WORD EMBEDDINGS

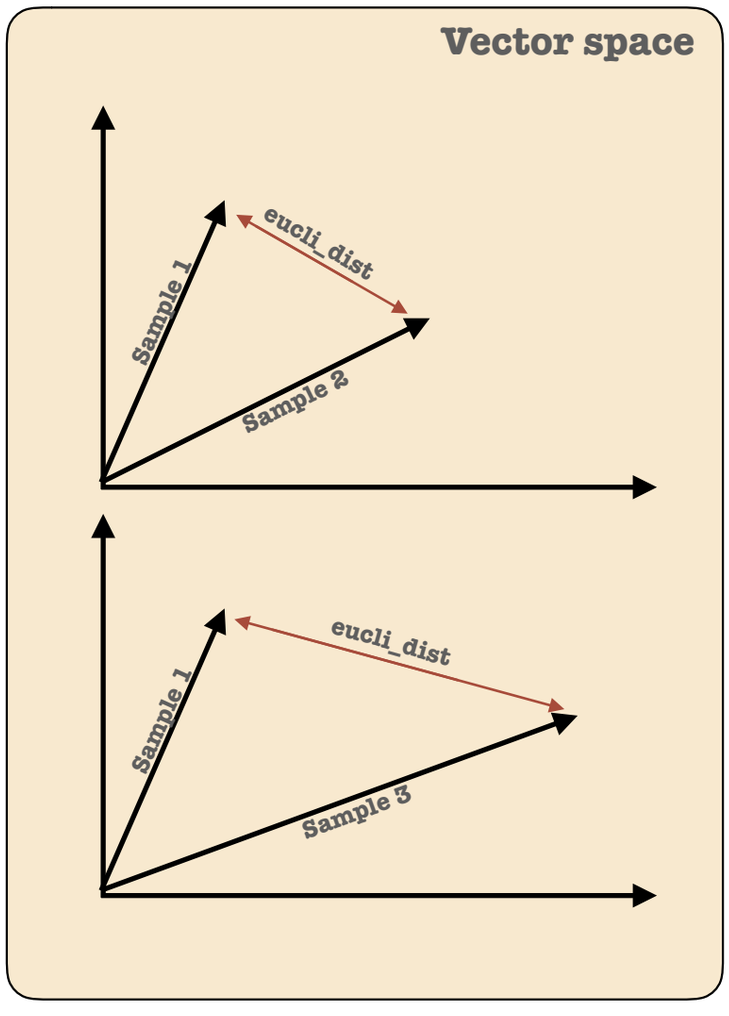
Topics

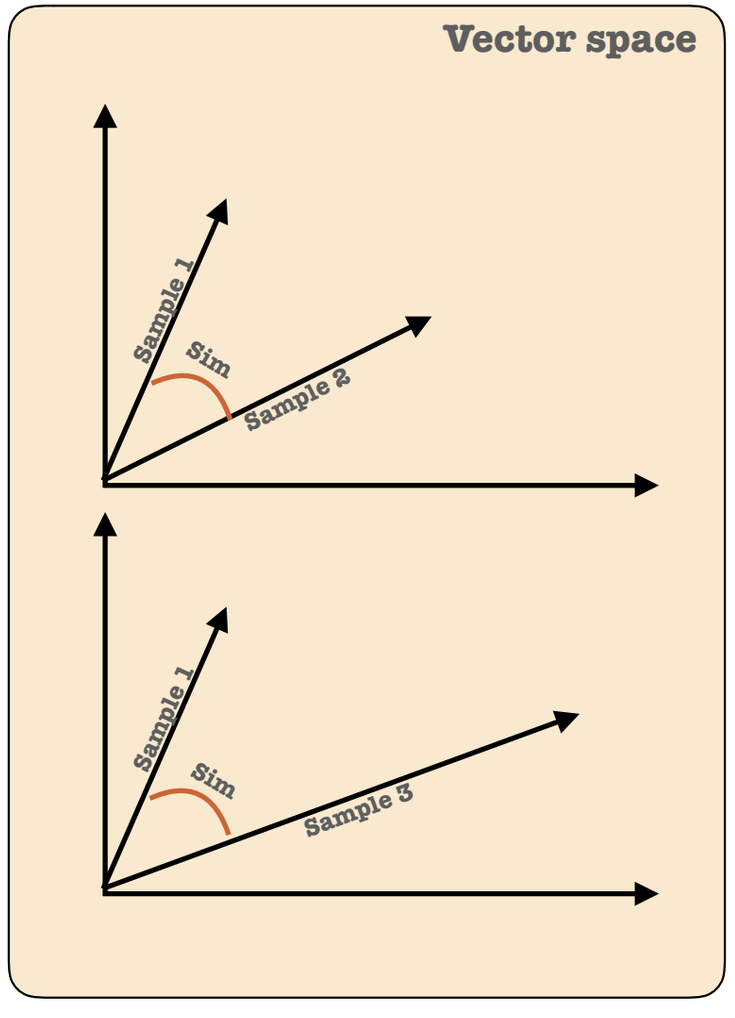
1. Document embeddings using TF-IDF / CV
2. Word embeddings
3. Skip-gram
4. Continuous bag of words
5. Limitations of ANN

