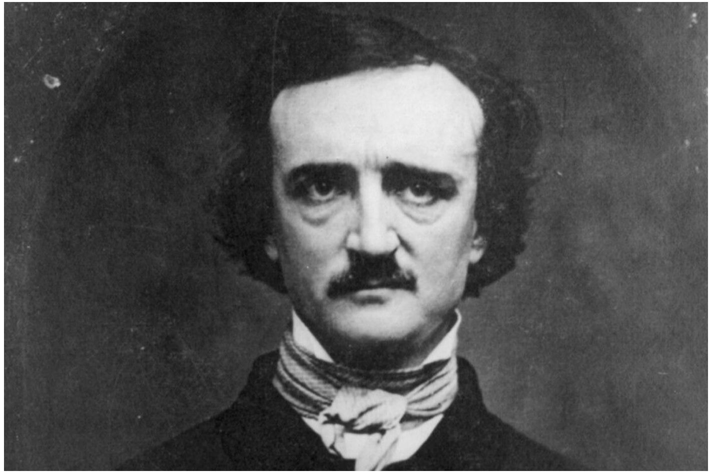
Text Classification with Horror Authors

Caleb Ho, Vickram Rajendran

Overview

- Given snippets of text, identify the author
    - Example: “It never once occurred to me that the fumbling might be a mere mistake.” - HPL
- 3 authors
    - Edgar Allan Poe
    - Mary Shelley
    - HP Lovecraft
- Kaggle competition

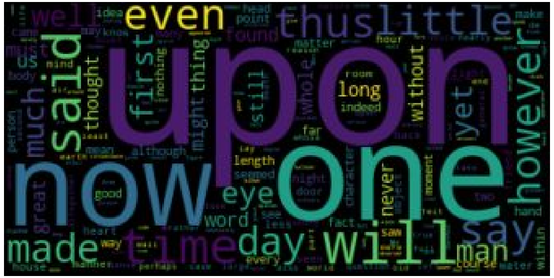


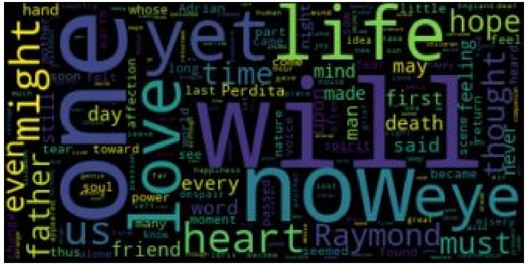
Background

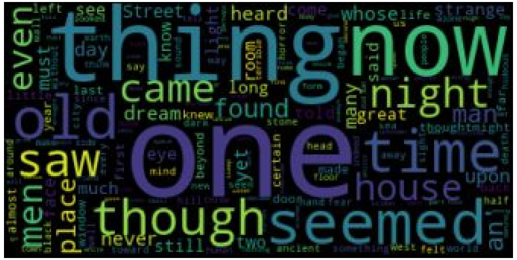
- Somewhat similar to what we’ve seen before…
  - Classification problem
  - Supervised Learning
- New twist - Semantics
  - Natural Language Processing (NLP)

Exploratory Data Analysis

- What “Metadata” do we need to know?

