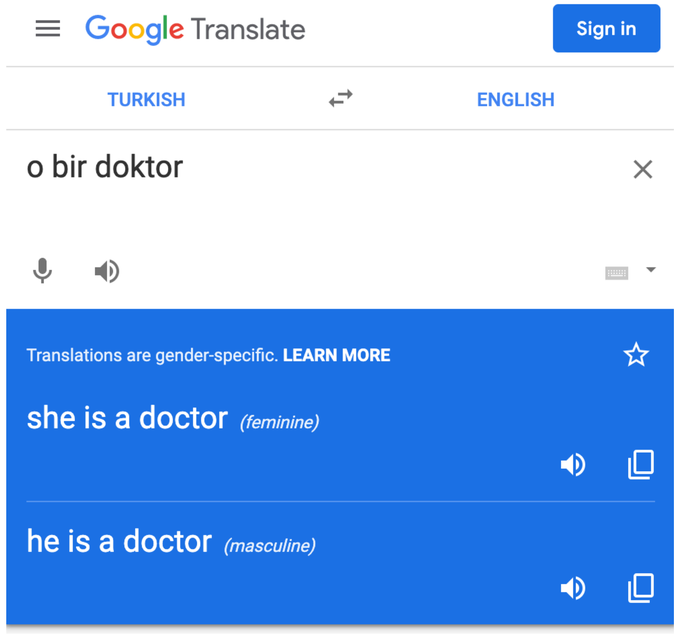
The Wonderful World of Embeddings

Franz Diebold, M.Sc. (TUM)

github.com/franzdiebold

linkedin.com/in/franz-diebold



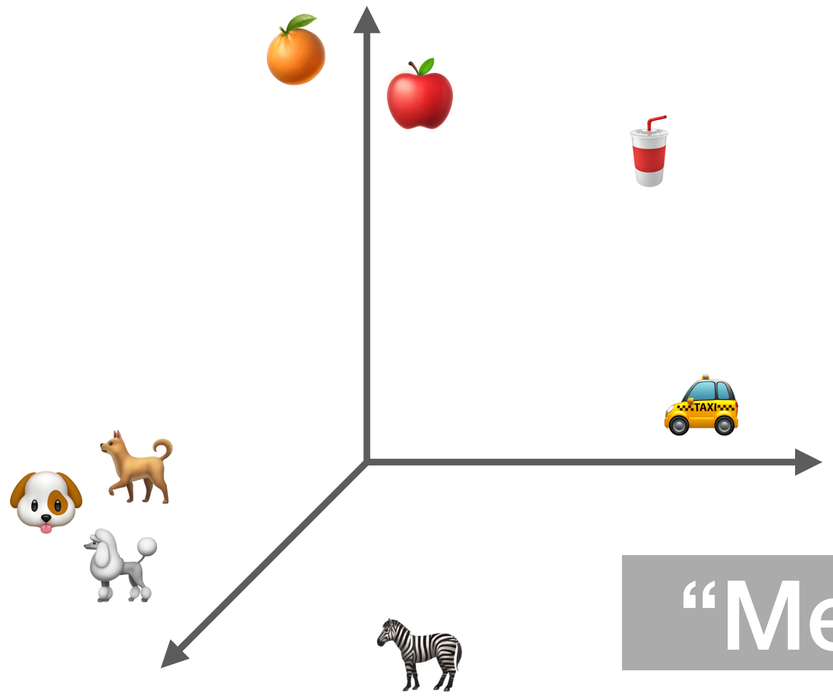
Embeddings are all over the place!

Word Embedding

Demo

Vocabulary V = {🍎, 🐶, 🐕, 🐩, 🍊, …, 🥤, 🚕, 🦓}

|V| = 100k



Embedding E: V → $\mathbb{R}^{50}$

“Meaning”

Properties of Word Embeddings

projector.tensorflow.org

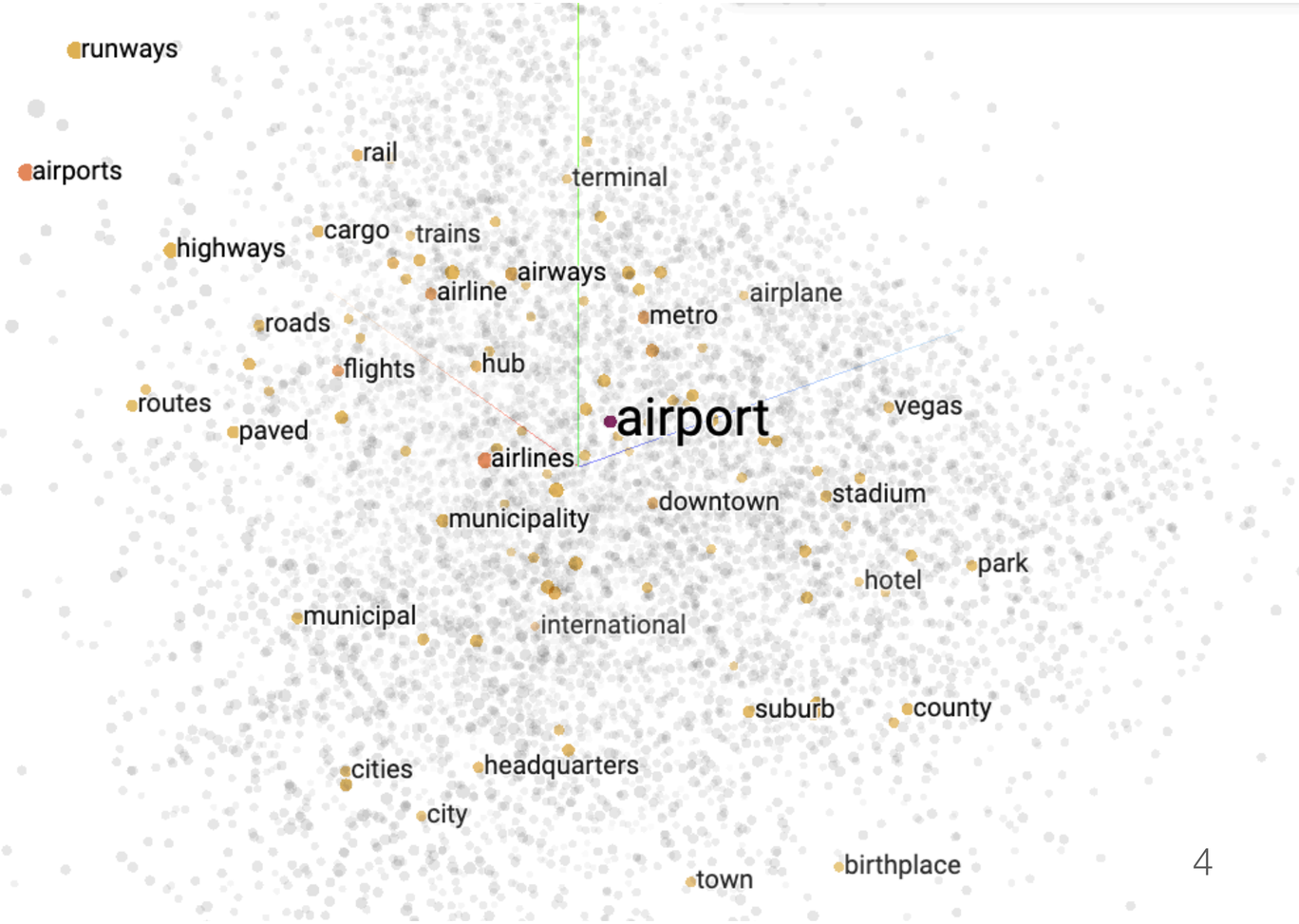
1. Nearest neighbours

Demo

Using Euclidean distance or cosine similarity

Nearest points in the original space:

airports	0.449
airlines	0.459
flights	0.493
airline	0.511
metro	0.525
international	0.546
ferry	0.550
downtown	0.556
airways	0.575
rail	0.578
passengers	0.579
hub	0.580



