slalom

Topic Modeling

Leveraging Machine Learning to identify underlying topics in a document.

Anny Truong – June 14, 2019

Today's Agenda

- Machine Learning with Text Data
- Text Preprocessing
- Topic Modeling using LDA
- Demo

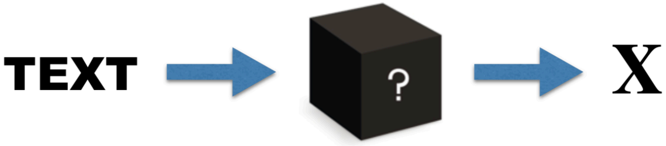
Machine Learning with Text Data

Use Cases!

- Text Classification, Clustering, Regression (scoring)
- Machine Translation
- Speech Processing (Recognition, Text-to-Speech, etc)
- Sentiment Analysis
- Topic Identification
- Document Recommendation
- ...And much more!

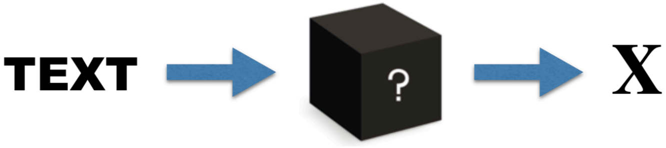
Machine Learning with Text Data

- Magical machine that can take raw text input and output feature vectors
  - With text, the rows in our X matrix will be documents



Machine Learning with Text Data

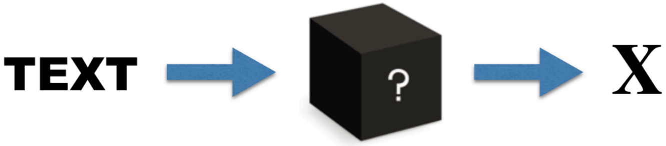
- Magical machine that can take raw text input and output feature vectors
  - With text, the rows in our X matrix will be documents



- Is it really magical?!

Machine Learning with Text Data

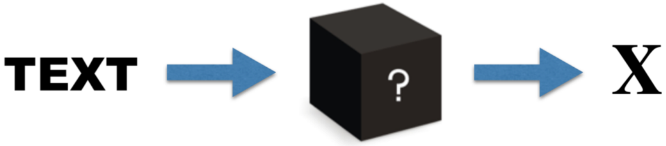
- Magical machine that can take raw text input and output feature vectors
  - With text, the rows in our X matrix will be documents



- Is it really magical?!

Machine Learning with Text Data

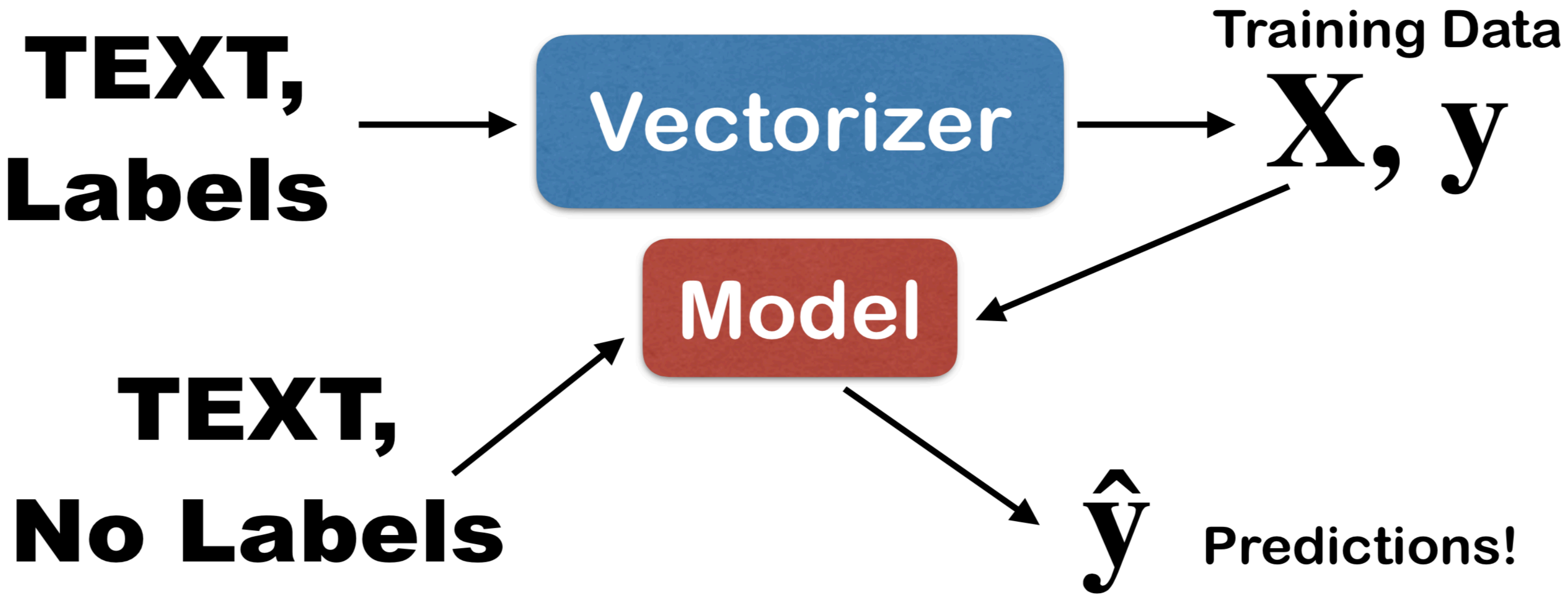
- Magical machine that can take raw text input and output feature vectors
  - With text, the rows in our X matrix will be documents