Document representation

There are various ways to represent documents. These representations allow documents with similar meaning to have similar representations.

	camera	good	best	cloth	material
Sample 1: good camera	**1.**	**1.**	**0.**	**0.**	**0.**
Sample 2: best camera	**0.**	**1.**	**1.**	**0.**	**0.**
Sample 3: cloth material	**0.**	**0.**	**0.**	**1.**	**1.**





EucliDis(S1 to S2) < EucliDis(S1 to S3)

Similarity(S1 to S2) > Similarity(S1 to S3)

Word embeddings

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation

1. **Skip-gram**
2. **CBOW- Continuous Bag Of Words**

1 Skip-gram

Corpus

S1 Camera is good
S2 Camera resolution is good
S3 Battery is good. Battery is super
S4 Battery life is good

Training Data

	Camera	is	good	resolution	battery	super	life	Camera	is	good	resolution	battery	super	life
1. Camera is	1	0	0	0	0	0	0	0	1	0	0	0	0	0
2. Is good	0	1	0	0	0	0	0	0	0	1	0	0	0	0
3. Camera resolution	1	0	0	0	0	0	0	0	0	0	1	0	0	0
4. Resolution is	0	0	0	1	0	0	0	0	1	0	0	0	0	0
5. Is good	0	1	0	0	0	0	0	0	0	1	0	0	0	0
6. Battery is	0	0	0	0	1	0	0	0	1	0	0	0	0	0
7. Is good	0	1	0	0	0	0	0	0	0	1	0	0	0	0
8. Battery is	0	0	0	0	1	0	0	0	1	0	0	0	0	0
9. Is super	0	1	0	0	0	0	0	0	0	0	0	0	1	0
10. Battery life	0	0	0	0	1	0	0	0	0	0	0	0	0	1
11. Life is	0	0	0	0	0	0	1	0	1	0	0	0	0	0
12. Is good	0	1	0	0	0	0	0	0	0	1	0	0	0	0

Camera	is	good	resolution	battery	super	life
1	0	0	0	0	0	0
0	1	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	0	0
0	0	0	0	1	0	0
0	1	0	0	0	0	0
0	0	0	0	1	0	0
0	1	0	0	0	0	0
0	0	0	0	1	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	1
0	1	0	0	0	0	0

