NUI Galway
OÉ Gaillimh

Introduction to NLP

Vector Space Model & Word Embeddings

Dr. Paul Buitelaar
Data Science Institute, NUI Galway

Learning Outcomes of This Lecture

Understand the distributional (vector space) model for NLP

Learn how to construct and use word vectors in semantic similarity

Insight into use of linguistic context in constructing word vectors

Introduction to word embeddings and pre-trained models in NLP

Overview

Vector Space Model

Distributional Semantics

Context in Distributional Representations

Word Embeddings

Overview

Vector Space Model

Distributional Semantics

Context in Distributional Representations

Word Embeddings

Vector Space Model

Foundations in Information Retrieval (Salton et al., 1975)

Represent document set by distribution of index terms

Represent distributional model in vector space

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

Document Set

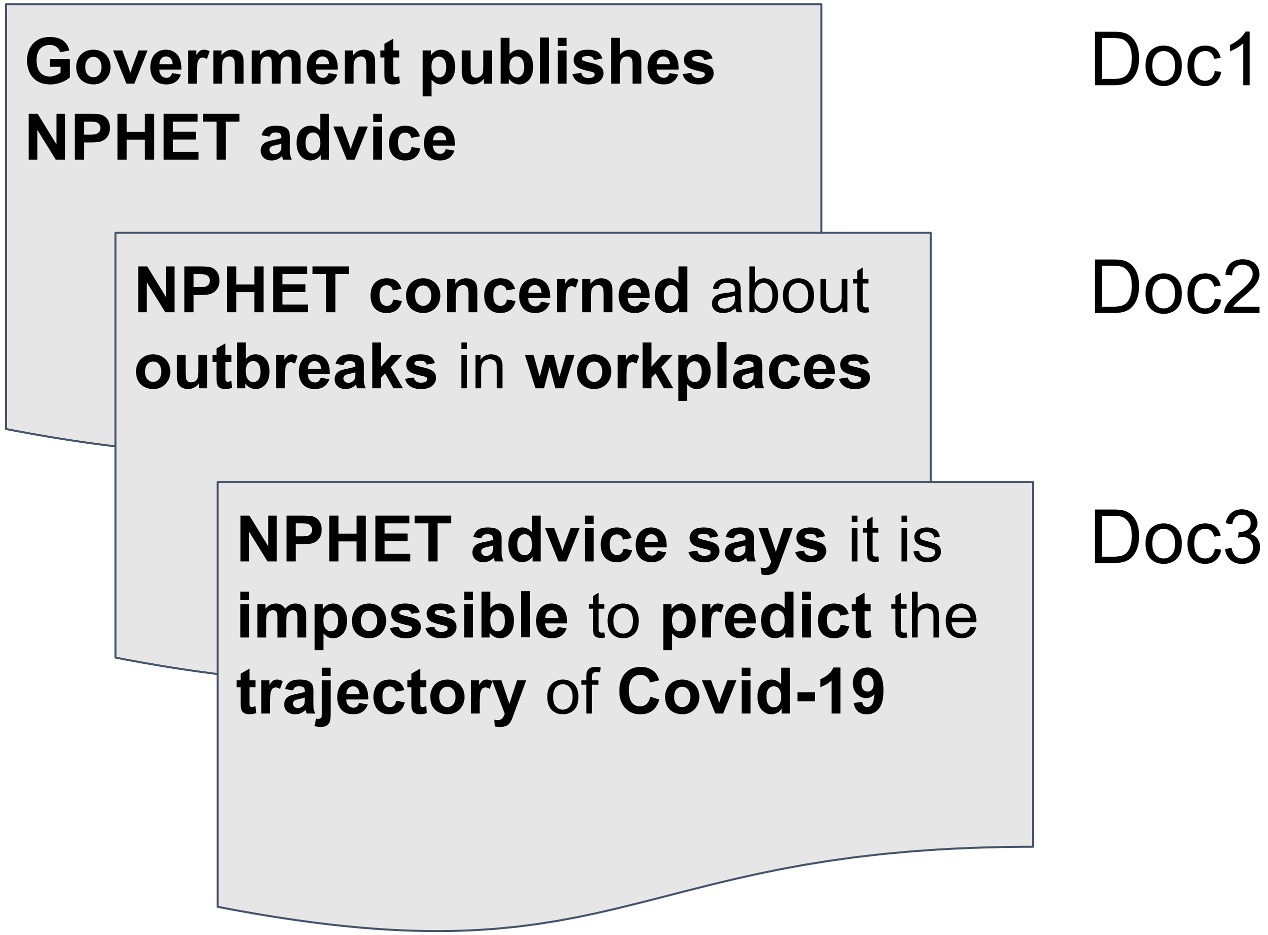
Doc1: Government publishes NPHET advice

Doc2: NPHET concerned about outbreaks in workplaces

Doc3: NPHET advice says it is impossible to predict the trajectory of Covid-19

Index Terms

Content words (nouns, verbs, adjectives), ignoring stopwords (about, in, it, is, ...)



Government publishes NPHET advice