- Is it really magical?!

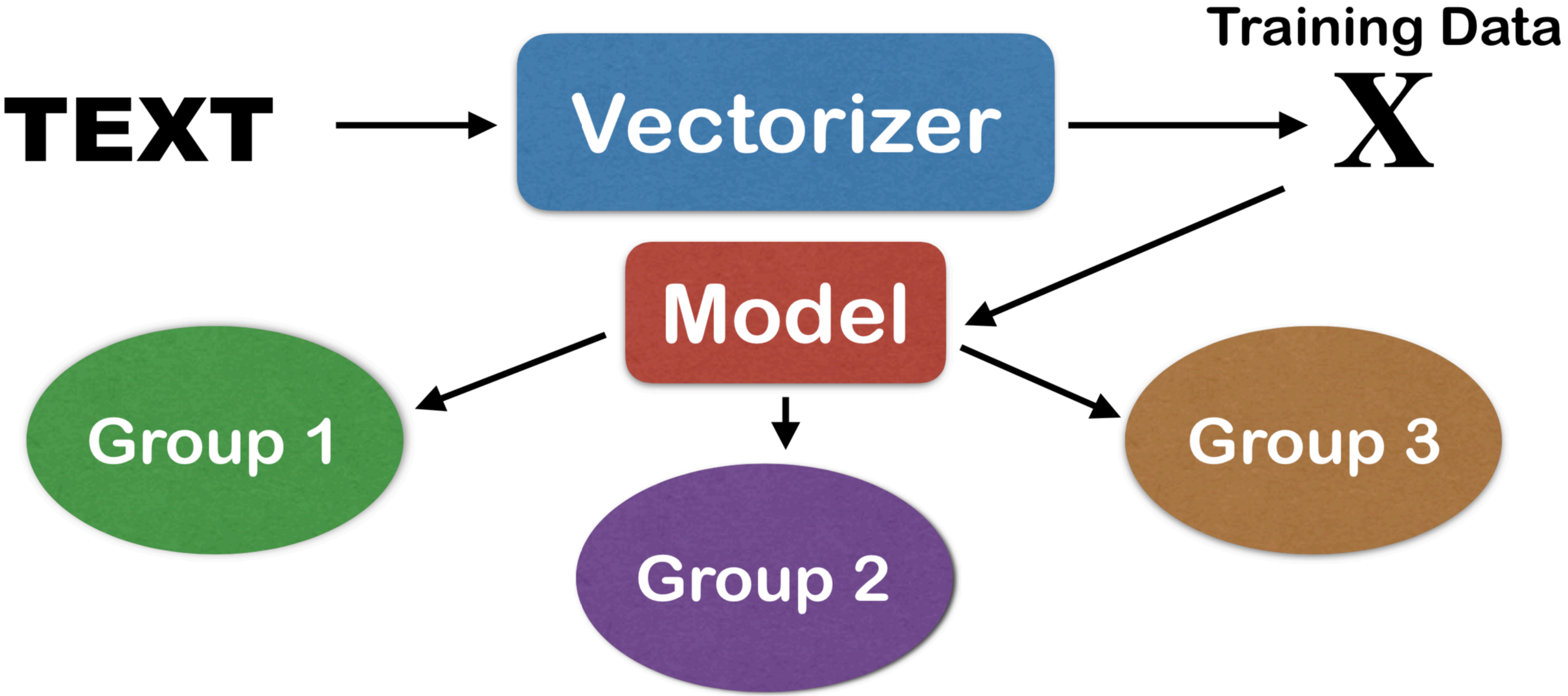
Machine Learning with Text Data

Supervised Learning



Machine Learning with Text Data

Unsupervised Learning



Vector Space Models

- Models that map raw text documents into vector spaces for mathematical comparison
  - These vectors usually extract some form of meaning
- Input: Corpus of raw text documents
- Output: Vectors for text documents (and usually terms)
- How do VSM work?
  - Start with raw text
  - Reduce the space
    - Using dimensionality reduction (like SVD, NMF)
    - Using neural networks
    - Using probabilistic models (LDA!)
  - Once we have ”semantic”, or meaning, vectors
    - We can do all sorts of further Machine Learning!

Step 1: Text Preprocessing

- Text is unstructured data
  - A lot of noise present
- Text preprocessing to remove noise and standardize it for analysis
  - Noise Removal
  - Lexicon Normalization
  - Object Standardization

Text Preprocessing: Noise Removal

Stopwords

- Words with little to semantic value, so we usually ignore them
- Can be domain specific
- Reduces complexity without loss of information

Anny eats the apples.
<n>Anny<n> <v>eat<v> <n>apple<n>.

Removing Punctuation

- Doesn’t add semantic value

Anny eats the apples.
<n>Anny<n> <v>eat<v> <n>apple<n>

Lowercasing

- Makes case insensitive without loss of semantic value

