# CS6120

Jiaji Huang

https://jiaji-huang.github.io

# About myself

- PhD in electrical engineering from Duke University
- Thesis: statistical signal processing and machine learning
- Worked at Baidu Research on NLP and Speech
- Now at AWS, on Large Language Models

# Motivating Task: Positive or Negative Reivew?

Roll over image to zoom in

Cyphis

★☆☆☆☆ **One Star is Too Much for This Product**

Reviewed in the United States on September 7, 2012

I don't know if this is a scam or if mine was broken, but it doesn't work and I am still getting abducted by UFO's on a regular basis.

Amazon Customer

★★★★★ **it works**

Reviewed in the United States on March 21, 2018

**Verified Purchase**

it works great

Marcie G.

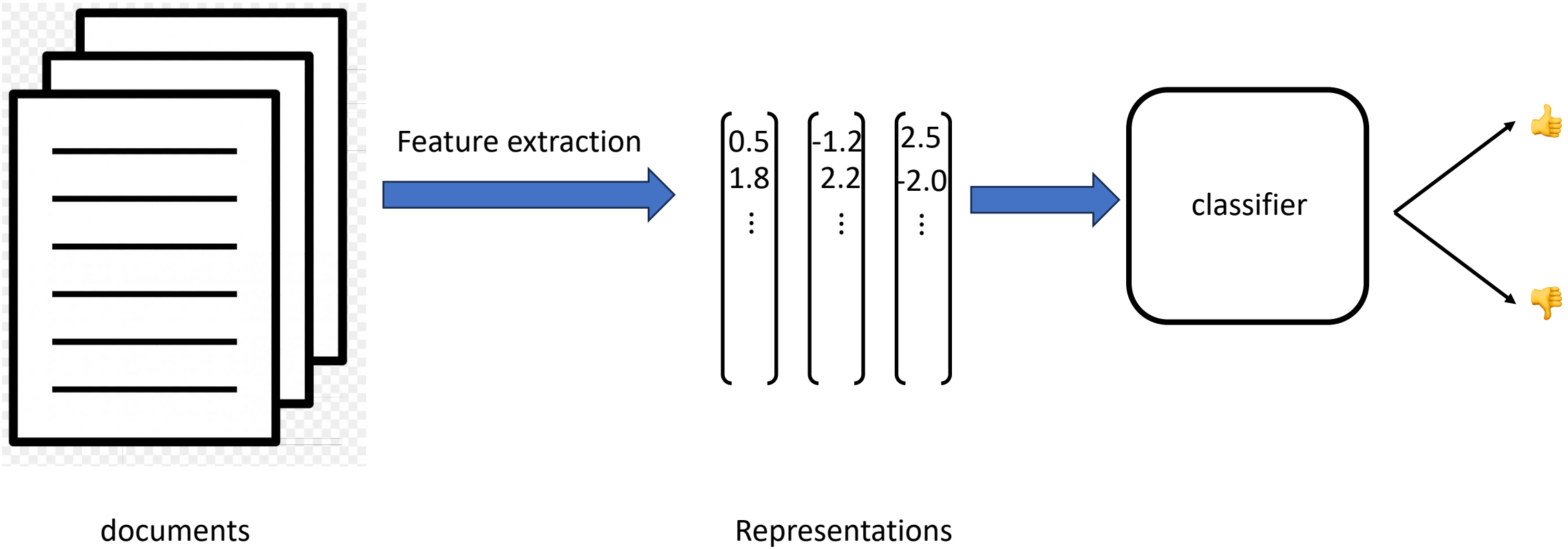
★★★★☆ **Freedom!**

Reviewed in the United States on December 12, 2013

Knowing when "they" are in the area this allows me to take off my tin foil hat a little more frequently. I've even peeked out the foil drapes occasionally, but still not that often.

I gave it 4 stars as I believe it could alert you sooner. I was almost teleported during one of my peeking episodes.