Camera	0
is	0
good	0
resolution	1
battery	0
super	0
life	0

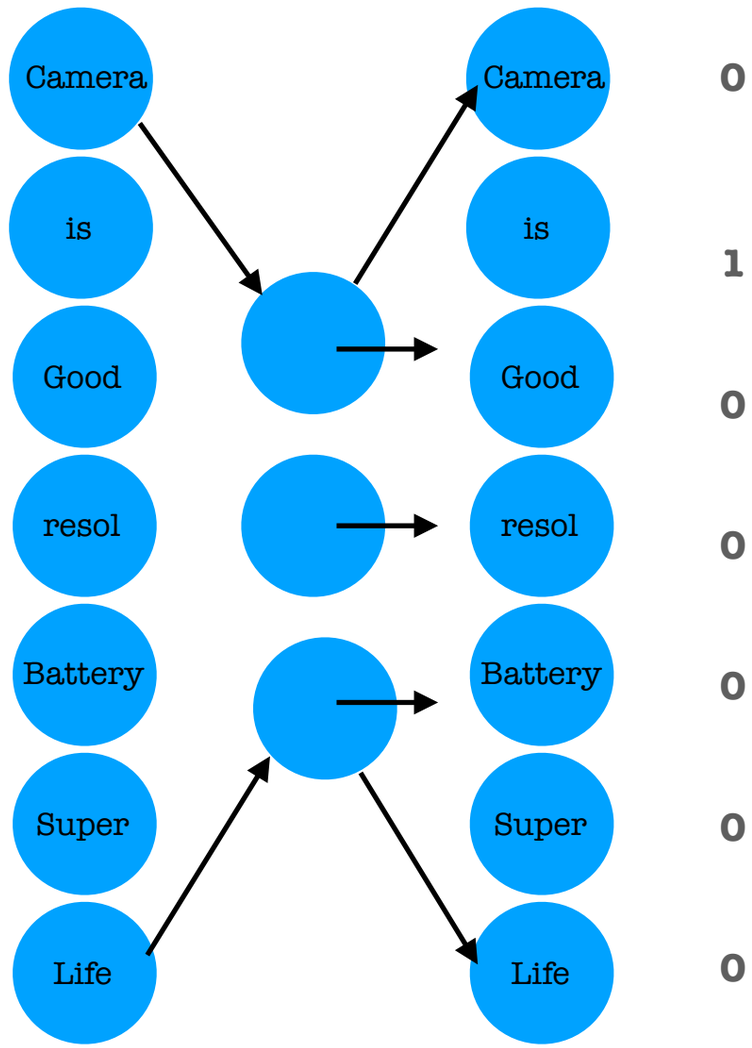
1

Camera	is	good	resolution	battery	super	life
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1
0	1	0	0	0	0	0
0	0	1	0	0	0	0

0

Input

0
0
0
0
0
0



1
0
0
0
0
0

Target

2 CBOW- Continuous Bag Of Words

Corpus

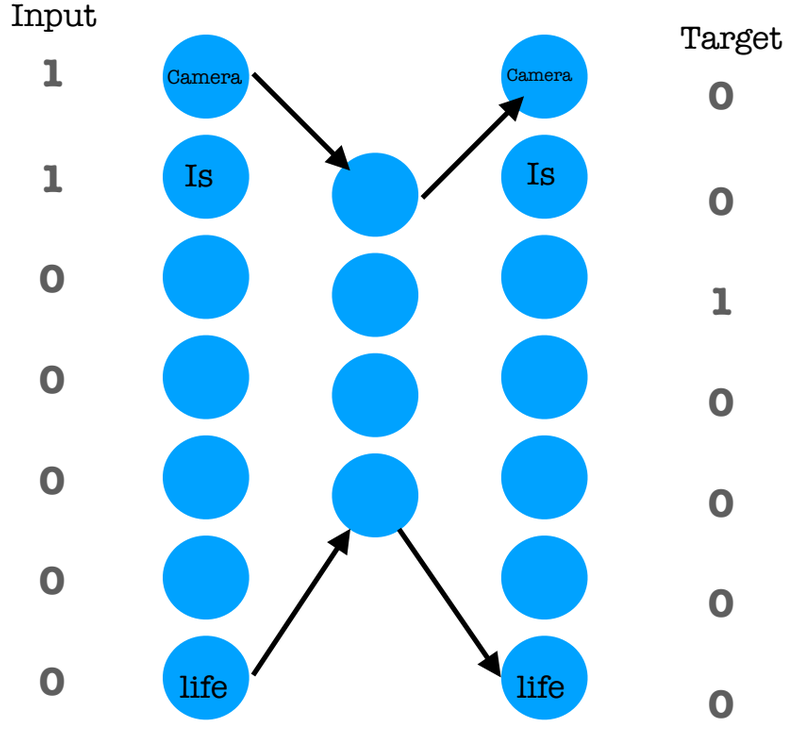
S1 Camera is good
S2 Camera resolution is good
S3 Battery is good. Battery is super
S4 Battery life is good