Doc1

NPHET concerned about **outbreaks** in **workplaces**

Doc2

NPHET advice says it is **impossible** to **predict** the **trajectory** of **Covid-19**

Doc3

Inverted Index

Term-Document matrix (alphabetically sorted)

Index Term	Document ID
advice	1,3
concerned	2
Covid-19	3
Government	1
impossible	3
NPHET	1,2,3
outbreaks	2
predict	3
publishes	1
says	3
trajectory	3
workplaces	2

Document Vectors

	T1,T2,T3,T4,…,T11, T12
Doc1	[1,0,0,1,0,1,0,0,1,0,0,0]
Doc2	[0,1,0,0,0,1,1,0,0,0,0,1]
Doc3	[1,0,1,0,1,1,0,1,0,1,1,0]

Index Term	Term ID	Document ID
advice	T1	1,3
concerned	T2	2
Covid-19	T3	3
Government	T4	1
impossible	T5	3
NPHET	T6	1,2,3
outbreaks	T7	2
predict	T8	3
publishes	T9	1
says	T10	3
trajectory	T11	3
workplaces	T12	2

Document Retrieval

Doc1 [1,0,0,1,0,1,0,0,1,0,0,0]

Doc2 [0,1,0,0,0,1,1,0,0,0,0,1]

Doc3 [1,0,1,0,1,1,0,1,0,1,1,0]

What is the NPHET Government advice?

Index Term	Term ID	Document ID
advice	T1	1,3
concerned	T2	2
Covid-19	T3	3
Government	T4	1
impossible	T5	3
NPHET	T6	1,2,3
outbreaks	T7	2
predict	T8	3
publishes	T9	1
says	T10	3
trajectory	T11	3
workplaces	T12	2

Document Retrieval

Doc1 [1,0,0,1,0,1,0,0,1,0,0,0]

Doc2 [0,1,0,0,0,1,1,0,0,0,0,1]

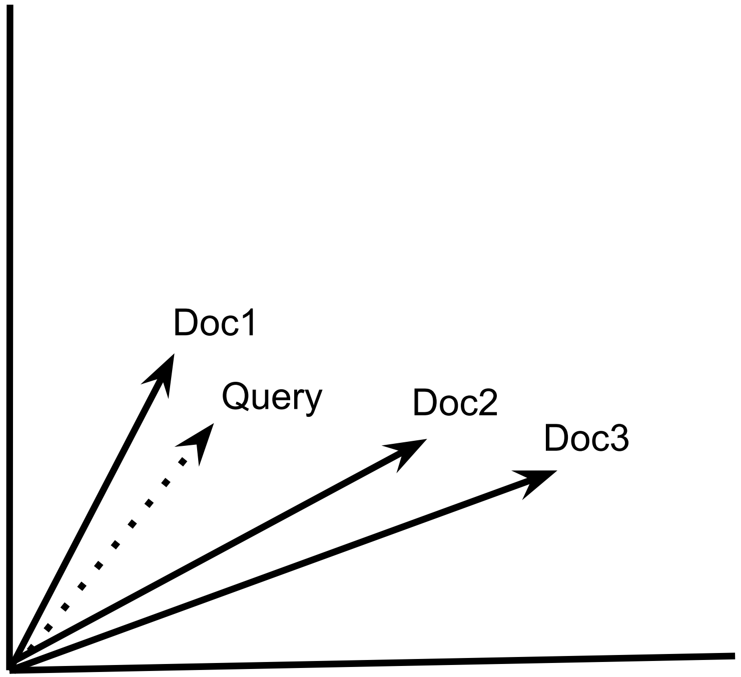
Doc3 [1,0,1,0,1,1,0,1,0,1,1,0]

