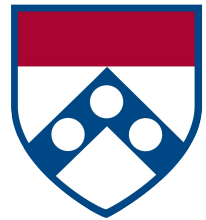
Wharton PhD Tech Camp

Session 7

Alex Miller

Ph.D. Student, Information Systems

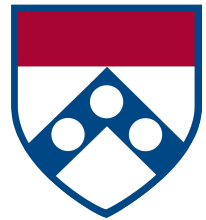
Wharton, OID



Goals for today

- Working with raw text data
- Natural language tools in Python
- Lab Exercise

Text Data



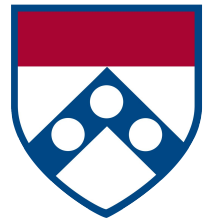
Introduction

- Where does text data come from?
  - The internet!
  - Open-ended survey responses
- Challenges of working with text
  - Very messy
  - “High dimensional”
- But there is a lot of rich information contained in unstructured text

Tools for working with text

- NLTK
  - Python’s “natural language toolkit”
- RegEx
  - For cleaning or basic feature extraction
  - Good at determining, “does this text contain X”
- Machine Learning
  - Will cover next lecture

Text Analysis



Terminology

- **Corpus**
  - the entire body of text you are working with
- **Documents**
  - individual entries in your corpus
  - Could be tweets, reviews, or novels
- **Vocabulary**
  - The set of unique words in your entire corpus
- **Tokens**
  - Individual “words” in a given document

Text Analysis Workflow

- Organize data into a set of documents
- Pre-process
  - Clean your data
  - Tokenization and Stemming
  - Part-of-speech tagging
- Analysis
  - Extracting information from your text:
    - Factual information
    - Subject of document
    - Sentiment… Many more

Cleaning Text Data

- Computers don’t know that:
  - “Dog” is the same thing as “dog”
  - “Yes” is the same thing as Yes
  - “Danny’s” should count the same as “Danny”
- This means we need to perform basic **“sanitizing”** operations before doing real analysis
- In some situations, you may want to go a step further:
  - **Stemming!**
  - Pares words down to their raw “stem”
  - “Loving”, “Love”, and “Lovingly” would all get stemmed to “Lov*”

Resources

- Lecture notebook
  - https://github.com/alexmill/techcamp_2017/blob/master/session7/session7_notebook.ipynb
  - Movies dataset walkthrough
- Exercises
  - https://github.com/alexmill/techcamp_2017/blob/master/session7/exercises.md