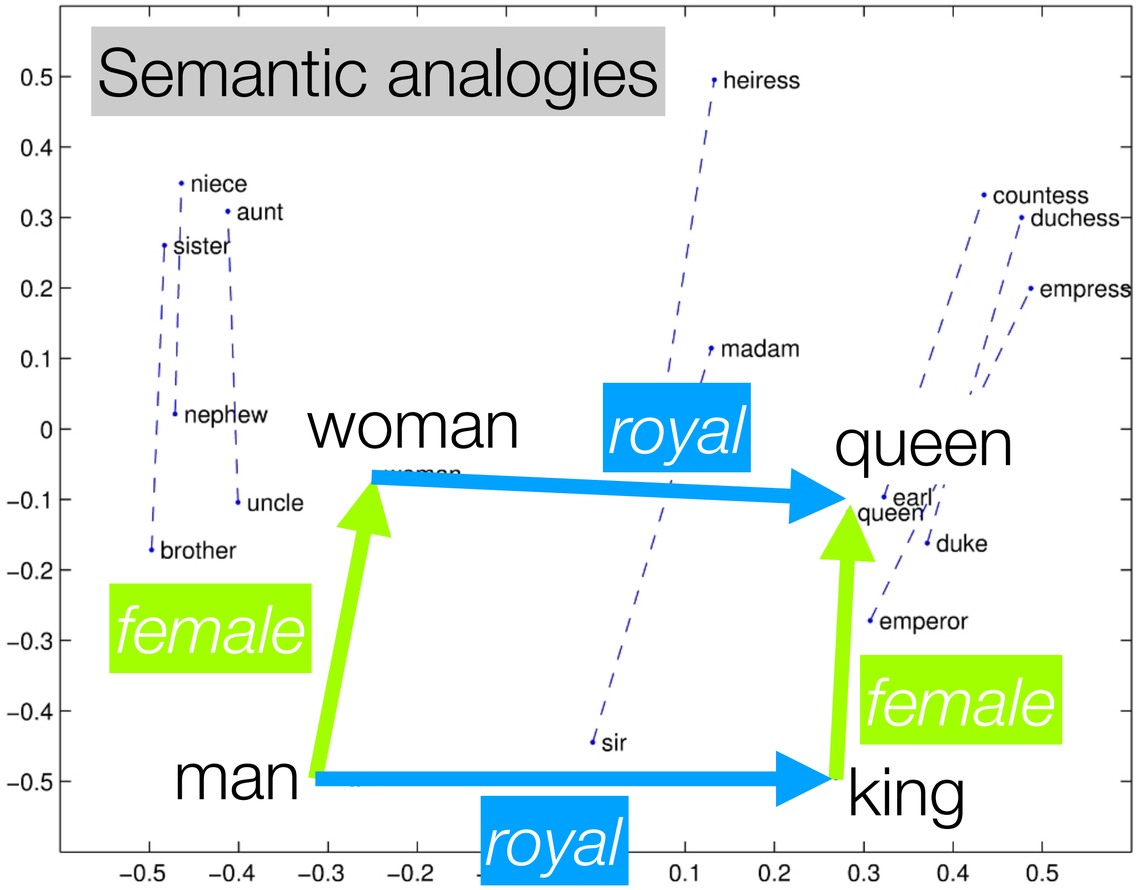
Properties of Word Embeddings

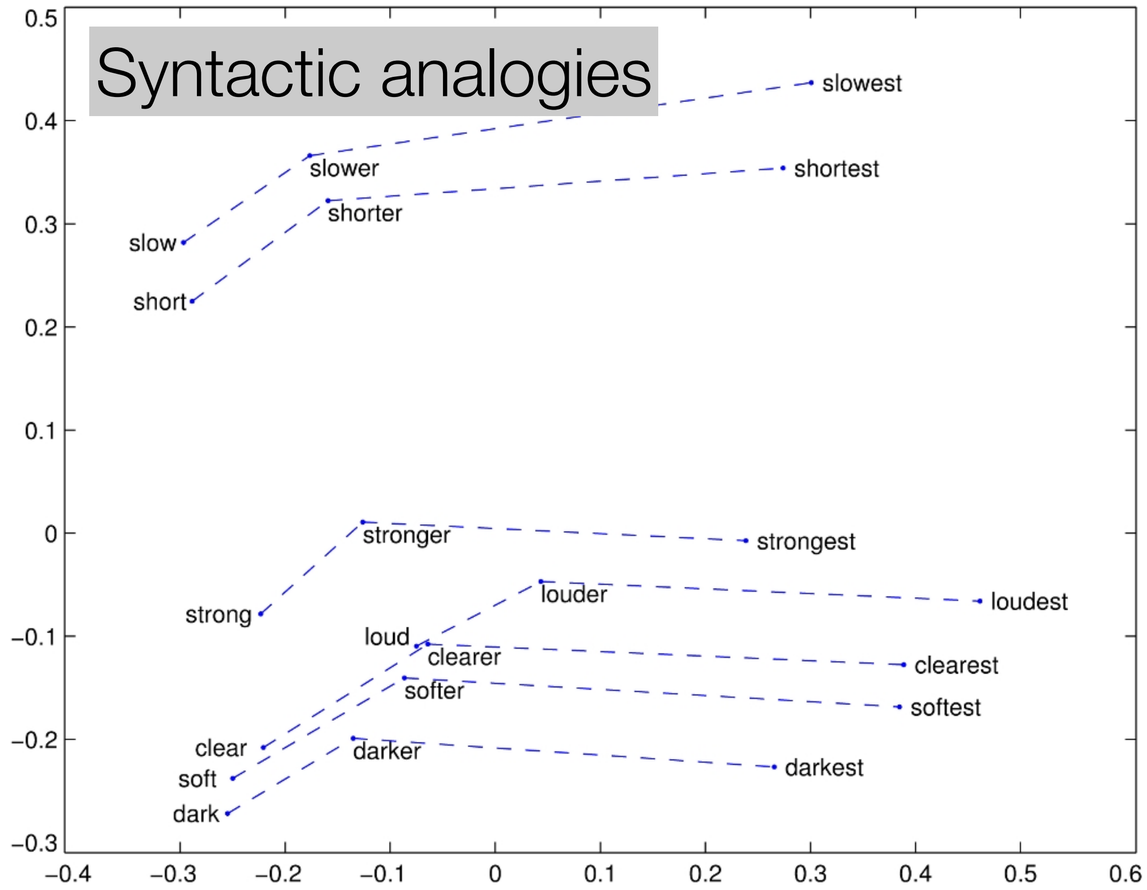
nlp.stanford.edu/projects/glove

2. Linear substructures

Demo

For visualisation purposes: dimensions further reduced using or t-SNE





Word2Vec

Zebra

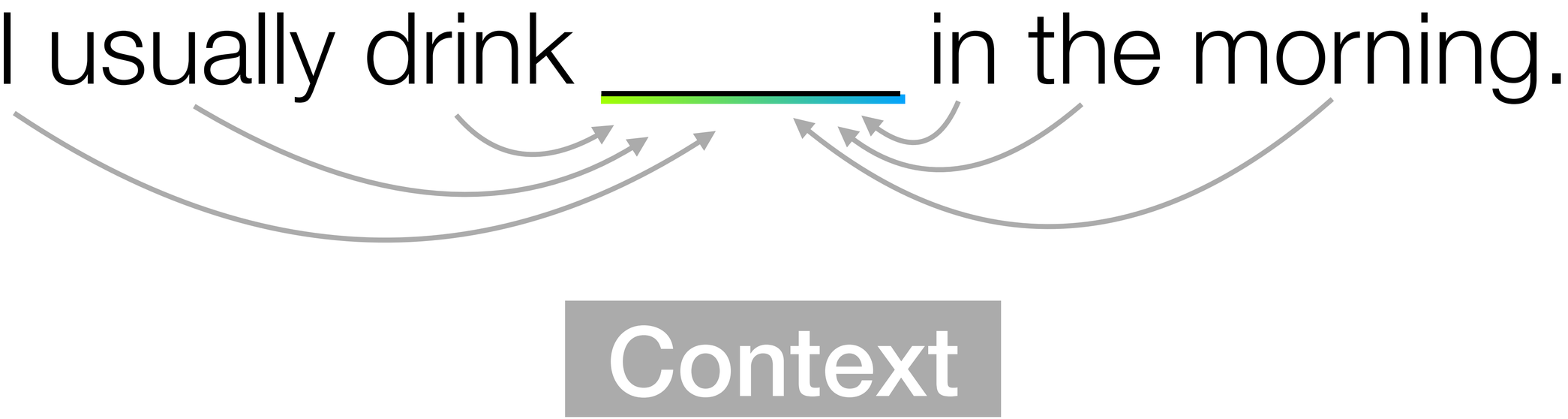
0.73	-0.42	0.07	0.87

- Group of models to compute "*continuous vector representations of words*" (Word Embeddings)
- Two model architectures
  - Continuous Bag-of-Words (CBOW)
  - Continuous Skip-gram

Word2Vec

Continuous Bag-of-Words (CBOW)

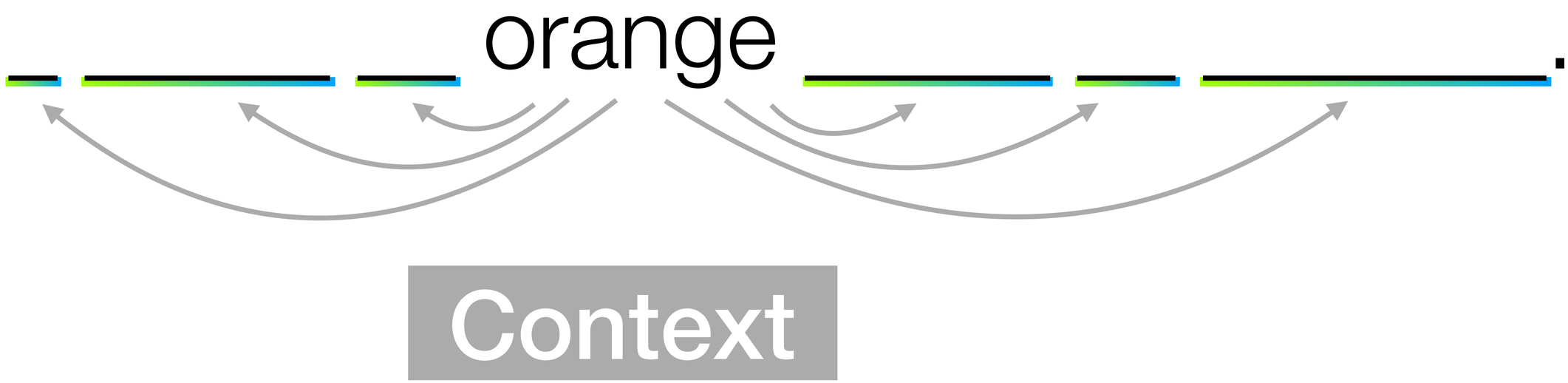
- Given the surrounding words, predict the middle word.



Word2Vec

Continuous Skip-gram

- Given a word, predict the surrounding words.



Word2Vec

Continuous Skip-gram

- Context window size: 2

Jupiter	is	the	largest	planet	in	the	Solar	System

Word2Vec

Continuous Skip-gram

- Context window size: 2

Jupiter	is	the	largest	planet	in	the	Solar	System

Input	Label
Jupiter	is
Jupiter	the

Word2Vec

Continuous Skip-gram

- Context window size: 2

