Week 0:

1. [Class Syllabus](https://docs.google.com/document/d/1pzzQYX-sDtgvPOdD-fMy_yGUqh8V4pQrtMsJ9N1u7YE/edit)
2. [The NLTK book](http://www.nltk.org/book/) its [1st ed](http://www.nltk.org/book_1ed/)
3. [NLTK download](http://nltk.org/)
4. [Python 3 highlight](https://docs.python.org/dev/whatsnew/3.0.html)
5. [Python 2 to 3](https://docs.python.org/2/library/2to3.html)
6. [NLTK Table VIII.1](http://www.nltk.org/book/ch00.html#tab-modules)
7. [class github](https://github.com/datasci-w266/main/)
8. [Final project guideline](https://docs.google.com/document/d/1IR-R8PsEGriB1sVNCwotSe52m5fiO3UYA6DmRtESmso/preview)
9. [Final project hunt](https://docs.google.com/document/d/1rW5HGoC__dxaP3LHYkAQ1Nscu2Axca9e4jeqcfOEQeo/edit)
10. [Cross-Section](https://docs.google.com/document/d/1lF6BRxXzL85ZTENHg8OoJOvIjlH2BehH8yJVHpiOCYk/preview)
11. [Tensorflow github](https://github.com/tensorflow)

**MIDS: Natural Language Processing**

Dan Gillick

Kuzman Ganchev

**Course Summary**

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.

**Assignments and Grading**

Course grades will be based mostly on 4 python programming assignments: 3 guided projects, and 1 more open-ended final project.

**Course Resources**

General resources:

* [NLTK Book](http://www.nltk.org/book/)
* Jurafsky and Martin: Speech and Language Processing

**Week 1 [Introduction]**

* Skim: NLTK book chapter 1 (python and basics)
* Skim: NLTK book chapter 2 (data resources)
* Optional: [The Interpreter (New Yorker)](http://www.newyorker.com/magazine/2007/04/16/the-interpreter-2)
* Optional: [Introduction to Linguistic Typology](https://www.uio.no/studier/emner/hf/ikos/EXFAC03-AAS/h05/larestoff/linguistics/Chapter%204.(H05).pdf)
* Overview of NLP applications
* Ambiguity in language
* General concepts
* TensorFlow tutorial
* <https://github.com/datasci-w266/main> github

**Week 2 [Language Modeling I]**

* Skim: [Chen and Goodman Survey](http://www.cs.berkeley.edu/~klein/cs294-5/chen_goodman.pdf)
* Skim: [1 Billion Word Benchmark](http://arxiv.org/pdf/1312.3005.pdf)
* Speech Recognition, Translation, Phone Keyboards, Handwriting Recognition
* N-gram models
* Smoothing methods
* Text generation

**Week 3 [Clusters and Distributions]**

* Read: [Brown Clustering](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.9919&rep=rep1&type=pdf)
* Read: [CBOW and SkipGram](http://arxiv.org/pdf/1301.3781.pdf)
* Optional: [Tensorflow Word2Vec Tutorial](https://www.tensorflow.org/versions/master/tutorials/word2vec/index.html)
* Representations of meaning
* Word classes
* Word vectors via co-occurrence counts
* Word vectors via prediction (word2vec)
* Comments: for Brown Clustering, for C and C’, why the same k?

**Week 4 [Language Modeling II]**Question: 4.7 huffman coding: not uniform length, problem?

* QUestion: 4.7 Quiz training only? No prediction?
* Question: Forget? Why not negative?
* Read: [Neural Net Language Model](http://machinelearning.wustl.edu/mlpapers/paper_files/BengioDVJ03.pdf)
* Read or Skim: [How the backpropagation algorithm works](http://neuralnetworksanddeeplearning.com/chap2.html)
* Optional: [Tensorflow LSTM Language Model Tutorial](https://www.tensorflow.org/versions/master/tutorials/recurrent/index.html#recurrent-neural-networks)
* Also: [Tensorflow Playground](http://playground.tensorflow.org/)
* Neural Net LMs
* Word embeddings
* Hierarchical softmax
* State of the art: Recurrent Neural Nets

**Week 5 [Basics of Text Processing]**

* Q5.5: D(i,j) = D(1,i) D(1,j)? Why 1?
* Q5.5.1: 0 cost for substitution?
* Q56: del probability? Out of what kind of corpus? Or type database?
* Q5.8: Types vs Tokens for Chinese?
* Q5.9: last quiz: speech recognition. Longer gap…? “Blanks”
* JOKE: should Asynch be more NLP oriented on the grading?
* Skim: NLTK book chapter 3 (processing raw text)
* Skim: [Peter Norvig’s chapter on ngrams](http://norvig.com/ngrams/ch14.pdf)
* Read: [Sentence Boundary Detection and the Problem with the U.S.](http://www.dgillick.com/resource/sbd_naacl_2009.pdf)
* Edit distance for strings
* Tokenization
* Sentence splitting

**Week 6 [Information Retrieval]**

* Q6.8: log x= 0 vs. frequency TF, IDF = 0
* Q6.8: Quiz: one word IDF irrelavent.
* Q6.10: Mean Average Precision
* Building a Search Index
* Ranking
* TFIDF
* Leveraging clicks

**Week 7 [Part-of-Speech Tagging I]**

* Tag sets
* Most frequent tag baseline
* HMM/CRF models

**Week 8 [Part-of-Speech Tagging II]**

* Feature engineering
* Leveraging unlabeled data
* Low resource languages

**Week 9 [Dependency Parsing]**

* Read: NLTK book chapter 8 (analyzing sentence structure)
* Dependency trees
* Transition-based parsing: Arc-standard, arc-eager
* Graph based parsing: Eisner Algorithm, Chu-Liu-Edmonds

**Week 10 [Constituency Parsing]**

* Context-free grammars
* CYK algorithm
* Probabilistic context-free grammar
* Lexicalized grammars, split-merge and EM

**Week 11 [Information Extraction]**

* Syntax -> Semantics
* Mention chunking
* Type tagging
* Coreference resolution

**Week 12 [Coreference Resolution]**

* Mentions and Entities
* Mention pair classification
* Clustering algorithms
* Rule-based systems
* Statistical systems

**Week 13 [Machine Translation]**

* Word Alignment
* Alignment heuristics
* Alignment with EM
* Phrase table translation

**Week 15 [Sentiment and Summarization]**

* Classification from reviews
* Sentence selection criteria
* Sentence compression
* Multi-document summarization