Training Data

	Camera	is	good	resolution	battery	super	life	Camera	is	good	resolution	battery	super	life
Camera is good	1	1	0	0	0	0	0	0	0	1	0	0	0	0
Camera resolution is	1	0	0	1	0	0	0	0	1	0	0	0	0	0
Resolution is good	0	1	0	1	0	0	0	0	0	1	0	0	0	0
Battery is good	0	1	0	0	1	0	0	0	1	0	0	0	0	0
Battery is super	0	1	0	0	1	0	0	0	0	0	0	0	1	0
Battery life is	0	1	0	0	1	0	0	0	1	0	0	0	0	0
Life is good	0	1	0	0	0	0	1	0	0	1	0	0	0	0

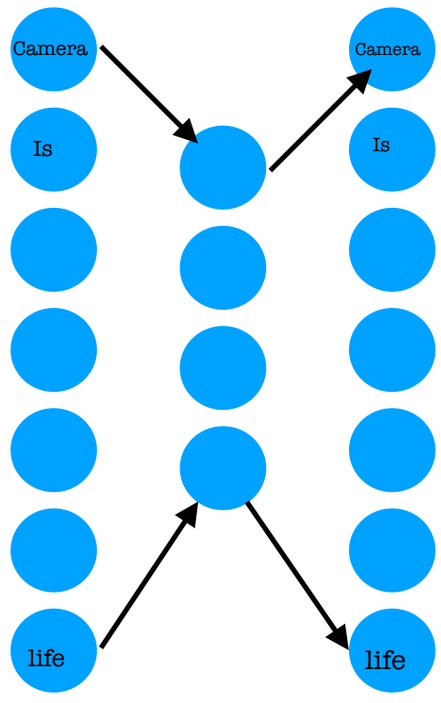
Camera	is	good	resolution	battery	super	life
1	1	0	0	0	0	0
1	0	0	1	0	0	0
0	1	0	1	0	0	0
0	1	0	0	1	0	0
0	1	0	0	1	0	0
0	1	0	0	1	0	0
0	1	0	0	0	0	1

Camera	is	good	resolution	battery	super	life
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	1	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0



2. **How we can get document embedding for a particular document using ANN?**

   A. **Train Skip-gram or Continuous bag of words**
   B. **Represent your document using Count vectorizer or Tf-Idf**
   C. **Pass this represented document to trained model**
   D. **The output of hidden layer represents the document**