Jupiter	is	the	largest	planet	in	the	Solar	System

Input	Label
Jupiter	is
Jupiter	the
is	Jupiter
is	the
is	largest

Word2Vec

Continuous Skip-gram

- Context window size: 2

Jupiter	is	the	largest	planet	in	the	Solar	System

Input	Label
Jupiter	is
Jupiter	the
is	Jupiter
is	the
is	largest
the	Jupiter
the	is
the	largest
the	planet

Word2Vec

Continuous Skip-gram

- Context window size: 2

Jupiter	is	the	largest	planet	in	the	Solar	System

Input	Label
Jupiter	is
Jupiter	the
is	Jupiter
is	the
is	largest
the	Jupiter
the	is
the	largest
the	planet

Input	Label
largest	is
largest	the
largest	planet
largest	in

Word2Vec

Continuous Skip-gram

- Context window size: 2

Jupiter	is	the	largest	planet	in	the	Solar	System

Input	Label
Jupiter	is
Jupiter	the
is	Jupiter
is	the
is	largest
the	Jupiter
the	is
the	largest
the	planet

Input	Label
largest	is
largest	the
largest	planet
largest	in
planet	the
planet	largest
planet	in
planet	the

Word2Vec

Continuous Skip-gram

- Context window size: 2

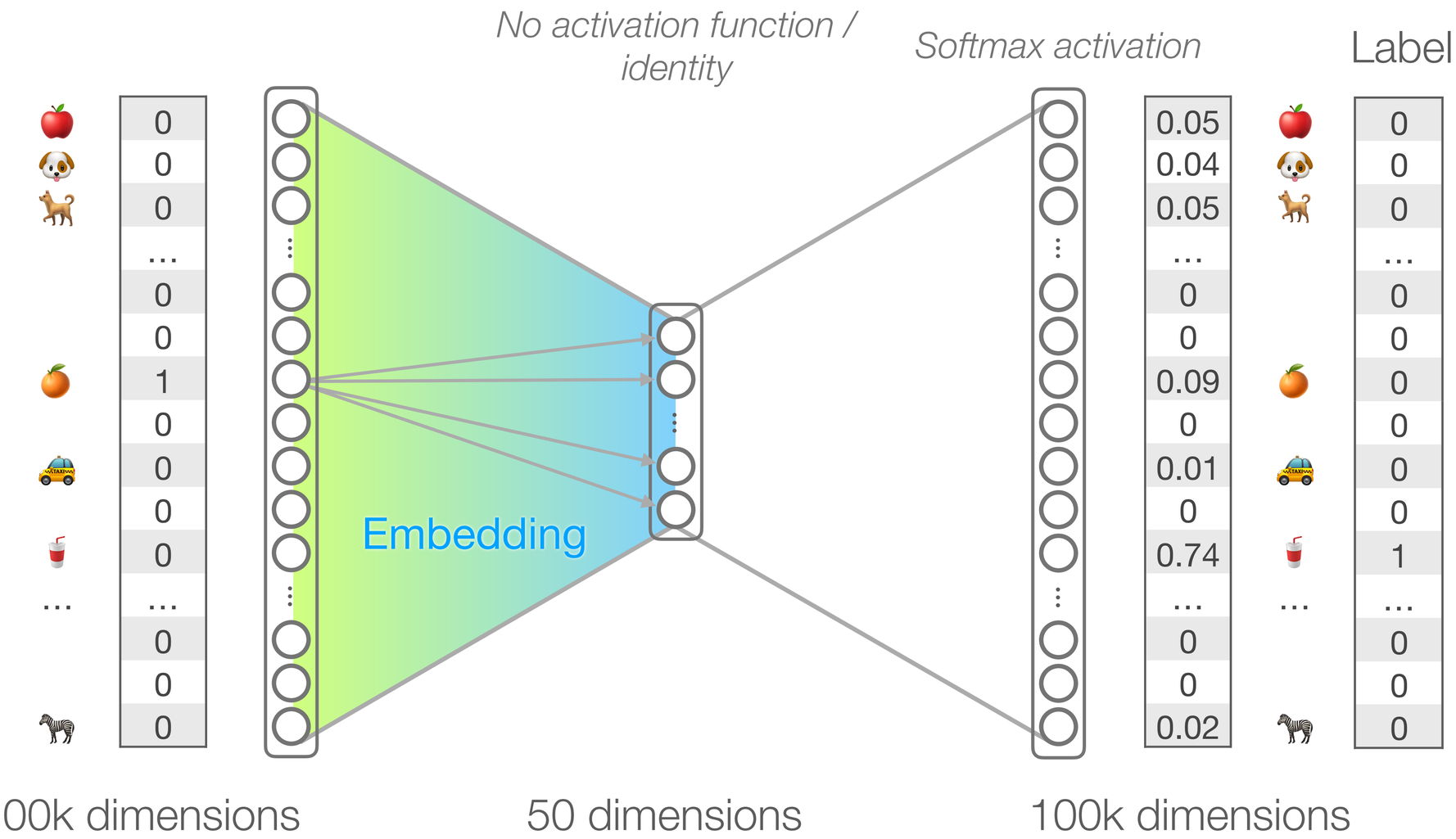
Jupiter	is	the	largest	planet	in	the	Solar	System