*What is the **NPHET Government advice**?*

Query [1,0,0,1,0,1,0,0,0,0,0,0]

Index Term	**Term ID**	**Document ID**
advice	T1	1,3
concerned	T2	2
Covid-19	T3	3
Government	T4	1
impossible	T5	3
NPHET	T6	1,2,3
outbreaks	T7	2
predict	T8	3
publishes	T9	1
says	T10	3
trajectory	T11	3
workplaces	T12	2

Vector Space



Government publishes NPHET advice

Doc1 [1,0,0,1,0,1,0,0,1,0,0,0]

*What is the **NPHET Government advice**?*

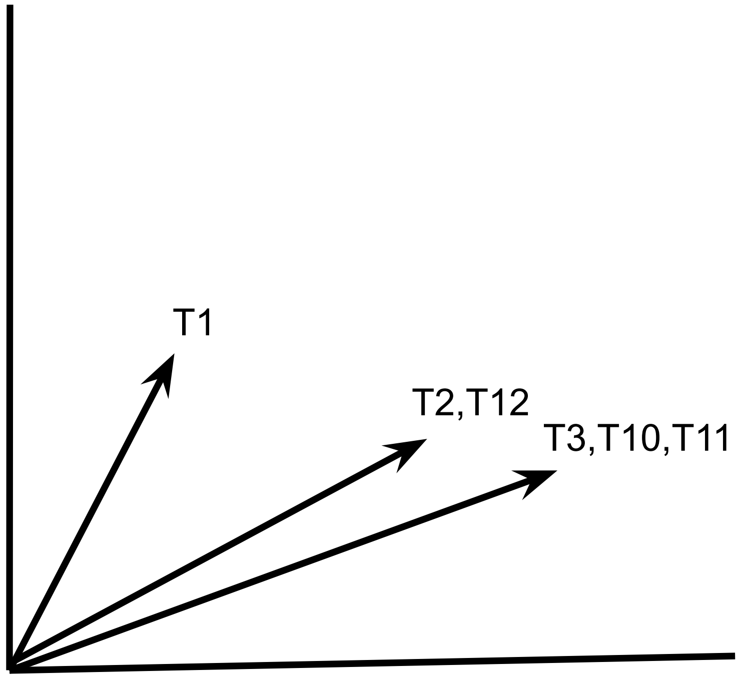
Query [1,0,0,1,0,1,0,0,0,0,0,0]

Term Vectors

	Doc1,Doc2, Doc3
T1	[1,0,1]
T2	[0,1,0]
T3	[0,0,1]
….	
T10	[0,0,1]
T11	[0,0,1]
T12	[0,1,0]

Index Term	**Term ID**	**Document ID**
advice	T1	1,3
concerned	T2	2
Covid-19	T3	3
Government	T4	1
impossible	T5	3
NPHET	T6	1,2,3
outbreaks	T7	2
predict	T8	3
publishes	T9	1
says	T10	3
trajectory	T11	3
workplaces	T12	2

Vector Space



T1	[1,0,1]
T2	[0,1,0]
T3	[0,0,1]
….	
T10	[0,0,1]
T11	[0,0,1]
T12	[0,1,0]

Latent Semantics

Represent index terms by their distribution over documents

Index **terms with similar meaning will show similarity in distributions**

‘Latent Semantic Analysis’ (LSA) or ‘Latent Semantic Indexing’ (LSI)

(Deerwester et al., 1990)

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), pp.391-407.

