LING 300 - Topics in Linguistics:
Introduction to Programming and Text Processing for Linguists

Week 9

Python for Text 2 (and Beyond)

Roadmap for This Week

Monday

- Assignment 6 Notes
- Content:
  - Dependency Parsing
  - WordNet
  - Word Vectors
- Final Assignment

Wednesday

- Assignment 6 Notes
- Final Self-Evaluation
- Content:
  - Classification
- Where To Go From Here
- Breakout Rooms / OH (as time allows)

More Notes from Assignment 6

- Run POS taggers (and other models) on full sentences - What tag is “run” if we have:
  - Just “run”
    - Verb
  - “I went on a run”
    - Noun

More Notes from Assignment 6

- Careful with negative indexing!
- In `left_adjectives`:

```
for idx, token in enumerate(doc):
    if token.text == target_word and
        doc[idx - 1].tag_ == 'JJ':
        adj_counts[doc[idx - 1].text] += 1
```

Classification is the task of assigning labels

Which is spam?

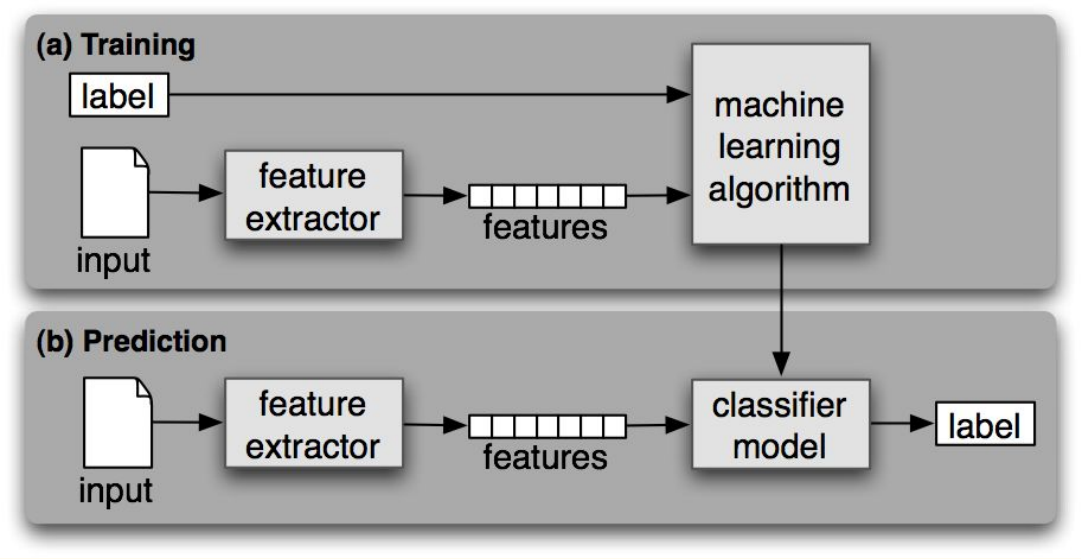
Congrats Andy Spellman!!!! You have won the sweepstakes!!! Click here to receive your FREE $50 Costco gift card!

Andy Spellman,

Thank you for your purchase of a $50 Costco giftcard. Your order details are listed below.

Classification is the task of assigning labels

- Use known input-label pairs to train an algorithm to decide which category a previously unseen input belongs to



https://www.nltk.org/book/ch06.html

Features are leveraged to make predictions

- Features can take many forms:
  - Counts of particular words
  - Counts of *n*-grams
    - multi-word phrases of length *n*:
      e.g. trigrams are three-word phrases (“so it goes”)
  - Numerical values (e.g., average concreteness)
  - Word vector dimensions