Input	Label
Jupiter	is
Jupiter	the
is	Jupiter
is	the
is	largest
the	Jupiter
the	is
the	largest
the	planet

Input	Label
largest	is
largest	the
largest	planet
largest	in
planet	the
planet	largest
planet	in
planet	the
…	…

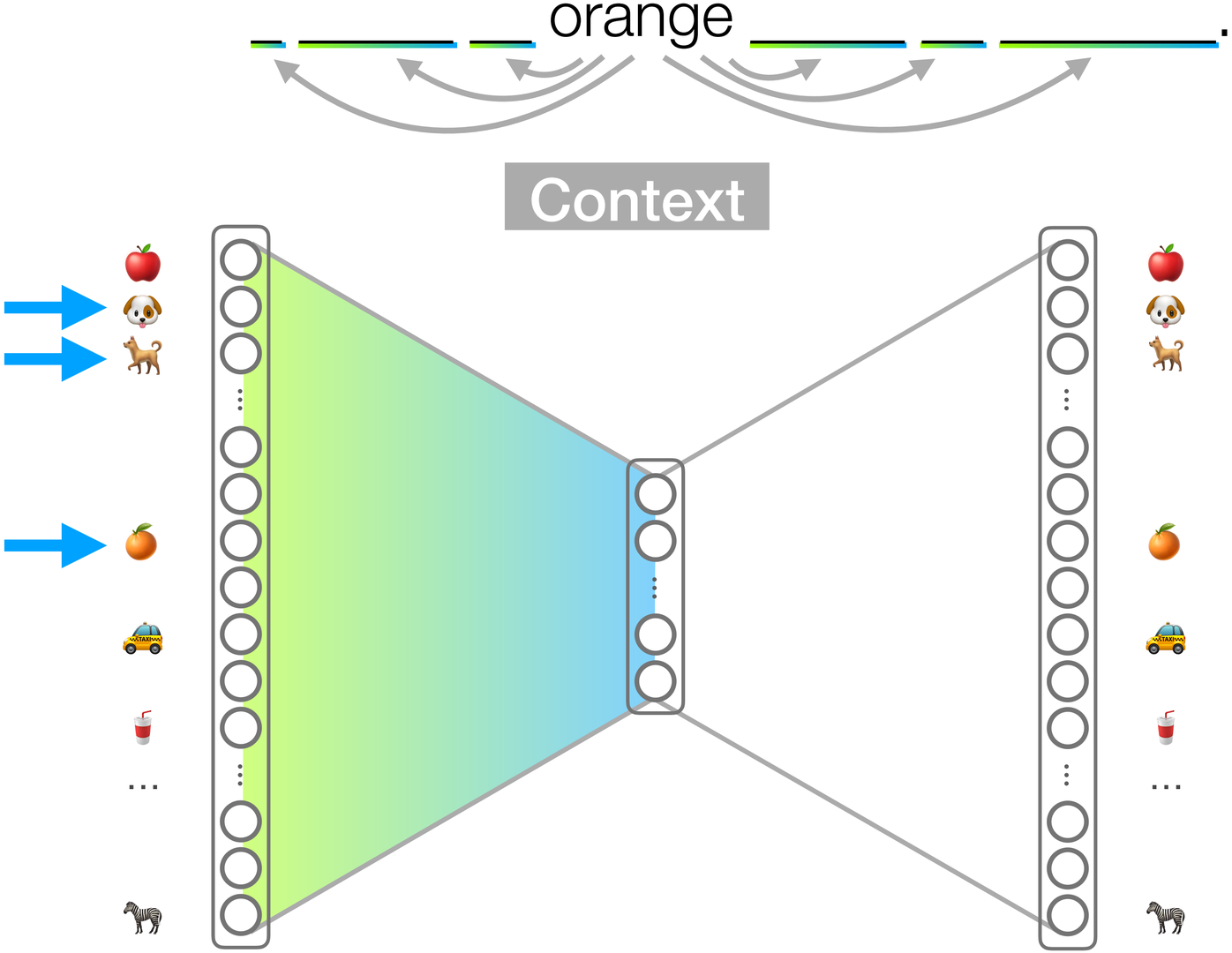
Word2Vec

Continuous Skip-gram - The Model



(Word) Embeddings

Intuition



Word2Vec

Hyperparameters

- Model architecture
  - Continuous Bag-of-Words (CBOW)
  - Continuous Skip-gram
- Vocabulary
- Embedding dimension
  - Typically 25 - 300
- Context window size
  - Typically 5 - 10
- Sampling techniques
  - Subsampling of frequent words
  - Negative sampling

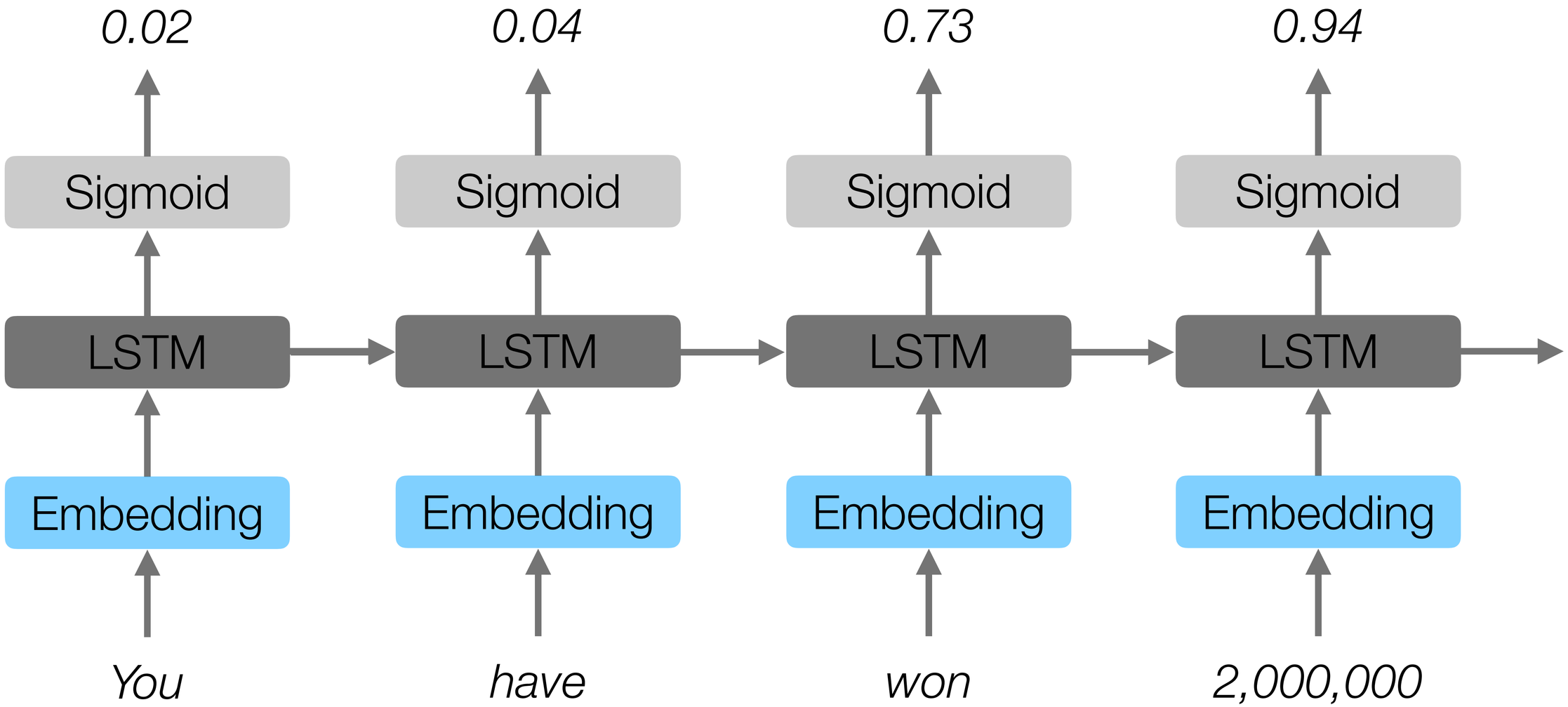
Embeddings in Machine Learning Frameworks

```
from keras.layers import Embedding

# ...
model.add(Embedding(input_dim=len(vocabulary), output_dim=50))
# ...
```

```
from torch import nn

# ...
self.embedding = nn.Embedding(num_embeddings=len(vocabulary),
                              embedding_dim=50)
# ...
```

