207 (Machine Learning): We assume you know what gradient descent is. We'll review simple linear classifiers and softmax at a high level, but make sure you've at least heard of these! You should also be comfortable with linear algebra, which we'll use for vector representations and when we discuss deep learning.

Contacts and resources:

- Course website: GitHub datasci-w266/2018-fall-main
- Piazza - we'll use this for Q&A, and this will be the fastest way to reach the course staff. Note that you can post anonymously, and/or make posts visible only to instructors for private questions.
- Email list for course staff: mids-nlp-instructors@googlegroups.com

Live Sessions:

- Tuesday 4 - 5:30p Pacific (Daniel Cer)
- Tuesday 6:30 - 8p Pacific (James Kunz)
- Wednesday 6:30 - 8p Pacific (Blake Lemoine)
- Thursday 4 - 5:30p Paciifc (Joachim Rahmfeld)
- Thursday 6:30 - 8p Pacific (Mark Butler)
- Friday 4 - 5:30p Pacific (Sid J Reddy)

Teaching Staff Office Hours:

- **Daniel Cer**: Wednesday at noon Pacific.
- **Drew Plant / Legg Yeung**: Saturday at 1:30 - 2:30pm Pacific.
- **James Kunz**: Tuesday immediately after his live session (8pm Pacific).
- **Joachim Rahmfeld**: Thursday immediately after his live session (5:30pm Pacific).
- **Mark Butler**: Thursday immediately after his live session (8pm Pacific).
- **Sid J Reddy**: Friday at 3pm Pacific.

Office hours are for the whole class; students from any section are welcome to attend any of the times above.

Async Instructors:

- Dan Gillick
- James Kunz
- Kuzman Ganchev

Grading

Breakdown

Your grade report can be found at https://w266grades.appspot.com.

Your grade will be determined as follows:

- **Assignments**: 40%
- **Final Project**: 60%
- **Participation**: Up to 10% bonus

There will be a number of smaller assignments throughout the term for you to exercise what you learned in async and live sessions. Some assignments may be more difficult than others, and may be weighted accordingly.

Participation will be graded holistically, based on live session participation as well as participation on Piazza (or other activities that improve the course this semester or into the future). Do not stress about this part.

Letter Grades

We curve the numerical grade to a letter grade. While we don't release the curve, it usually results in about a quarter of the class each receiving A, A-, B+, and B. Exceptional cases receive A+, C, or F, as appropriate.

A word of warning: Given that we (effectively) release solutions to assignments in the form of unit tests, it shouldn't be surprising that most students earn near perfect scores. Since the variance is so low, assignment scores aren't the primary driver of the final letter grade for most students. A good assignment score is necessary, but not sufficient, for a strong grade in the class. A well structured, novel project with good analysis is what makes the difference between a high B/B+ and an A-/A.

As mentioned above: this course is a lot of work. Give it the time it deserves and you'll be rewarded intellectually and on your transcript.

Late Day Policy

We recognize that sometimes things happen in life outside the course, especially in MIDS where we all have full time jobs and family responsibilities to attend to. To help with these situations, we are giving you **5 "late days"** to use throughout the term as you see fit. Each late day gives you a 24 hour (or any part thereof) extension to any deliverable in the course **except** the final project presentation or report. (UC Berkeley needs grades submitted very shortly after the end of classes.)

Once you run out of late days, each 24 hour period (or any part thereof) results in a **10 percentage point deduction** on that deliverable's grade.

You can use a **maximum of 2 late days** on any single deliverable. We will **not be accepting any submissions more than 48 hours past the original due-date**, even if you have late days. (We want to be more flexible here, but your fellow students also want their graded assignments back promptly!)

We don't anticipate granting extensions beyond these policies. Plan your time accordingly!

More serious issues

If you run into a more serious issue that will affect your ability to complete the course, please email the instructors mailing list and cc MIDS student services. A word of warning though: in previous sections, we have had students ask for INC grades because their lives were otherwise busy. Mostly we have declined, opting instead for the student to complete the course to the best of their ability and have a grade assigned based on that work. (MIDS prefers to avoid giving INCs, as they have been abused in the past.) The sooner you start this process, the more options we (and the department) have to help. Don't wait until you're suffering from the consequences to tell us what's going on!