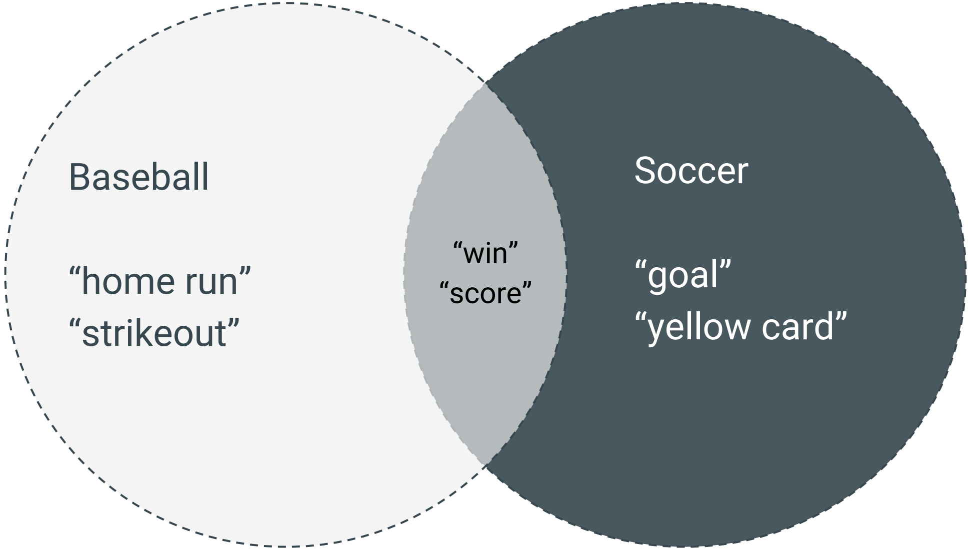






Topic Modeling

- Given a collection of documents, learn the topics that occur in the documents
- Topic Modeling is essentially Dimensionality Reduction.



Generative Latent Dirichlet Allocation

- Documents are a **mixture** of topics that randomly generate words according to some distribution of words in the topics.
- Pick a topic from a distribution, pick a word from that topic.
- Essentially an EM procedure to maximize the likelihood of your corpus occurring.

Latent Dirichlet Allocation: Pseudocode

1. Randomly assign each word to a topic for an initial distribution.
2. For each word
    a. For each topic
        i. In the sample sentence of the word, calculate proportion of words in this sentence in this topic
        ii. Calculate the proportion of sentences containing this word where we classified the word in this topic.
    b. Reassign the word a topic based on the distribution we just made.
3. Repeat a large amount of times.

Topics from LDA

```
Topic #15:life man friend mind nature time idea human taken year great better father say death adrian w
onder state power existence affection knowledge passion felt result idris ill hope desire new point ord
er sorrow sister spirit world gave received regard saw
======================================================================

Topic #16:heart come love body look hope hand tale seen pain self lost lip thou fate place corpse labou
r living hard brow equal amidst descent forward listen marie frequent girl inquiry pause boy stage murd
erer struggle discover marsh friendship energy head
======================================================================

Topic #17:thought shall say hand kind head like mean general attention time took natural just instant a
partment turned human longer entered really stranger don murder bear second dare nose believed odd supp
ose glass affair proper sad exertion face finger suddenly patient
======================================================================

Topic #18:night way thing character event terror account minute began point home leave save uncle desig
n horse remain certainly strange added slowly felt noise quiet whateley anxiety horrible page slope unk
nown crossed urged caused difficulty finally louder brought steadily surprise ve
======================================================================
```

Non-negative Matrix Factorization

- Treats it mathematically
- Intuition: The word embedding (bag of words) is a matrix W, where each row is a sentence and each index is a word.
- What if we tried to factor this matrix?
- Idea: Find a “sentence x topic” matrix M and a “topic x word” matrix T such that MT = W.

- A ~ WH

- Tweet 1
- Tweet 2
- Tweet 3

Term-Tweet Matrix

	Word 1	Word 2	Word n
Tweet 1	1	0	2
Tweet 2	0	1	0
Tweet 3	0	1	1

Specify No Themes (k)