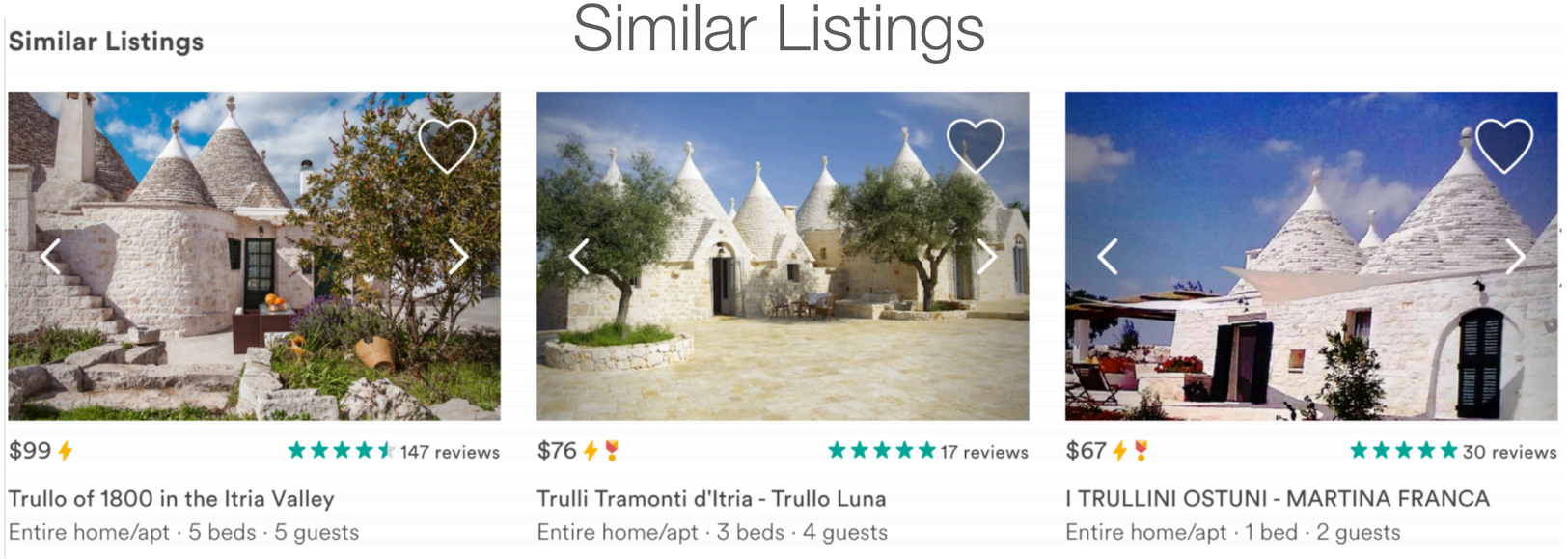
NLP Use Case: Spam Detection

Demo

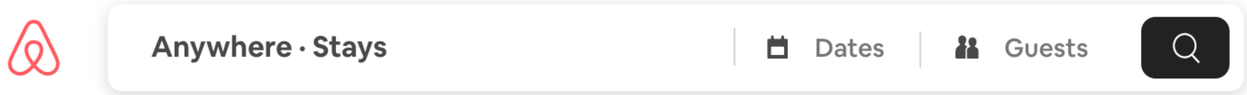


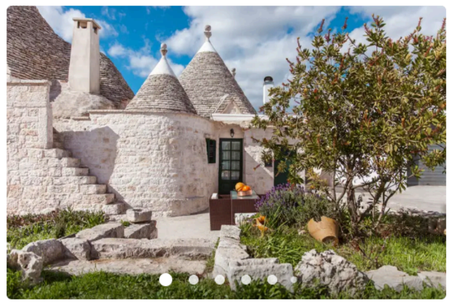
Airbnb

Similar Listings



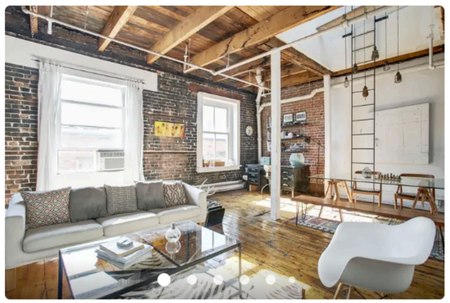
Airbnb





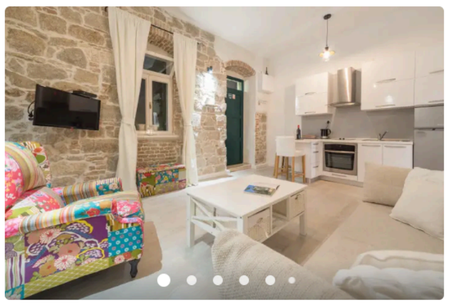
Cisternino, Brindisi ★4.71
Trullo del 1800 in Valle d'Itria