Overview

Vector Space Model

Distributional Semantics

Context in Distributional Representations

Word Embeddings

Distributional Hypothesis

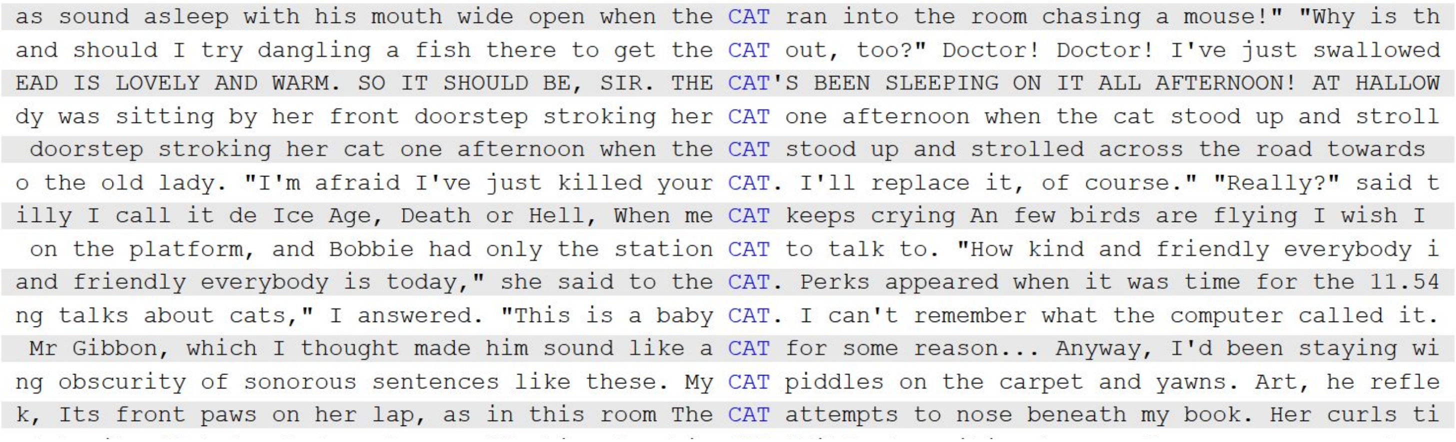
“you shall know a word by the company it keeps” - (Firth 1957)