Anny eats the apples.
<n>anny<n> <v>eat<v> <n>apple<n>

Text Preprocessing: Lexicon Normalization

Tokenization

- Breaking up text into words, phrases, symbols, or other meaningful elements called **tokens**
- **Tokens** become input for further ML processing and allows us to put text information into data vectors

Anny eats the apples.
<t>Anny<t> <t>eats<t> <t>the<t> <t>apples<t>.

Stemming

- Reducing words to their root form (verb forms, plurals, etc)
- Generally, the “semantic content” is in the root form

Anny eats the apples.
<t>Anny<t> <t>eat<t> <t>the<t> <t>apple<t>.

POS-tagging

- Tagging the parts of speech for a sentence

Anny eats the apples.
<n>Anny<n> <v>eat<v> <t>the<t> <n>apple<n>.

Text Preprocessing: Object Standardization

Words or phrases which are not present in any standard lexical dictionaries

Acronyms

- **RT**: retweet(?), roundtrip(?)

Anny visits MTV company.
lookup_dict = {'rt':'roundtrip', 'MTV':'Mountain View', "..."}

Colloquial slangs

- Pop vs. soda vs. Coke
  - "Can I get a Coke?" "Sure, which kind?" "Dr.Pepper, please!"

Anny visits MTV company.
lookup_dict = { 'MTV company':'Google', "..."}

Misspellings

- Abcense, absance

Anny has no ragrets visiting MTV company.
lookup_dict = { 'ragret':'regret', "..."}

Step 2: Topic Modeling

- **Topic modeling** is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents.