23420821



Old Montreal · Montreal ★4.78
Skylit Loft in Old Montreal

5070992



SUPERHOST Varoš · Split ★4.90
Twins Apartment in Center 4*

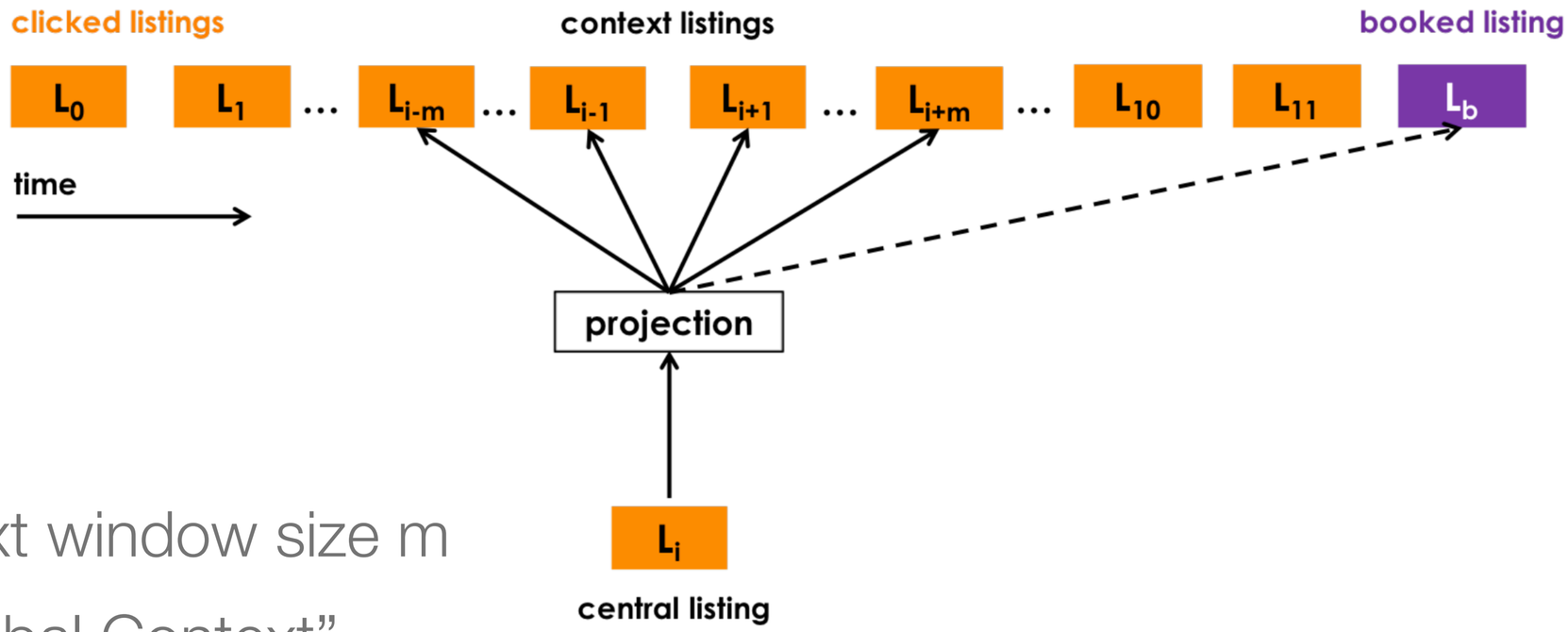
11031421

LB of Hammersmith & Fulham ·... ★4.87
Stylish house close to river thames

1012534

Listing Embeddings @ Airbnb

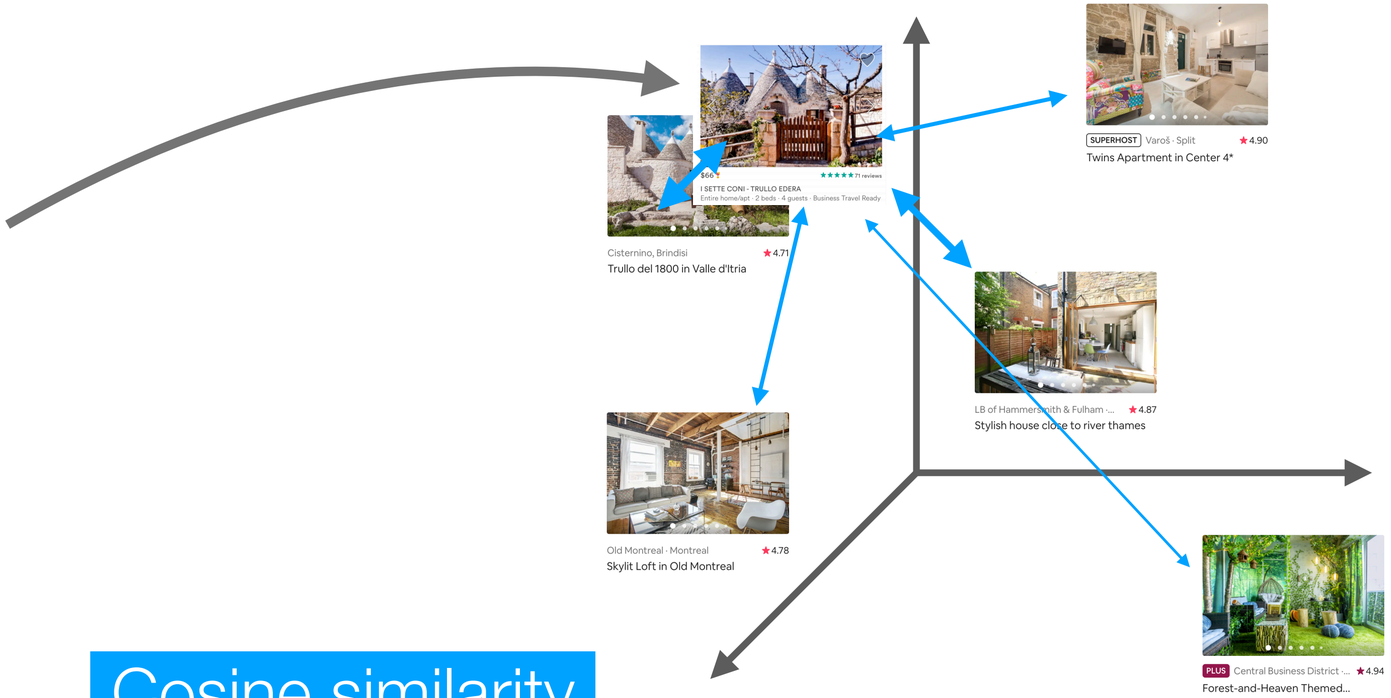
"Click Session"



- Skip gram with context window size m
- Booked listing as "Global Context"

Listing Embeddings @ Airbnb

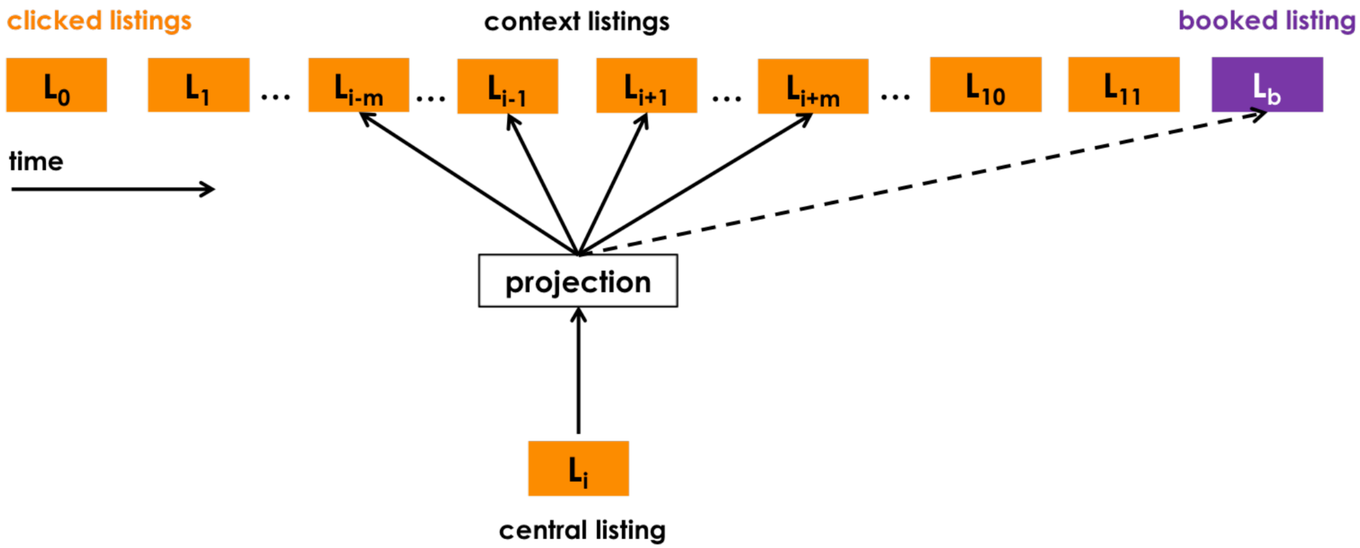
Finding similar Listings



Cosine similarity

$$\text{similarity}(L_i, L_j) = \cos(\theta_{L_i,L_j}) = \frac{\mathbf{L_i} \cdot \mathbf{L_j}}{\|\mathbf{L_i}\| \|\mathbf{L_j}\|} \quad \in [-1,1]$$