- Each is part of a mathematical representation of a document

Features are leveraged to make predictions

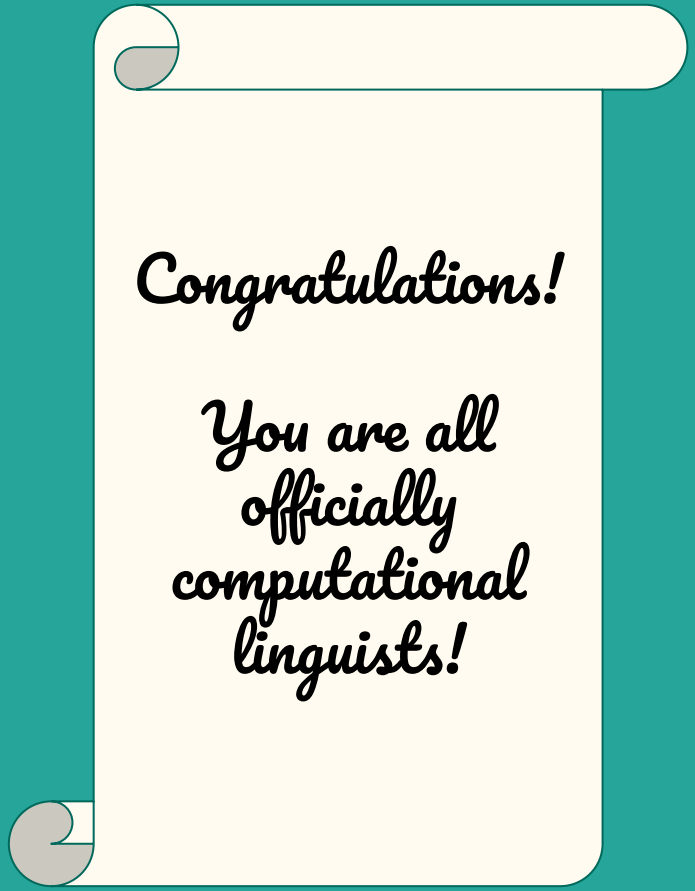
- "Learning" is most frequently the process of assigning numerical weights to each feature

NLTK movie review classification example:

```
>>> print(nltk.classify.accuracy(classifier, test_set)) ❶
0.81
>>> classifier.show_most_informative_features(5) ❷
Most Informative Features
   contains(outstanding) = True              pos : neg    =     11.1 : 1.0
        contains(seagal) = True              neg : pos    =      7.7 : 1.0
   contains(wonderfully) = True              pos : neg    =      6.8 : 1.0
         contains(damon) = True              pos : neg    =      5.9 : 1.0
        contains(wasted) = True              neg : pos    =      5.8 : 1.0
```

https://www.nltk.org/book/ch06.html

Where To Go From Here



Programming is very useful

- The skills you've learned are broadly applicable to linguistic and non-linguistic applications

- Try out your new computational tools and thinking in other parts of your life!

Other things you are now well-equipped to start learning

- Version control (git, see these lectures)
- Data science (see e.g. pandas and numpy)
- Machine learning (see e.g. scikit-learn)
- Web scraping (see e.g. BeautifulSoup)
- Dynamic web programming (see e.g. Flask or Django)
- App development (see e.g. Kivy)
- Game programming (see e.g. pygame or Godot)

Natural Language Processing (NLP) and Computational Linguistics (CL)

- NLP = more engineering, everything is a “task”, focus on system performance

- CL = computational social science, using and developing NLP tools for social, linguistic, humanistic questions

- No need, of course, to strictly pick a camp!

AI and Deep Learning

- Modern “neural networks” - I recommend this book: https://d2l.ai/

- and these more advanced lectures (Stanford CS224N): https://www.youtube.com/playlist?list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z

Interacting with me!

- *Next Quarter*

  Ling 334 - Introduction to Computational Linguistics
  - Covering fundamental algorithms and problems
  - This class is sufficient background
  - But expect some trickiness!
    - More independent programming projects, more math

- Always interested to chat about research projects etc!

Thank you!

It’s been a privilege and a joy to teach this class.