Final Project

See the Final Project Guidelines

Course Resources

We are not using any particular textbook for this course. We'll list some relevant readings each week. Here are some general resources:

- Speech and Language Processing (2nd edition) (Jurafsky and Martin)
- Speech and Language Processing (3rd edition draft) (Jurafsky and Martin) - *free online!*
- NLTK Book - Accompanies NLTK (Natural Language ToolKit) and includes useful, practical descriptions (with python code) of basic concepts.
- Deep Learning (Goodfellow, Bengio, and Courville)

We'll be posting materials to the course GitHub repo.

Note: the syllabus below might be subject to change. We'll be sure to announce anything major on Piazza.

Code References

The course will be taught in Python, and we'll be making heavy use of NumPy, TensorFlow, and Jupyter (IPython) notebooks. We'll also be using Git for distributing and submitting materials. If you want to brush up on any of these, we recommend:

- **Git tutorials:** Introduction / Cheat Sheet, or interactive tutorial
- **Python / NumPy:** Stanford's CS231n has an excellent tutorial.
- **TensorFlow:** We'll go over the basics of TensorFlow in Assignment 1.
  Effective TensorFlow is a great reference, ranging from the absolute basics through advanced topics like multi-GPU training, `tf.learn`, and debugging.
  You can also check out the tutorials on the TensorFlow website, but these can be somewhat confusing if you're not familiar with the underlying models.

Misc. Deep Learning and NLP References

A few useful papers that don’t fit under a particular week. All optional, but interesting!

- (optional) Chris Olah’s blog and Distill
- (optional) GloVe: Global Vectors for Word Representation (Pennington, Socher, and Manning, 2014)

Schedule and Readings

We'll update the table below with assignments as they become available, as well as additional materials throughout the semester. Keep an eye on GitHub for updates!

Dates are tentative: assignments in particular may change topics and dates. (Updated slides for each week will be posted during the live session week.)

	Async to Watch	Topics	Materials
Week 1 (September 3 - 9)	Introduction 5.3 Softmax Classification 5.4 Neural network recap 5.5 Neural network training loss	- Overview of NLP applications - Ambiguity and grounding in language - Information theory and linear algebra review - ML models: Logistic regression and feed forward networks	- Skim: NLTK book chapter 1 (python and basics) - Skim: NLTK book chapter 2 (data resources) - Read: AI's Language Problem (Technology Review) - Read: The Rise and Fall of the English Sentence - *Optional:* The Interpreter (New Yorker) - *Optional:* Introduction to Linguistic Typology
Assignment 0 released September 3 due September 9	**Course Set-up**	- GitHub - Piazza - Google Cloud	Assignment 0
Week 2 (September 10 - 16)	Classification and Sentiment (up to 2.6)	- Sentiment lexicons - Aggregated sentiment applications - Bag-of-word models	- Skim: Opinion Mining and Sentiment Analysis (Pang and Lee 2008) - focus on Chapters 1-4 - Read: Understanding Convolutional Neural Networks for NLP - Read: Convolutional Neural Networks for Sentence Classification (Yoon Kim, 2014) - *Optional:* Natural Language Processing (almost) from Scratch (Collobert et al., 2011)

		• Introduction to Word embeddings	
Assignment 1 released September 7 due September 16	**Background and TensorFlow**	• Information Theory • Dynamic Programming • TensorFlow Introduction	Assignment 1
Week 3 (September 17 - 23)	Classification and Sentiment (2.7 onwards) *Note: you may want to review Async 5.3, 5.4, and 5.5.*	• Convolutional neural networks for NLP	•
Week 4 (September 24 - 30)	Language Modeling I, 4.1-4.4, 4.8 - 4.11	• LM applications • N-gram models • Smoothing methods • Representations of meaning	Language model introduction: • Skim: Chen and Goodman Survey • Skim: 1 Billion Word Benchmark • *Optional:* Natural Language Corpus Data (Peter Norvig)

		• Distributed representations	[Language Modeling Notebook] Distributed representations: • Read: Brown Clustering (Brown et al. 1992) • Read: CBOW and SkipGram (Mikolov et al. 2013) • *Optional:* Deep Learning, NLP, and Representations (Chris Olah's blog) • *Optional:* Tensorflow Word2Vec Tutorial(just the parts on word2vec_basic.py - don't bother with the "Optimizing the Implementation" part or anything in C++) • *Optional:* How Vector Space Mathematics Reveals the Hidden Sexism in Language (and the original paper) [Word Embeddings Notebook][TensorFlow Embedding Projector]
Assignment 2 released September 21 due September 30	**Text Classification**	• Exploration & Naive Bayes • Neural Bag-of-Words • Convolutional neural networks	Assignment 2
Week 5 (October 1 - 7)	Language Modeling II	• Neural Net LMs • Word embeddings • Hierarchical softmax	• Read: A Neural Probabilistic Language Model (Bengio et al. 2003) • Read or skim: How the backpropagation algorithm works • *Optional:* Understanding LSTM Networks (Chris Olah's blog)