Features Matrix

	Word 1	Word 2	Word n
Theme 1	0.5	0	1
Theme 2	0	0.5	0

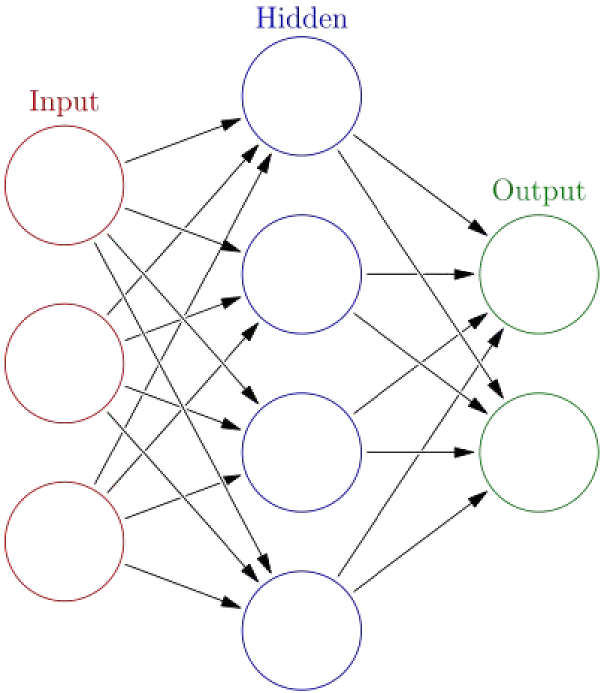
Weights Matrix

	Theme 1	Theme 2
Tweet 1	1	0
Tweet 2	0	1
Tweet 3	0	1

https://medium.com/mlreview/topic-modeling-with-scikit-learn-e80d33668730

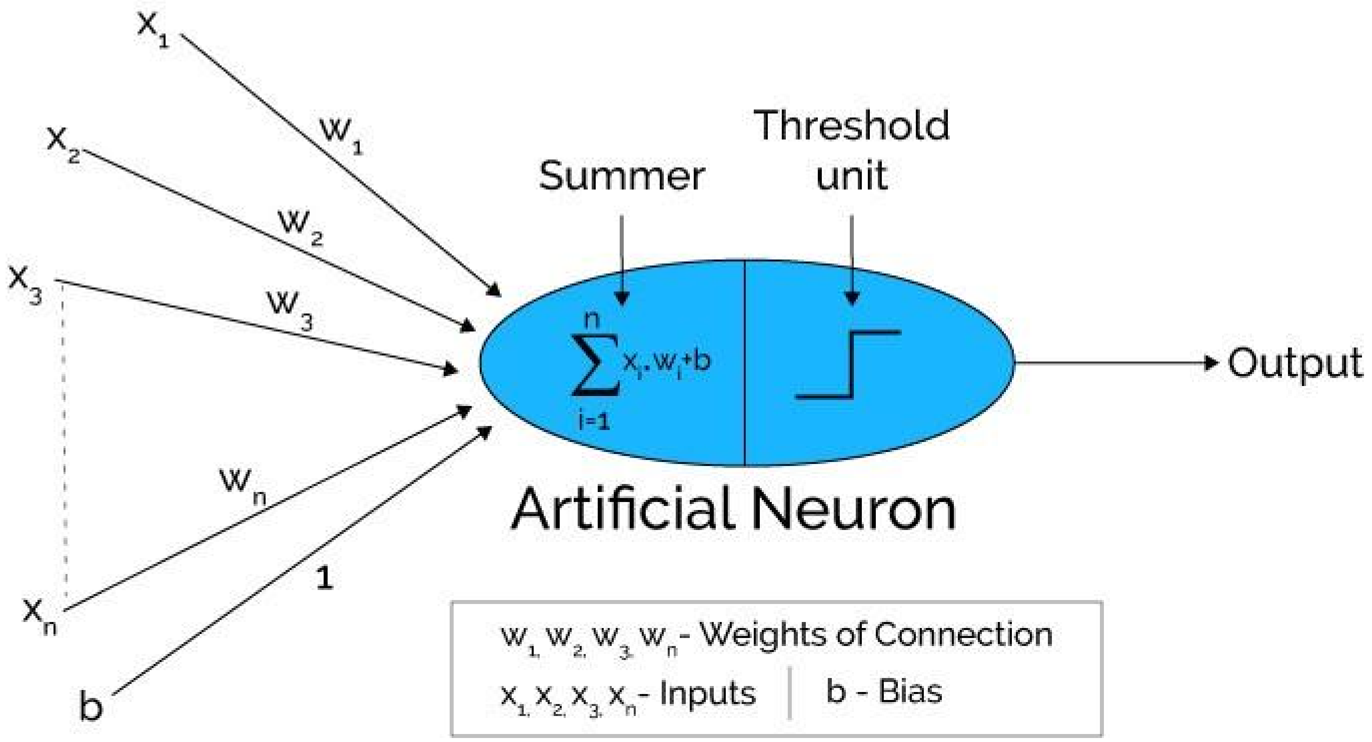
Neural Networks

- Weighted directed acyclic graph of artificial neurons
- Training using gradient descent and backpropagation
- Implemented using Keras