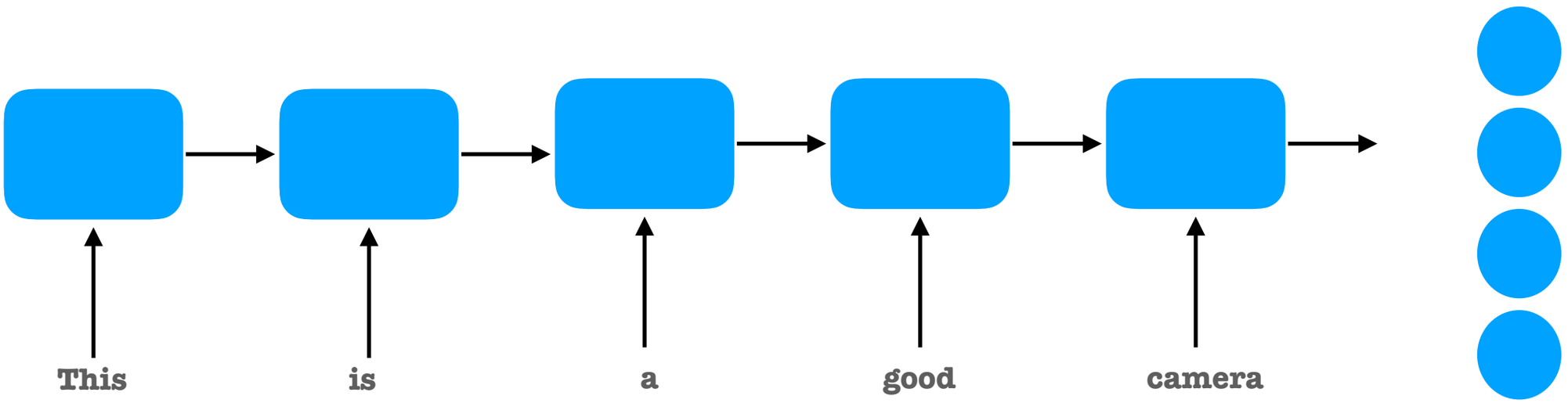


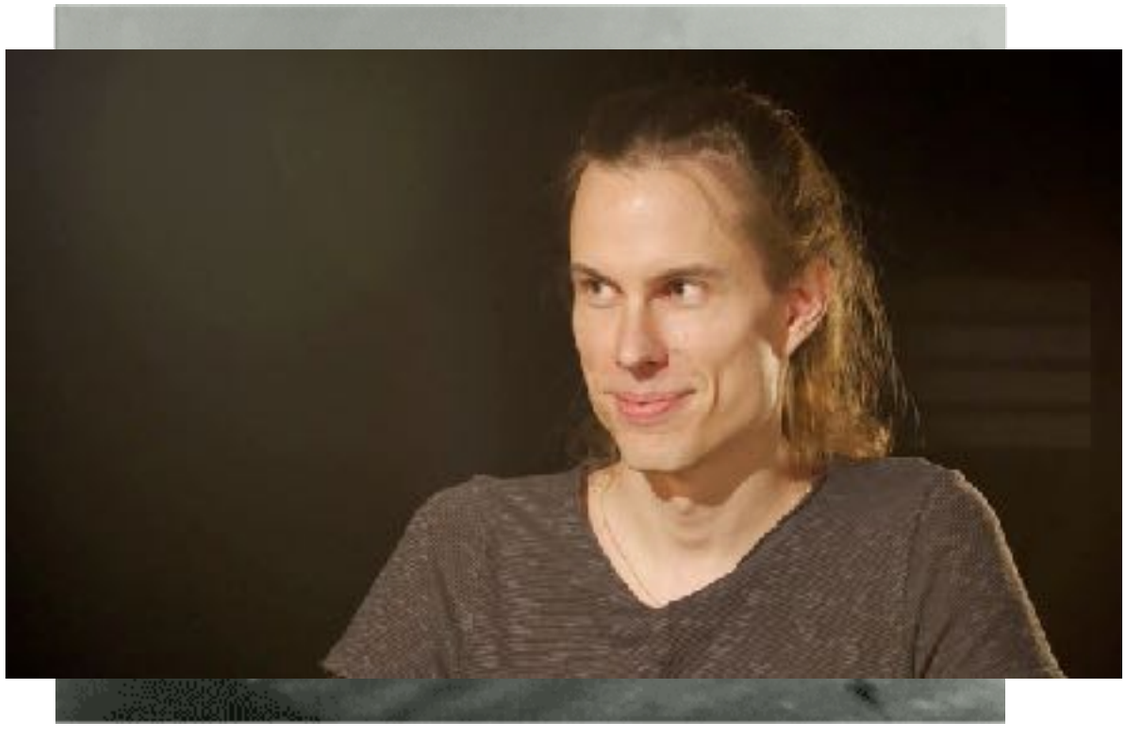
Limitation of ANN

	camera	good	best	bad	material	Target
Sample 1: good camera	1.	1.	0.	0.	0.	1
Sample 2: best camera	0.	1.	1.	0.	0.	1
Sample 3: bad material	0.	0.	0.	1.	1.	0

ANN doesn't look at the context of sentence what you pass but it will just see the words individually.
If the sentence looses its context then there will be high chance that it might loose the meaning also.

Example: This is a good camera





Summary

1 **Artificial Neural Networks**

Count Vectorizer 1943

A document can be represented as vector using CV

2 **Tf-IDF**

Karen Spärck Jones 1972

A document can be represented as vector using Tf-IDF

Can we represent words as vectors using ANN?

3 **Word2Vec**

Tomas Mikolov 2013

A word can be represented as vector using word2vec

Can we represent document as vectors using ANN?

4 **Word2Vec**

Tomas Mikolov 2013

Can we represent document as vectors using NN without disturbing the context?

5 **RNN**