
H2O
WORLD
2017

Megan Kurka

Customer Data Scientist

megan@h2o.ai

H2O WORLD 2017

NLP with H2O

Supervised Learning with Unstructured Text Data

H2O WORLD 2017

Our Use Case

The Data

The Amazon Fine Food Reviews dataset consists of 568,454 food reviews Amazon users left up to October 2012

- J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews

Column	Example
Product ID	B006K2ZZ7K
User ID	A1UQRSCLF8GW1T
Helpfulness Numerator	1
Helpfulness Denominator	1
Score	5
Time	1350777600
Summary	Great taffy
Text	*“Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.”*

The Goal

Predict whether a food product has a good rating based on the reviews

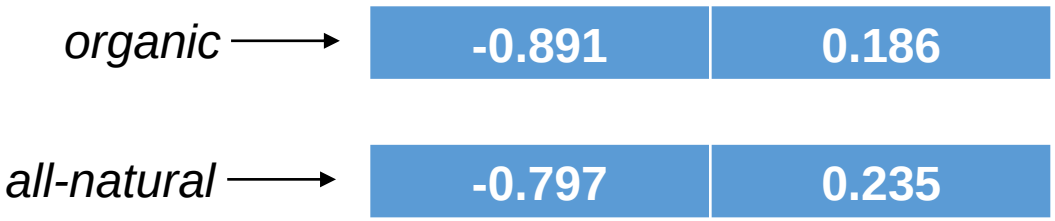
“Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.”