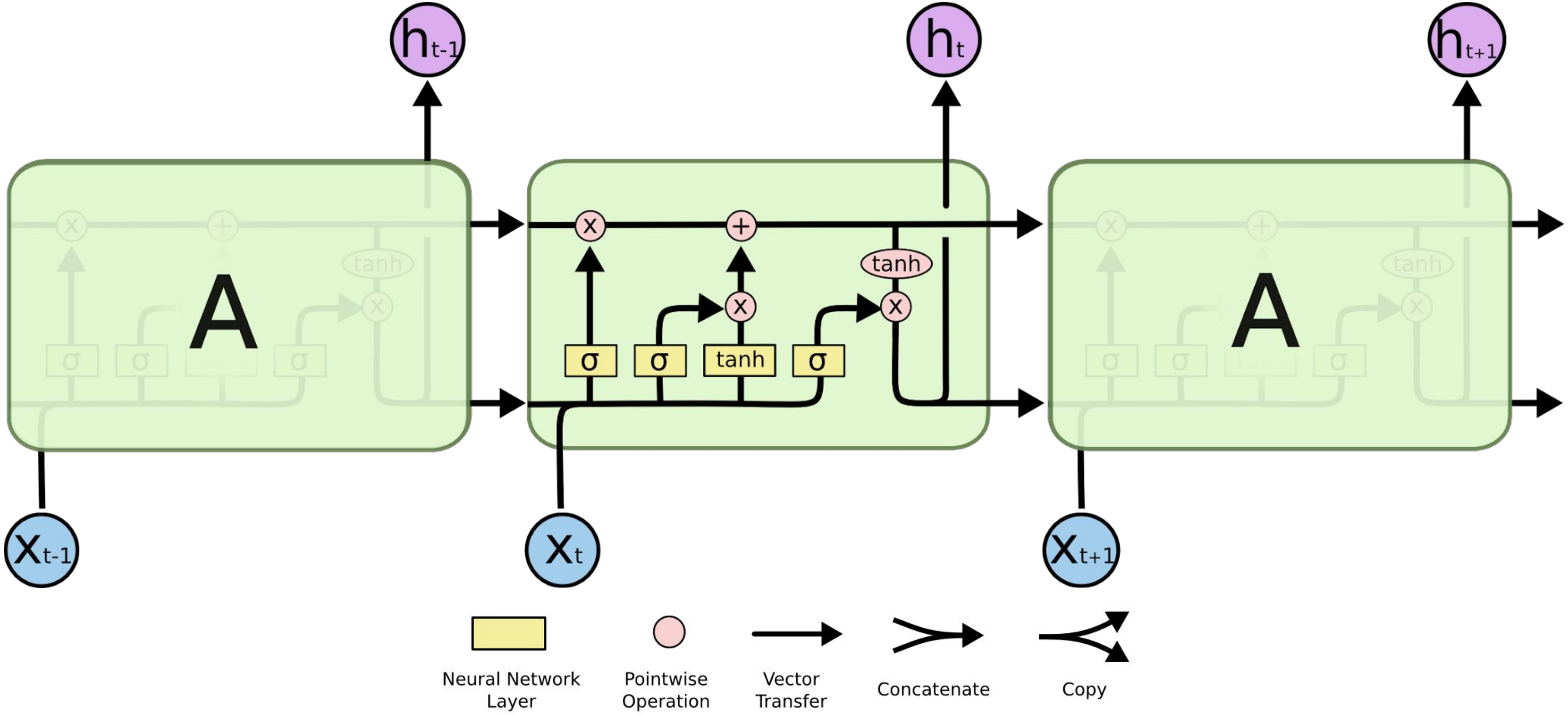
https://upload.wikimedia.org/wikipedia/commons/thumb/4/46/Colored_neural_network.svg/399px-Colored_neural_network.svg.png