cat and *dog* occur in similar contexts, therefore similar in meaning

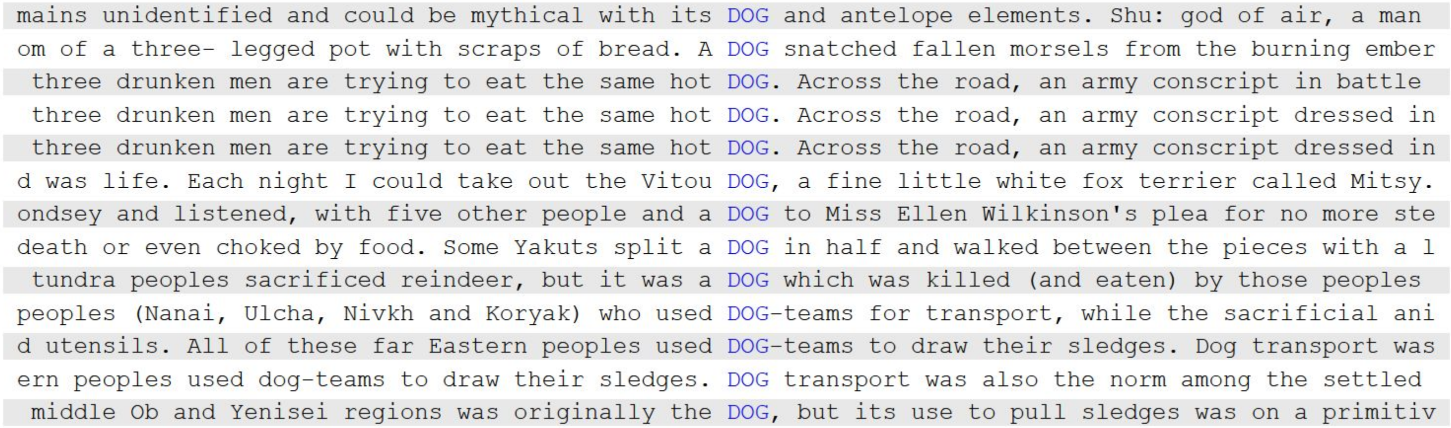
Distributional Semantics

Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis: 1–32

Distributional Semantics - Context Words



```
as sound asleep with his mouth wide open when the CAT ran into the room chasing a mouse!" "Why is th
and should I try dangling a fish there to get the CAT out, too?" Doctor! Doctor! I've just swallowed
EAD IS LOVELY AND WARM. SO IT SHOULD BE, SIR. THE CAT'S BEEN SLEEPING ON IT ALL AFTERNOON! AT HALLOW
dy was sitting by her front doorstep stroking her CAT one afternoon when the cat stood up and stroll
 doorstep stroking her cat one afternoon when the CAT stood up and strolled across the road towards
o the old lady. "I'm afraid I've just killed your CAT. I'll replace it, of course." "Really?" said t
illy I call it de Ice Age, Death or Hell, When me CAT keeps crying An few birds are flying I wish I
 on the platform, and Bobbie had only the station CAT to talk to. "How kind and friendly everybody i
and friendly everybody is today," she said to the CAT. Perks appeared when it was time for the 11.54
ng talks about cats," I answered. "This is a baby CAT. I can't remember what the computer called it.
 Mr Gibbon, which I thought made him sound like a CAT for some reason... Anyway, I'd been staying wi
ng obscurity of sonorous sentences like these. My CAT piddles on the carpet and yawns. Art, he refle
k, Its front paws on her lap, as in this room The CAT attempts to nose beneath my book. Her curls ti
```



```
mains unidentified and could be mythical with its DOG and antelope elements. Shu: god of air, a man
om of a three- legged pot with scraps of bread. A DOG snatched fallen morsels from the burning ember
 three drunken men are trying to eat the same hot DOG. Across the road, an army conscript in battle
 three drunken men are trying to eat the same hot DOG. Across the road, an army conscript dressed in
 three drunken men are trying to eat the same hot DOG. Across the road, an army conscript dressed in
d was life. Each night I could take out the Vitou DOG, a fine little white fox terrier called Mitsy.
ondsey and listened, with five other people and a DOG to Miss Ellen Wilkinson's plea for no more ste
death or even choked by food. Some Yakuts split a DOG in half and walked between the pieces with a l
 tundra peoples sacrificed reindeer, but it was a DOG which was killed (and eaten) by those peoples
peoples (Nanai, Ulcha, Nivkh and Koryak) who used DOG-teams for transport, while the sacrificial ani
d utensils. All of these far Eastern peoples used DOG-teams to draw their sledges. Dog transport was
ern peoples used dog-teams to draw their sledges. DOG transport was also the norm among the settled
 middle Ob and Yenisei regions was originally the DOG, but its use to pull sledges was on a primitiv
```

Concordance examples generated with https://www.lextutor.ca/conc/

Target and Context Words in a Corpus

Corpus

The cat lies on the mat.

The dog lies on the floor.

The cat sits near the door.

The dog lies near the door

Targets

cat, dog

Vocabulary of context words (V)

the, lies, on, mat, floor, sits, near, door

Co-occurrence Matrix (Context N=1)

Corpus

The cat lies *on the mat.*

The dog lies on the floor.

The cat sits near the door.

The dog lies near the door

Targets

cat, dog

Vocabulary of context words (V)

the, lies, on, mat, floor, sits, near, door

V	*the*	*lies*	*on*	*mat*	*floor*	*sits*	*near*	*door*
cat	1	1						
dog								

Co-occurrence Matrix (Context N=1)

Corpus

The cat lies on the mat.

The dog lies on the floor.

***The cat sits** near the door.*

The dog lies near the door

Targets

cat, dog

Vocabulary of context words (V)

the, lies, on, mat, floor, sits, near, door

V	*the*	*lies*	*on*	*mat*	*floor*	*sits*	*near*	*door*
cat	2	1				1		
dog								

Co-occurrence Matrix (Context N=1)

Corpus

The cat lies on the mat.

***The dog lies** on the floor.*

The cat sits near the door.