Listing Embeddings @ Airbnb



- Context window size m = 5
- Embedding dimension: 32
- Vocabulary: 4.5M listings
- Training data: 800M click sessions
- Daily re-training from scratch with sliding window training data

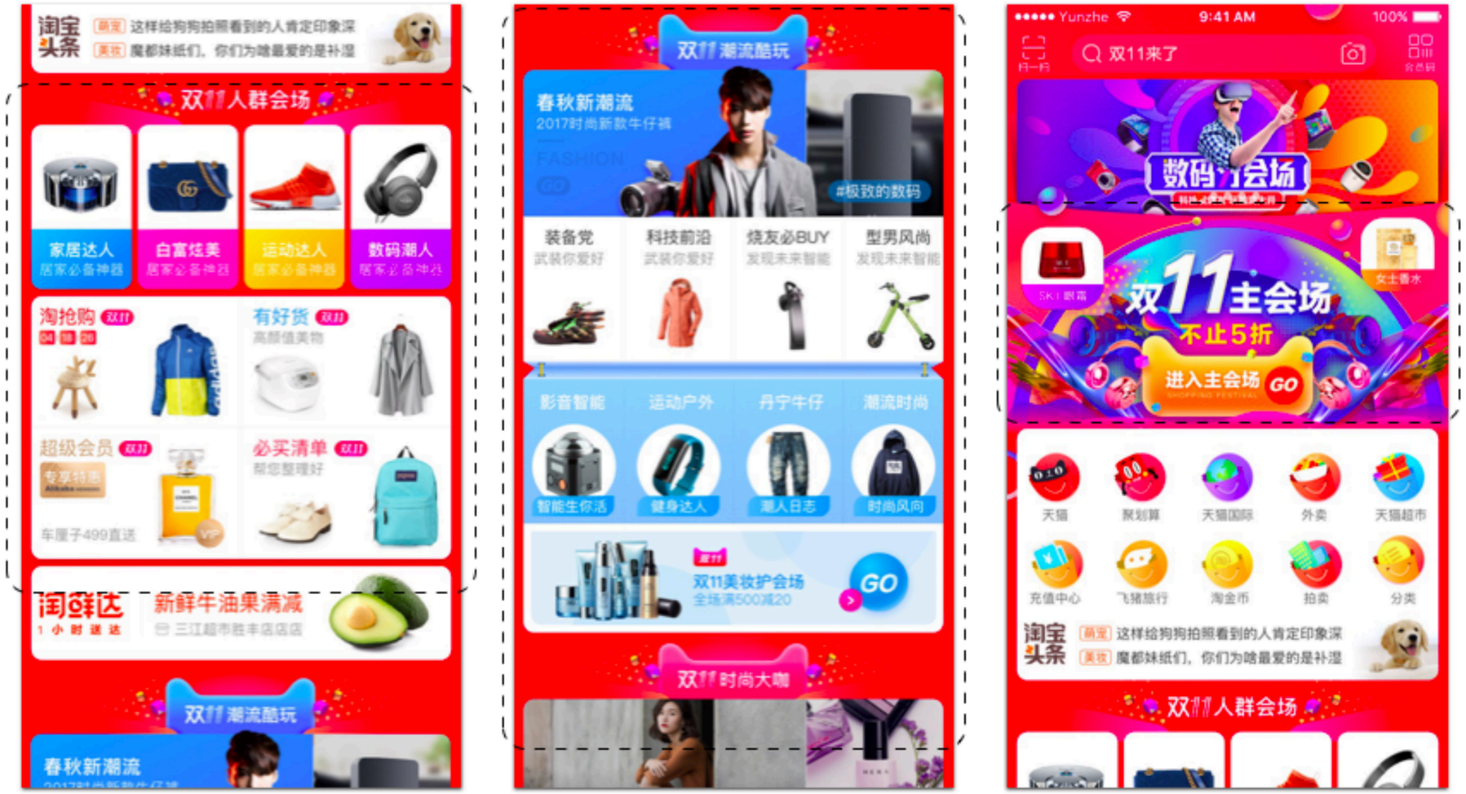
⇒ “21% increase in Similar Listing carousel CTR”

Alibaba (Taobao)

- “eBay” of China

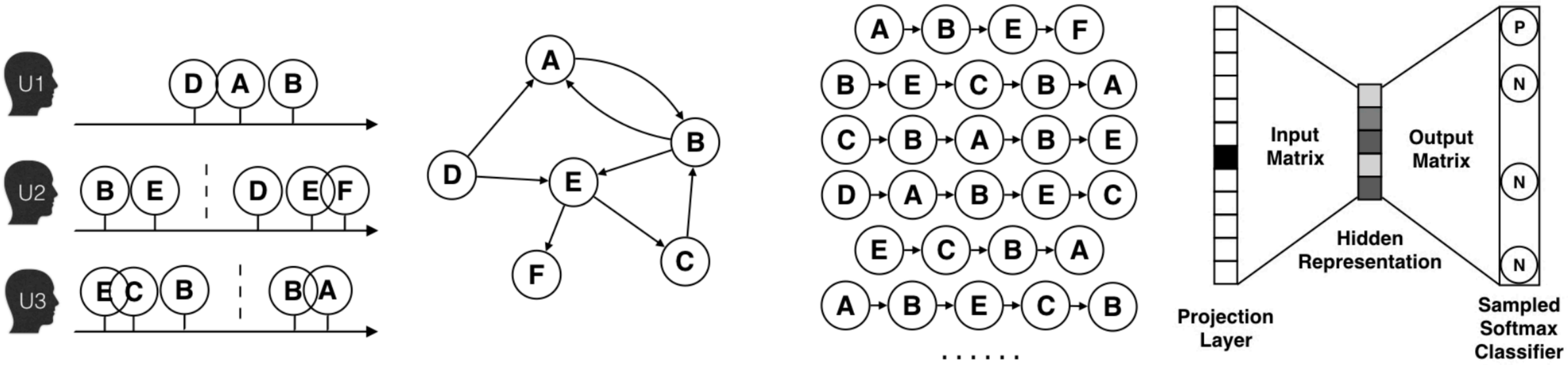
Alibaba Group

淘宝网 Taobao.com



Graph Embedding (Deep Walk)

淘宝网 Taobao.com



(a) Users' behaviours sequences

(b) Item graph construction

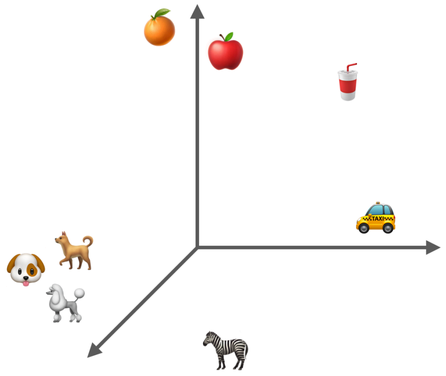
(c) Random walk generation

(d) Embedding with Skip-Gram

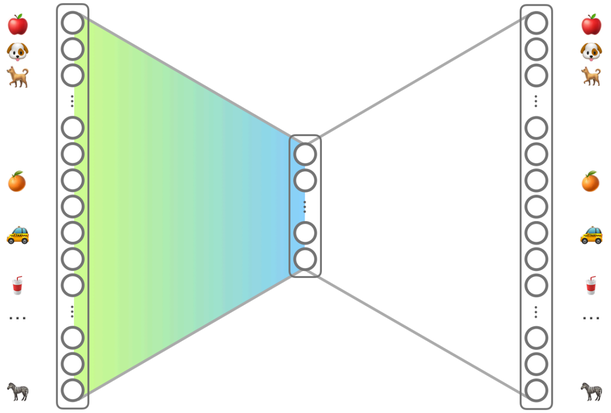
- Context window size: 5
- Embedding dimension: 160
- Vocabulary: ~2B items

