README.md

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:

- Python: 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 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-w266/2017-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:

- Monday 4p - 5:30p Pacific (James Kunz)
- Monday 6:30p - 8p Pacific (Melody Dye)
- Tuesday 6:30p - 8p Pacific (Melody Dye)
- Friday 4p - 5:30p Pacific (James Kunz)

Office Hours:

- Immediately after the live sessions.
- Wednesday or Thursday most weeks (see ISVC) with Drew.

Teaching Staff

- Arathi Mani
- Melody Dye
- Drew Plant
- Ian Tenney
- James Kunz

Async Instructors:

- Dan Gillick
- Kuzman Ganchev

Grading

Breakdown

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

Your grade will be determined as follows:

- **Participation**: 10%
- **Assignments**: 50%
- **Final Project**: 40%

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 will be weighted accordingly.

Participation will be graded holistically, based on live session attendance and participation as well as participation on Piazza. (Don't stress about this part.)

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 contact the instructors and 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.)

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:

- Jurafsky and Martin: Speech and Language Processing

- NLTK Book accompanies NLTK (Natural Language ToolKit) and includes useful, practical descriptions (with python code) of basic concepts.

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

Note: this is a relatively new class, and 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. 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
- (optional) Natural Language Processing (almost) from Scratch (Collobert and Weston, 2011)
- (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 1 - 7)	Introduction	• Overview of NLP applications • Ambiguity in language • General concepts	• Skim: NLTK book chapter 1 (python and basics) • Skim: NLTK book chapter 2 (data resources) • Read: AI's Language Problem (Technology Review) • *Optional:* The Interpreter (New Yorker) • *Optional:* Introduction to Linguistic Typology [Slides] [Tensorflow Intro notebook]
Assignment 0 due September 7	**Course Set-up**	• GitHub • Google Cloud	Assignment 0
Week 2 (September 8 - 14)	Classification and Sentiment	• Sentiment lexicons • Aggregated sentiment applications • Convolutional neural networks (CNNs)	• 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
Assignment 1 due September 14	**Background and TensorFlow**	• Information Theory • TensorFlow tutorial	Assignment 1 [Tutorial Slides]
Week 3 (September 15 - 21)	Language Modeling I	• LM applications • N-gram models • Smoothing methods • Text generation	• Skim: Chen and Goodman Survey • Skim: 1 Billion Word Benchmark • *Optional:* Natural Language Corpus Data (Peter Norvig) [Slides] [Language Modeling Notebook]
Project Proposal due September 21			**Final Project Guidelines** [Project Overview / Topic Slides] [Note on LDC Corpora Access]
Week 4 (September 22 - 28)	Clusters and Distributions	• Representations of meaning • Word classes • Word vectors via co-occurrence counts • Word vectors via prediction (word2vec)	• 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) [Slides] [Word Embeddings Notebook] [TensorFlow Embedding Projector]
Week 5-6 (September 29 - October 12)	Language Modeling II	• Neural Net LMs • Word embeddings • Hierarchical softmax • State of the art: Recurrent Neural Nets	• 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) • *Optional (skim):* Tensorflow LSTM Language Model Tutorial • *Optional / fun:* Tensorflow Playground

			[Slides] [NPLM Notebook]
Assignment 2 due October 5	**n-grams and Word Embeddings**	• Smoothed n-grams • Exploring embeddings	Assignment 2
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. Slides
Assignment 3 due October 12	**Dynamic Programming Intro**	• Dynamic programming	Assignment 3 *Note: this is a somewhat shorter assignment than usual.*
Week 7-8 (October 13 -November 2)	Part-of-Speech Tagging I	• Tag sets • Most frequent tag baseline • HMM/CRF models **Note:** Section 7.6 this week in the async is optional.	• Read: NLTK book chapter 5: Categorizing and Tagging Words [Tagging Slides] [Interactive HMM Demo]
Optional	Part-of-Speech Tagging II	• Feature engineering • Leveraging unlabeled data • Low resource languages **Note:** This week's async is optional (but good reference material if your project focuses on POS tagging). We will spend live session time to reinforce Week 7 material.	• Read: A Universal Part-of-Speech Tagset • Read: Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?
Assignment 4 due November 2	**RNN Language Model**	• RNNLM structure • TensorFlow implementation	Assignment 4
Week 9 (November 3 - 9)	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) [Parsing Slides]
Week 10 (November 10 - 16)	Constituency Parsing	• Context-free grammars (CFGs) • CYK algorithm	• Read: NLTK book chapter 8 (analyzing sentence structure)

		• Probabilistic CFGs • Lexicalized grammars, split-merge, and EM	• Skim: Accurate Unlexicalized Parsing (Klein & Manning 2003) • Play: Stanford parser (online demo) • *Optional / reference:* Penn Treebank Constituent Tags [Interactive CKY Demo]
Week 11-1 (November 17 - 23)	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) [Slides]
Week 11-2 (November 17 - 23)	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) [Slides]
Project Milestone due November 16			**Final Project Guidelines** [Note on LDC Corpora Access]
Assignment 5 due November 29	**Tagging and Parsing**	• HMMs / Forward-Backward and Viterbi • Parsing / CKY	Assignment 5 [HMM Demo] [CKY Demo]
Week 12 (November 24 - November 30)	Machine Translation I	• Word-based translation models • IBM Models 1 and 2 • HMM Models • Evaluation	• Skim: Statistical MT Handbook by Kevin Knight [MT Slides]
Week 13 (December 1 - 7)	Machine Translation II	• Phrase-based translation • Neural MT via sequence-to-sequence models • Attention-based models	• 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)
Week 14 (December 8 - 15)	Summarization	• Single- vs. multi-document summarization • Maximum marginal relevance (MMR) algorithm • Formulation of a summarization objective • Integer linear programming (ILP) for optimal solutions • Evaluation of 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) [Slides]
Project Presentations in-class December 16-22			**Final Project Guidelines**
Project Reports **due December 19** **(hard deadline)**			**Final Project Guidelines**