Neural Networks



https://cdn-images-1.medium.com/max/1600/0*OHlzxsoDSEBXW0iI.jpg

Long-short term memory



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

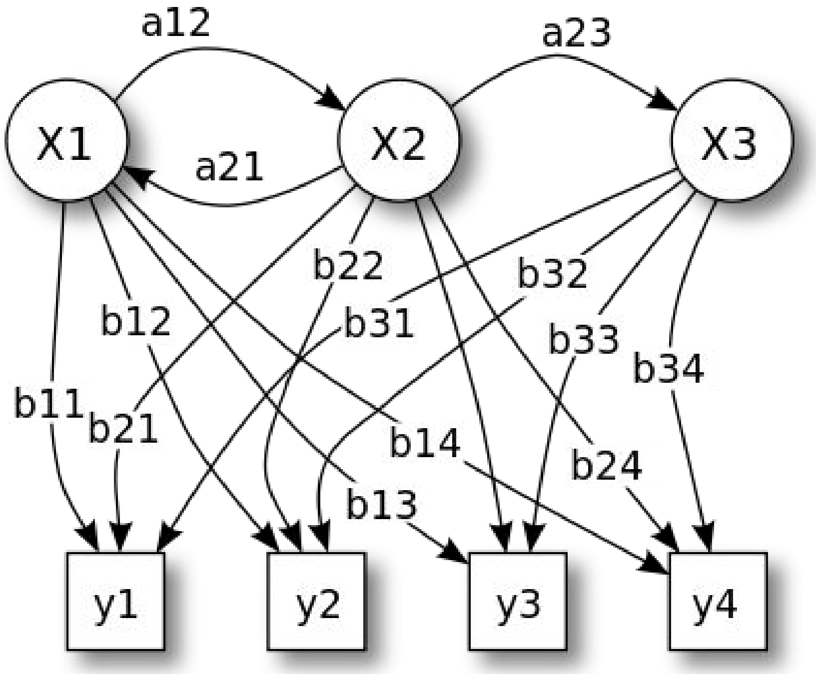
Network Topology

Embedding	LSTM	Dense	Output
Given a fixed-length sequence of integers, output a sequence (of the same length) of vectors.	Recurrent layer to learn sequence structure. Use dropout + recurrent dropout to control overfitting.	Fully connected layer.	Output layer of size 3.

Hidden Markov Models

- Graphical model
- Markov property: current hidden state x(t) depends only on x(t - 1); observed state y(t) depends only on x(t).
    - Can be generalized to higher orders
- Originally formulated for sequence learning, but can be adapted for sequence classification.
- Given a sequence of hidden states (words), what is the most likely sequence of observed states (authors)?
- Implemented using seqlearn

Hidden Markov Models



https://upload.wikimedia.org/wikipedia/commons/thumb/8/8a/HiddenMarkovModel.svg/500px-HiddenMarkovModel.svg.png

Log loss (a.k.a. categorical cross-entropy)

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

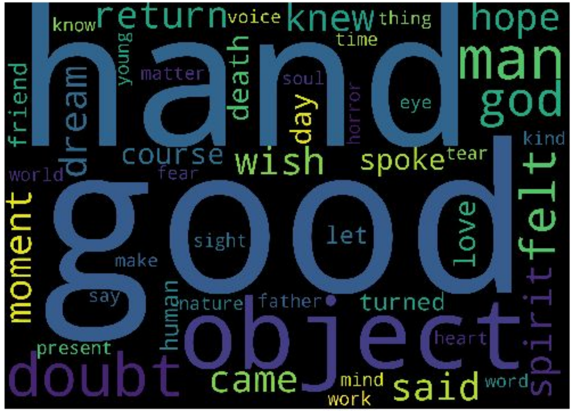
- Heavily penalize highly confident, but inaccurate predictions
- More nuanced than classification accuracy

Results

Algorithm	Accuracy	Log loss
Naive Bayes	0.823	0.474
NB + LDA	0.495	1.080
NB + NMF	0.556	6.615
LSTM	0.792	0.463
HMM	0.562	1.018

Analysis: Topic Models :(

- Similar topics
- Recall Exploratory Data Analysis - most used words were not in topics, just per word.
- How much distinction between the authors?
- Writing style rather than content?
- Log Loss Ratio - weird tuning on NMF?



Analysis: Sequential Models

- Why this works (ish)
- Markov Chains strictly worse than LSTM - 1 word memory?
- Some sequences could be ambiguous
- Likely performs better with more incongruous data, larger dataset.

Analysis: Naive Bayes

- Conditional independence assumption: Conditioned on author, appearance of a word does not change the probability of observing some other word
- Authors from same genre
- Poe and Shelley from similar time periods (early 1800s)
- Space of words has greater dimension than space of topics

Conclusions

- Naive Bayes is not very naive
- Topic modelling did not improve performance of Naive Bayes
- LSTM network achieves slightly lower accuracy than Naive Bayes, but slightly better log loss
- LSTM > Markov Chain
- Text classification of contemporaries who write about the same topic is particularly hard