H2O WORLD 2017

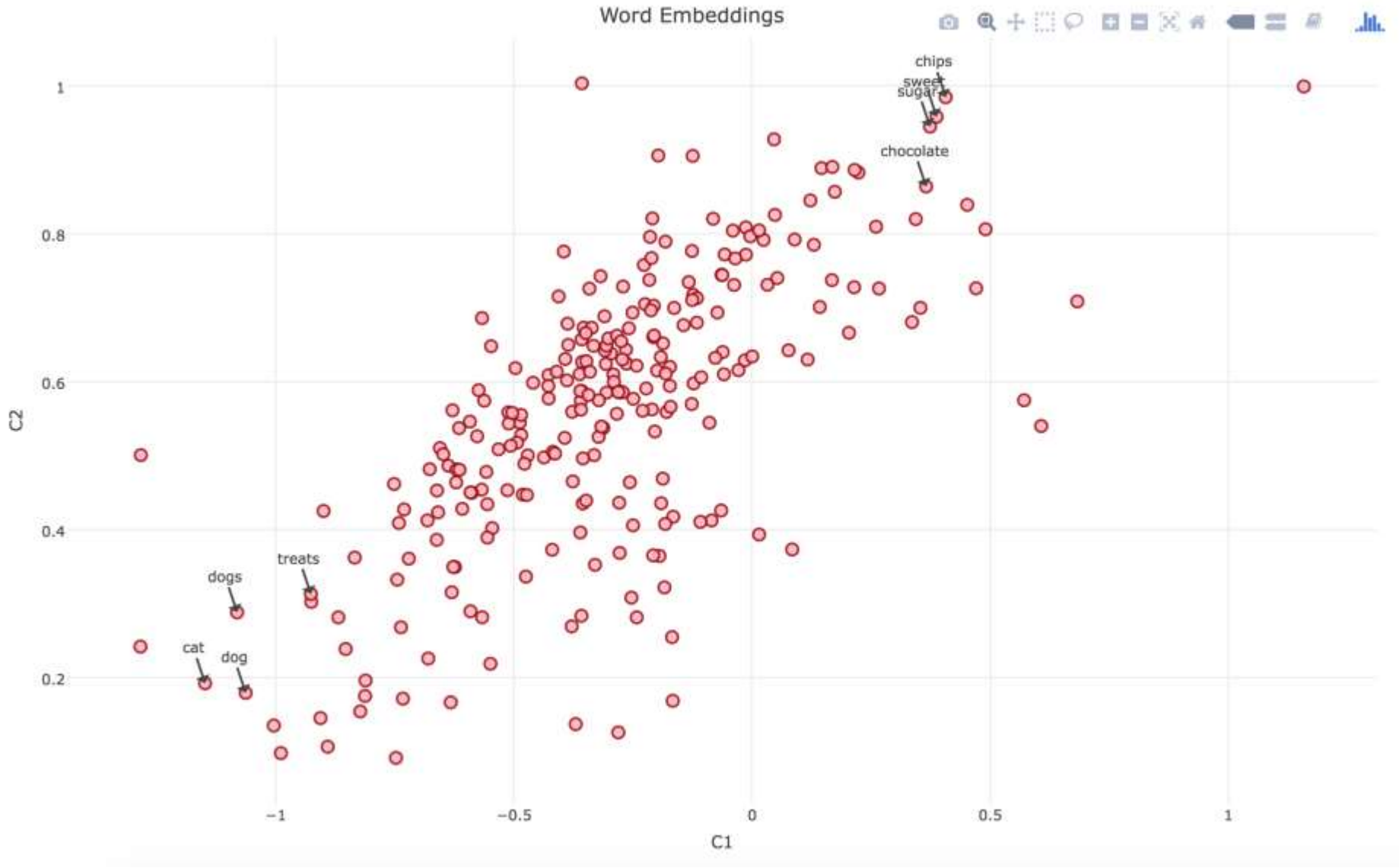
Natural Language Processing

Word Embeddings

- What?
  - Mapping of words to vectors from a high dimensional space (100-1000)
- Why?
  - Embeddings capture the meaning of the word
  - Semantically similar words are close to each other



Word Embeddings



Word2Vec Algorithm

How do we use a neural network to capture the semantic meaning of words?

- Frame the problem as a supervised learning problem
  - Given an input word predict the neighboring words

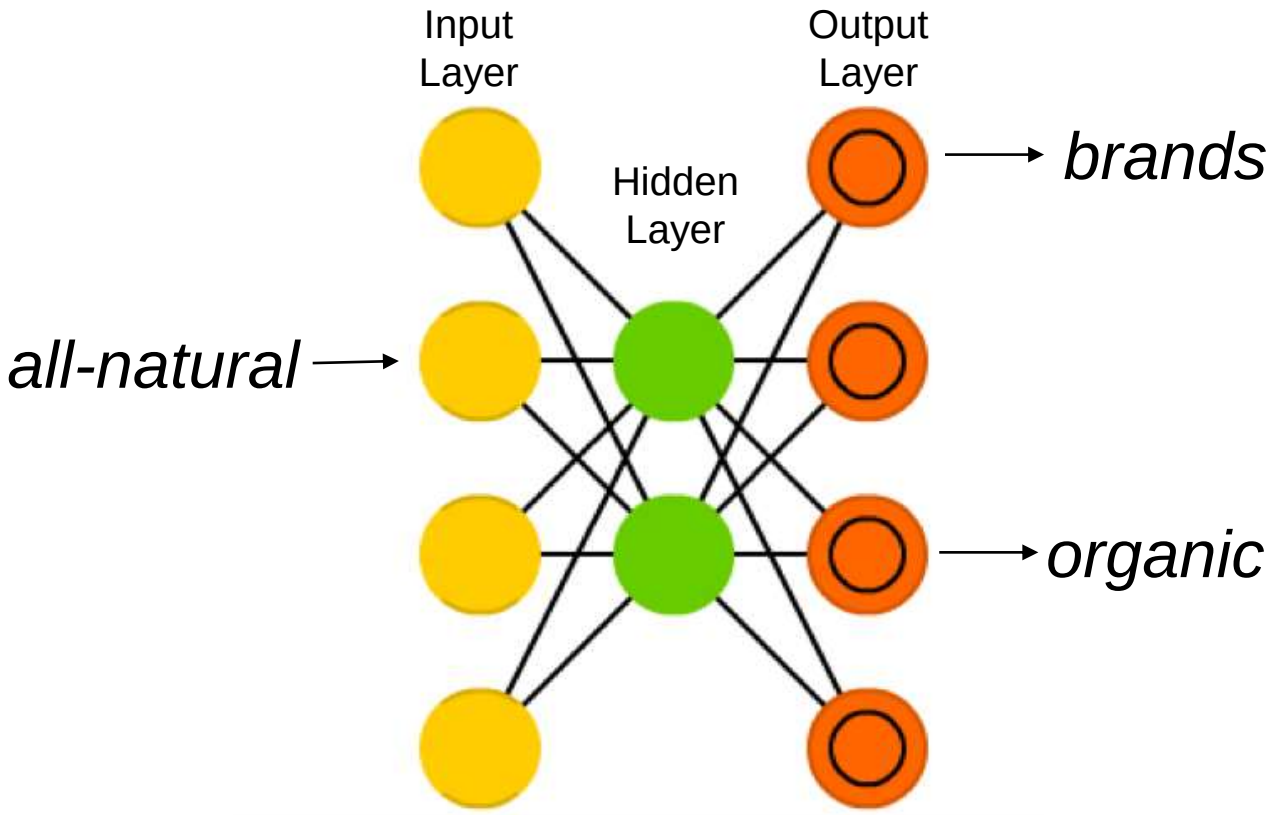
*“It's even better than the **organic, all-natural brands** I have tried.”*