# A Machine Learning Approach



# Roadmap

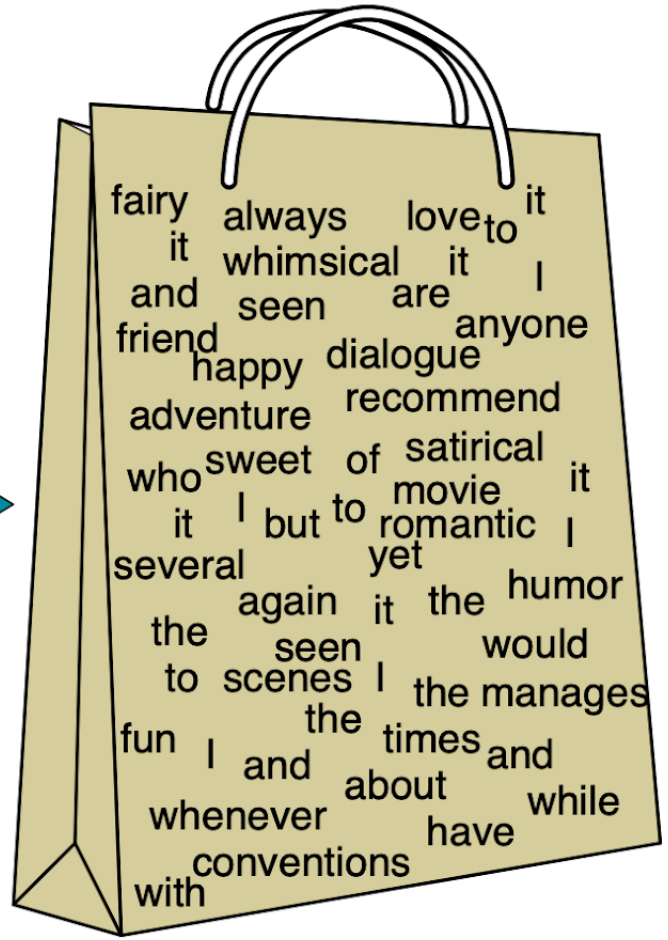
- Document representation
  - Bag of words
  - tf-idf
- Word representation
  - Word2vec and PMI
  - Applications
- Classifier
  - Naïve Bayesian classifier
  - Softmax classifier

# Roadmap

- Document representation
  - Bag of words
  - tf-idf
- Word representation
  - Word2vec and PMI
  - Applications
- Classifier
  - Naïve Bayesian classifier
  - Softmax classifier

# The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!




| | |
|---|---|
| it | 6 |
| I | 5 |
| the | 4 |
| to | 3 |
| and | 3 |
| seen | 2 |
| yet | 1 |
| would | 1 |
| whimsical | 1 |
| times | 1 |
| sweet | 1 |
| satirical | 1 |
| adventure | 1 |
| genre | 1 |
| fairy | 1 |
| humor | 1 |
| have | 1 |
| great | 1 |
| … | … |

https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf

# tf-idf

- Words have different importance, overlooked by simple count
- tf: term frequency (multiple ways to define)

$$tf_{t,d} = \frac{count(t,d)}{\sum_t count(t,d)}$$

$$tf_{t,d} = \log[1 + count(t,d)]$$

- idf: inverse document frequency

$$idf_t = \log\left(\frac{\#\ total\ docs}{\#docs\ that\ have\ term\ t}\right)$$

- tf-idf for word $t$ in document $d$: $tf_{t,d} \times idf_t$

# Example

# Brain storming

- Ways to improve bag of words representation?
  - Word order information: counts of tuple of n words (n-gram)
  - For “new” word: sub-word counts
  - More efficient: dimension reduction, PCA etc…
  - …

# Roadmap

- Document representation
  - Bag of words
  - tf-idf
- Word representation
  - Word2vec and PMI
  - Applications
- Classifier
  - Naïve Bayesian classifier
  - Softmax classifier

# word2vec

## Efficient Estimation of Word Representations in Vector Space

by T Mikolov · 2013 · Cited by 37327 — Download a PDF of the **paper** titled Efficient Estimation of Word Representations in Vector Space, by Tomas Mikolov and 3 other authors.