The Wonderful World of Embeddings

Embedding



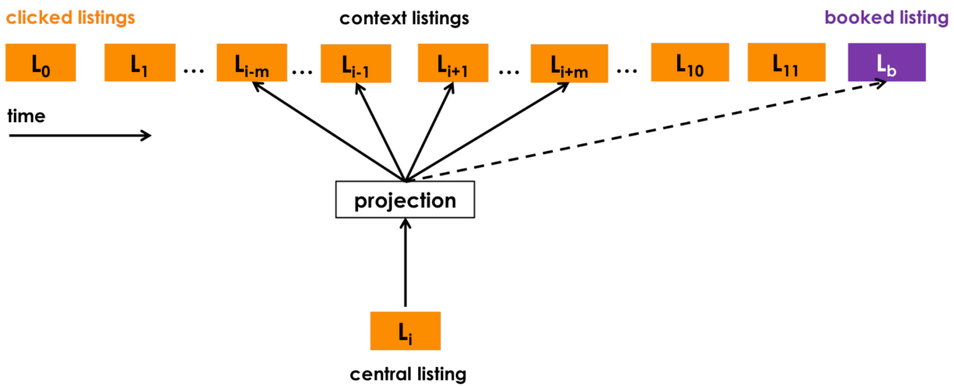
Skip-gram Model Architecture



Franz Diebold, M.Sc. (TUM)

Skip-gram

Jupiter	is	the	largest	planet	in	the	Solar	System



github.com/franzdiebold

linkedin.com/in/franz-diebold

Sources

- http://jalammar.github.io/illustrated-word2vec/
- http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

There is even more…

- Thought Vector
- Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings - Bolukbasi et al. (2016)

One Hot Encoding

- Vocabulary: {🐶, 🐰, 🐸, 🦄}
- Sequence to encode: 🐰 🐸 🐰 🦄

In general

- Set of items to be encoded
  - n items → **n bit representation**
- Comparable to statistics: “dummy variables”

	🐶	🐰	🐸	🦄
🐰	0	1	0	0
🐸	0	0	1	0
🐰	0	1	0	0
🦄	0	0	0	1

One Hot Encoding - Limitations

What if?

100k items

- 100k bits/dimensions

⇒ Highly sparse matrix

- Large vocabulary

vocabulary = {🐶, 🐕, 🐩, 🚕}

- (🐶, 🐕) vs. (🐶, 🚕)

⇒ Degree of similarity?

- Synonyms
- Ambiguities

⬇

Human Language

Word2Vec

Continuous Bag-of-Words (CBOW)

"Jupiter is the largest planet in the Solar System."

- Window size: 2

w(t-2)	w(t-1)	w(t+1)	w(t+2)	w(t) ≙ label
-	-	is	the	Jupiter
-	Jupiter	the	largest	is
Jupiter	is	largest	planet	the
is	the	planet	in	largest
the	largest	in	the	planet
largest	planet	the	Solar	in
planet	in	Solar	System	the
in	the	System	-	Solar
the	Solar	-	-	System

GloVe: Global Vectors for Word Representation

- Pre-trained word vectors

Source	#Tokens	Vocab size	Dimensionalities	Size
Wikipedia 2014 + Gigaword 5	6B	400K	50d, 100d, 200d, 300d	822 MB
Common Crawl	42B	1.9M	300d	1.75 GB
Common Crawl	840B	2.2M	300d	2.03 GB
Twitter	27B	1.2M	25d, 50d, 100d, 200d	1.42 GB

GloVe: Global Vectors for Word Representation

Example

Wikipedia 2014 + Gigaword 5, 50d

0	1	2	3	4	5	6	7	8	9	10	...	41	42
the	0.418000	0.249680	-0.41242	0.12170	0.34527	-0.044457	-0.49688	-0.17862	-0.00066	-0.656600	...	-0.298710	-0.15749
,	0.013441	0.236820	-0.16899	0.40951	0.63812	0.477090	-0.42852	-0.55641	-0.36400	-0.239380	...	-0.080262	0.63003
.	0.151640	0.301770	-0.16763	0.17684	0.31719	0.339730	-0.43478	-0.31086	-0.44999	-0.294860	...	-0.000064	0.06898
of	0.708530	0.570880	-0.47160	0.18048	0.54449	0.726030	0.18157	-0.52393	0.10381	-0.175660	...	-0.347270	0.28483
to	0.680470	-0.039263	0.30186	-0.17792	0.42962	0.032246	-0.41376	0.13228	-0.29847	-0.085253	...	-0.094375	0.01832

5 rows × 50 columns

Continuous Skip-gram - Improvements (1/2)

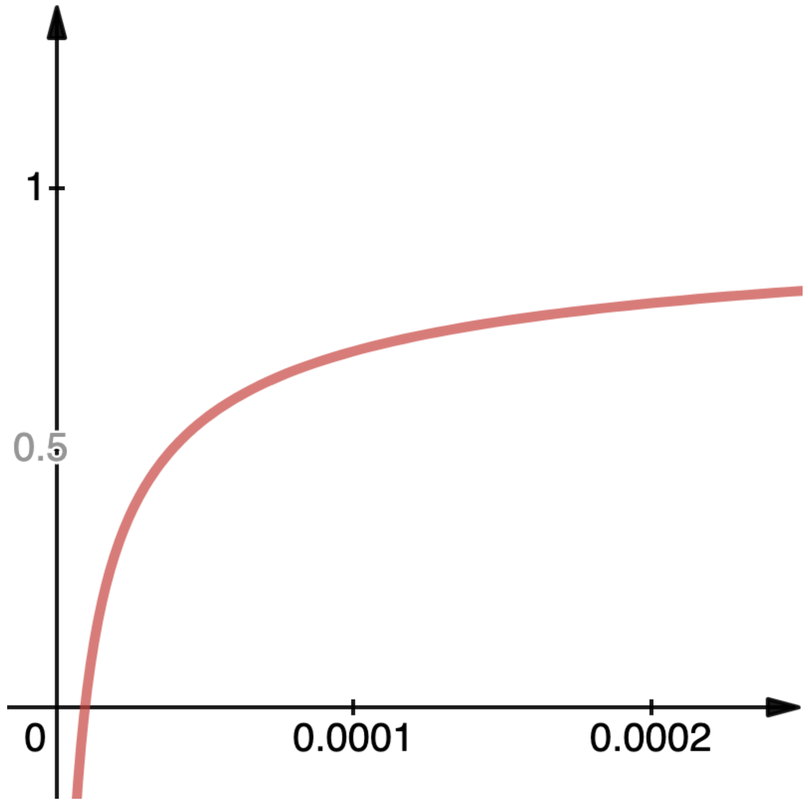
How to deal with frequent words like “the”, “of” or “to”?

⇒ Subsampling of Frequent Words

- Discard word w_i with probability

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

- Word frequency $f(w_i)$
- Threshold t, i.e. 10^{-5}



Continuous Skip-gram - Improvements (2/2)

Negative Sampling

- “Sampled Softmax”: Compute prediction for correct label and some random other labels

  ⇒ Estimate Softmax distribution

  ⇒ Much faster!