		• State of the art: Recurrent Neural Nets	• *Optional (skim):* Tensorflow LSTM Language Model Tutorial • *Optional / fun:* Tensorflow Playground [NPLM Notebook]
Project Proposal due October 7			**Final Project Guidelines**
Interlude (Extra Material)	Basics of Text Processing	• Edit distance for strings • Tokenization • Sentence splitting	• Skim: NLTK book chapter 3 (processing raw text) • Skim: Natural Language Corpus Data(Peter Norvig) *(if you didn't read in Week 2)* • Read: Sentence Boundary Detection and the Problem with the U.S.
Week 6 - 7 (October 8 - 21)	Machine Translation I Machine Translation II	• Word- and phrase-based MT • IBM alignment models • Evaluation • Neural MT with sequence-to-sequence models and attention	• Read: Sequence to Sequence Learning with Neural Networks • Read: Neural Machine Translation by Jointly Learning to Align and Translate • *Optional:* Google's Neural Machine Translation System • *Optional:* Attention and Augmented Recurrent Neural Networks (section on "Attentional Interfaces" has an awesome visualization of an MT example, showing alignments)

Assignment 3 released October 5 due October 21	**Language Models and Word Embeddings**	- Smoothed n-grams - Exploring embeddings - RNNLM	Assignment 3
Week 8 (October 22 - 28)	Summarization	- Single- vs. multi-document summarization - Extractive and abstractive summarization - Classical summarization algorithms - Evaluating generated summaries	- Skim: A Survey on Automatic Text Summarization (Das and Martins, 2007) - Read: A Neural Attention Model for Abstractive Sentence Summarization(Rush et al. 2015) - *Optional:* Get To The Point: Summarization with Pointer-Generator Networks (See et al. 2017)
Assignment 5 released October TBD due October TBD	**Assignment 4**	- TBD	Assignment 5
Week 9 (October 29 - November 4)	Part-of-Speech Tagging I	- Tag sets - Most frequent tag baseline	- Read: NLTK book chapter 5: Categorizing and Tagging Words

		• HMM/CRF models **Note:** Section 7.6 this week in the async is optional.	[Interactive HMM Demo] • Read: A Universal Part-of-Speech Tagset • Read: Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?
Week 10 (November 12 - 18)	Dependency Parsing	• Dependency trees • Transition-based parsing: Arc-standard, Arc-eager • Graph based parsing: Eisner Algorithm, Chu-Liu-Edmonds	• Read: SyntaxNet (Parsey McParseface) • Read: Dependency Parsing (J&M Chapter 14) • *Optional:* A Fast and Accurate Dependency Parser using Neural Networks (Chen & Manning 2014)
Week 11 (November 19 - 25)	Constituency Parsing	• Context-free grammars (CFGs) • CYK algorithm • Probabilistic CFGs • Lexicalized grammars, split-merge, and EM	• Read: NLTK book chapter 8 (analyzing sentence structure) • Skim: Accurate Unlexicalized Parsing(Klein & Manning 2003) • Play: Stanford parser (online demo) • *Optional / reference:* Penn Treebank Constituent Tags [Interactive CKY Demo]

Week 12 (November 26 - December 2)	Information Retrieval	• Building a Search Index • Ranking • TF-IDF • Click signals	• Read: Web Search for a Planet (Google) • Read: The Anatomy of a Large-Scale Hypertextual Web Search Engine • Skim: "An Introduction to Information Retrieval", sections 6.2 and 6.3 • *Optional:* PageRank (Page, et al. 1999)
Week 13 (December 3 - 9)	Entities	• From syntax to semantics • Named Entity Recognition • Coreference Resolution	• Read: NLTK Book Chapter 7 (Extracting Information from Text) • *Optional:* Simple Coreference Resolution with Rich Syntactic and Semantic Features (Haghighi and Klein 2009, rule-based coreference) • *Optional:* Improving Coreference Resolution by Learning Entity-Level Distributed Representations (Clark and Manning 2016, neural coreference)
Project Reports **due December 7** **(hard deadline)**			**Final Project Guidelines**
Project Presentations in-class December 10-14			**Final**

Live Session Slides: [available here]