Given: *all-natural*

Predict: *organic, brands*

Word2Vec Algorithm

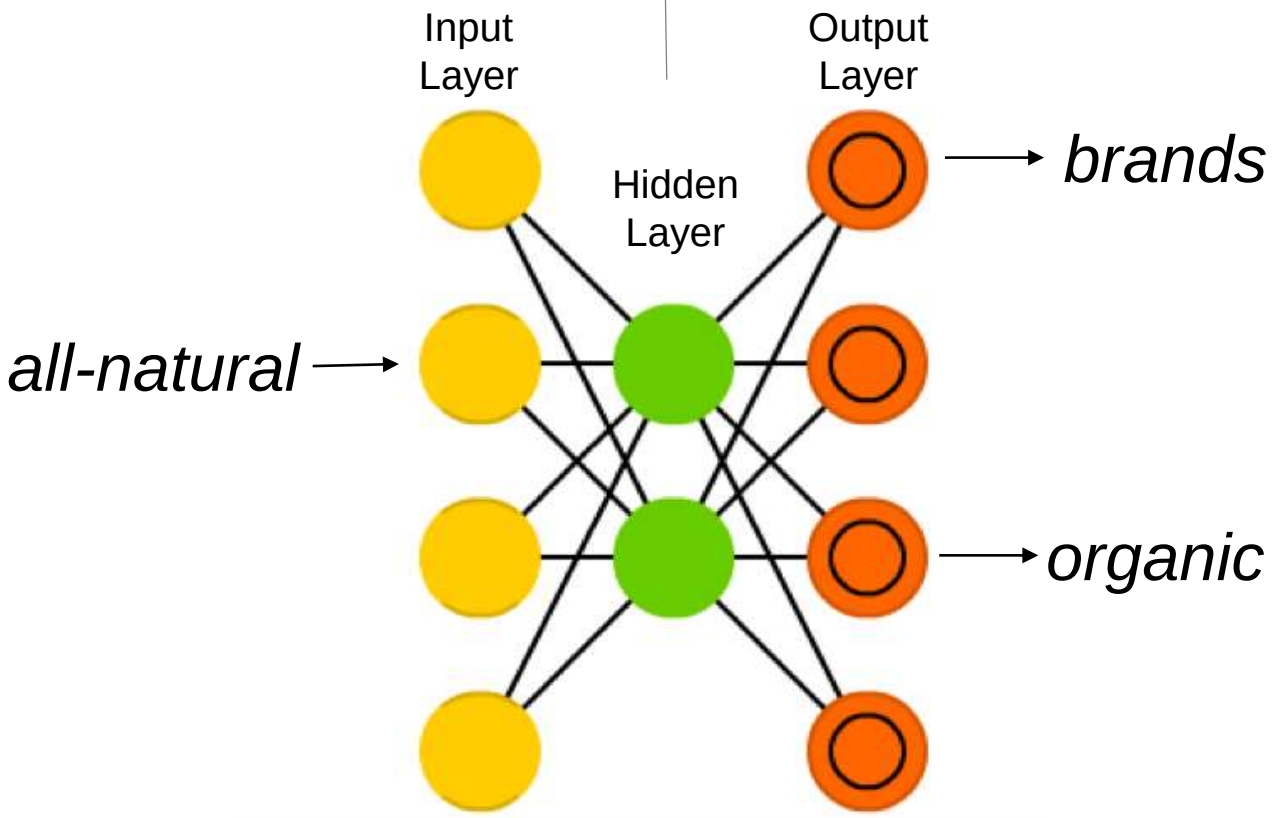
*"It's even better than the **organic, all-natural brands** I have tried."*



Word2Vec Algorithm

Matrix from Hidden Layer = Word Embeddings

Word	C1	C2
brands	0.647	0.235
all-natural	-0.797	0.235
organic	-0.891	0.186
tried	-0.751	0.409



Our Workflow

Use Case: Predict whether a food product has a good rating based on the review.

1. Tokenize Reviews
   - Break up reviews into separate words
   - Filter words: remove stop words like “the” and “if”
2. Train a Word2Vec Model
3. Use model to transform reviews to vectors
4. Train a supervised learning model to predict good rating