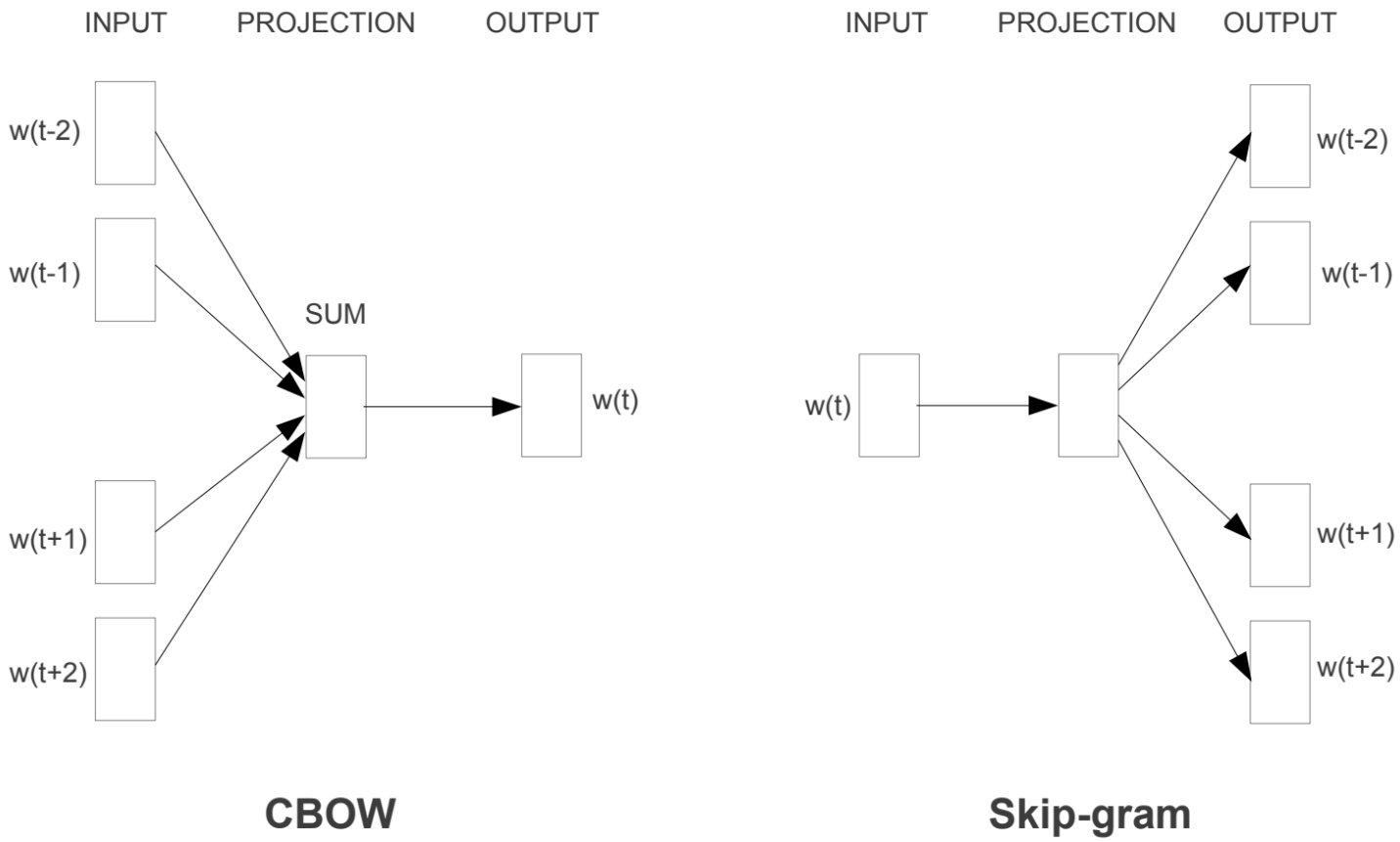



Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

# How it works

- Word2vec software
- Show a demo …

# PMI

Word Association Norms, Mutual Information, and ...

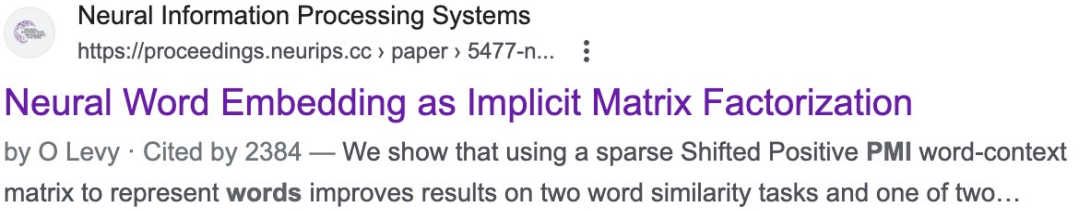
by K Church · 1990 · Cited by 6485 — Anthology ID: J90-1003; Volume: Computational Linguistics, Volume 16, Number 1, March 1990; Month: Year: 1990; Address: Venue: CL; SIG:…

- The PMI (point wise mutual information) between two words x, y

$$\operatorname{pmi}(x;y) \equiv \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{p(x|y)}{p(x)} = \log_2 \frac{p(y|x)}{p(y)}$$

- Practically:
  How associated are two words?

# Connections




- Do some math on whiteboard …

# Generalize: *2vec

- Create pairwise associations between entities, call it matrix $M$
- Then factorize $M$!

node2vec: Scalable Feature Learning for Networks

by A Grover · 2016 · Cited by 10317 — Here we propose **node2vec**, an algorithmic framework for learning continuous feature representations for nodes in networks. In **node2vec** ...

[1902.03545] Task2Vec: Task Embedding for Meta-Learning

by A Achille · 2019 · Cited by 250 — Abstract: We introduce a method to provide vectorial representations of visual classification tasks which can be used to reason about the ...

# Application: Word Analogy

# Application: Bilingual Lexicon Induction (BLI)

# More cools things about BLI …

- Unsupervised Machine Translation

# Connection with BERT

# Roadmap

- Document representation
  - Bag of words
  - tf-idf
- Word representation
  - Word2vec and PMI
  - Applications
- Classifier
  - Naïve Bayesian classifier
  - Softmax classifier

# Naïve Bayesian Classifier

- Switching to slides last year …

# Softmax classifer

# Brain storming …

- Strength and weakness of Naïve Bayesian v.s. softmax