Topic Modeling

Methods

- Latent Dirichlet Allocation (LDA)
  - Most common!
- Others
  - Hierarchical latent tree analysis (HLTA)
  - Pachinko Allocation
  - Probabilistic Latent Semantic Analysis (PLSA)

Latent Dirichlet Allocation (LDA)

- **Latent Dirichlet Allocation (LDA)** is an example of topic model
  - Classify text
  - Builds a topic per document model
  - Modeled as Dirichlet distributions

Latent Dirichlet Allocation (LDA)

- Suppose you have the following set of sentences:
  - Anny likes to eat broccoli and bananas.
  - Anny ate a banana and toast for breakfast.
  - Hedgehogs and kittens are cute.
  - My sister adopted a kitten yesterday.

Latent Dirichlet Allocation (LDA)

- Suppose you have the following set of sentences:
  - Anny likes to eat broccoli and bananas.
  - Anny ate a banana and toast for breakfast.
  - Hedgehogs and kittens are cute.
  - My sister adopted a kitten yesterday.

Topic A

Topic B

Latent Dirichlet Allocation (LDA)

- Suppose you have the following set of sentences:
  - Anny likes to eat broccoli and bananas.
  - Anny ate a banana and toast for breakfast.
  - Hedgehogs and kittens are cute.
  - My sister adopted a kitten yesterday.
  - Look at this cute hedgehog munching on a piece of broccoli.

Topic A

Topic B

Latent Dirichlet Allocation (LDA)

- Suppose you have the following set of sentences:
  - Anny likes to eat broccoli and bananas.
  - Anny ate a banana and toast for breakfast.
  - Hedgehogs and kittens are cute.
  - My sister adopted a kitten yesterday.
  - Look at this cute hedgehog munching on a piece of broccoli.

Topic A

Topic B

60% Topic A, 40% Topic B

Latent Dirichlet Allocation (LDA)

Wait, but how?!

- LDA assumes this is how you write your document
  - First, you decide on the number of words
  - Choose a topic mixture
  - Generate each word by:
    - Picking a topic
    - Using the topic to generate the word itself
- Using these assumptions, LDA then tries to backtrack from the documents to find a set of topics

Latent Dirichlet Allocation (LDA)

Wait, but how?!

- LDA assumes this is how you write your document
  - First, you decide on the number of words
  - Choose a topic mixture
  - Generate each word by:
    - Picking a topic
    - Using the topic to generate the word itself
- Using these assumptions, LDA then tries to backtrack from the documents to find a set of topics

- Document A
  - Number of words = 5
  - Topic mixture: 50% food, 50% cute animals
  - Generate each word:
    - Pick topic: Food
      - Words: broccoli, bananas, eat
    - Pick topic: Cute Animals
      - Words: Hedgehogs, kitten, adorable

Latent Dirichlet Allocation (LDA)

Wait, but how?!

- Collapsed Gibbs sampling:
  - For each word *w* in document *d*, compute:
    - P (topic *t* | document *d*) = the proportion of words in document *d* that are currently assigned to topic *t*
    - P (word *w* | topic *t*) = the proportion of assignments to topic *t* over all documents that come from this word *w*
  - Reassign *w* a new topic, where we choose topic *t* with probability
    - P (topic *t* | document *d*) * P (word *w* | topic *t*) = probability that topic *t* generated word *w*
  - Rinse and repeat until you get to a steady state of assignments
  - Use the assignment to estimate the topic mixtures of each document

Example!

- Scenario: Anny just moved to San Francisco and is a motorcycle and MMA enthusiast
  - Caveat: Anny is introverted and hates asking people to find communities
- What to do?
  - Step 1: Scope out establishments (**documents**), making note of the people (**words**) in each establishment, find typical interest groups of each establishment (**topics**)
  - Step 2: Pick some number of categories to learn, and make guess as to why people hang out where they do
    - Nate goes to the gym wearing a gi… He probably has an interest in jiu-jitsu!
    - Bobby goes to the park with a stack of board games… He is probably meeting with his friends to play board games.
  - Step 3: Improve on your guesses
    - Make a new guess as to why Nate is at the gym and Bobby is at the park. What are the probabilities of these interests?
  - Step 4: Go through each place and person over and over again
    - The gym also has a lot of other people with gi, probability that Nate's interest in jiu-jitsu is very high!

Demo!