The dog lies near the door

Targets

cat, dog

Vocabulary of context words (V)

the, lies, on, mat, floor, sits, near, door

V	***the***	***lies***	***on***	***mat***	***floor***	***sits***	***near***	***door***
cat	2	1				1		
dog	1	1						

Co-occurrence Matrix (Context N=1)

Corpus

The cat lies on the mat.

The dog lies on the floor.

The cat sits near the door.

***The dog lies** near the door*

Targets

cat, dog

Vocabulary of context words (V)

the, lies, on, mat, floor, sits, near, door

V	***the***	***lies***	***on***	***mat***	***floor***	***sits***	***near***	***door***
cat	2	1				1		
dog	2	2						

Co-occurrence Matrix (Context N=1)

Corpus

The cat lies *on the mat.*

The dog lies *on the floor.*

The cat sits *near the door.*

The dog lies *near the door*

Targets

cat, dog

Vocabulary of context words (V)

the, lies, on, mat, floor, sits, near, door

V	***the***	***lies***	***on***	***mat***	***floor***	***sits***	***near***	***door***
cat	2	1	0	0	0	1	0	0
dog	2	2	0	0	0	0	0	0

Word Vectors for *cat, dog*

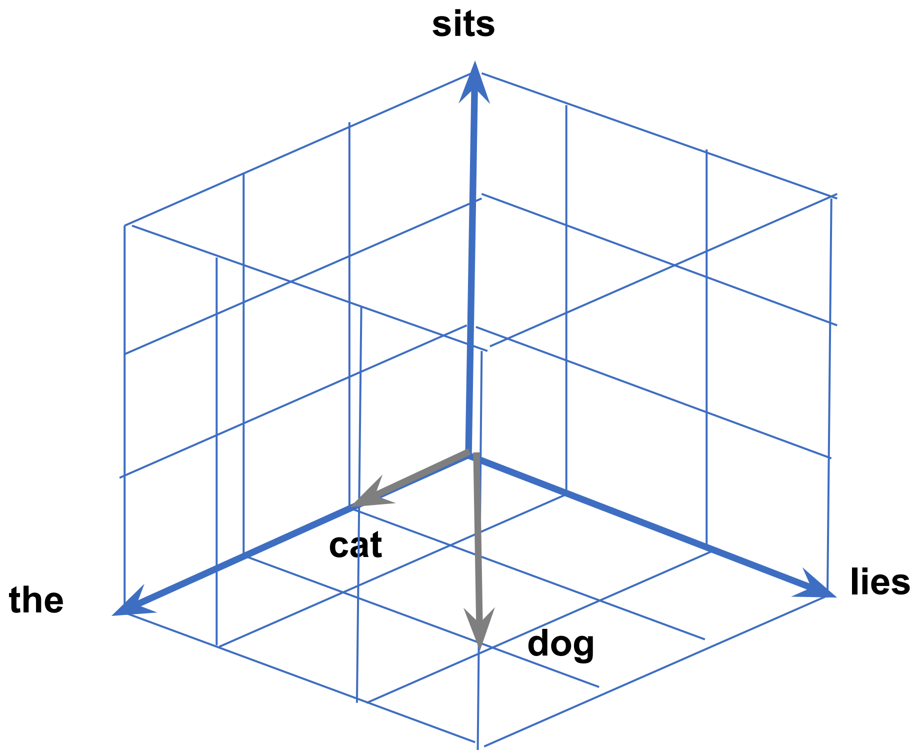
V	***the***	***lies***	***on***	***mat***	***floor***	***sits***	***near***	***door***
cat	2	1	0	0	0	1	0	0
dog	2	2	0	0	0	0	0	0

with 3 Non-Zero Dimensions

V	*the*	*lies*	*on*	*mat*	*floor*	*sits*	*near*	*door*
cat	2	1	0	0	0	1	0	0
dog	2	2	0	0	0	0	0	0

Notice many zeros (**sparse vector**)

Vector Space for *cat, dog* with 3 Dimensions



V	*the*	*lies*	*on*	*mat*	*floor*	*sits*	*near*	*door*
cat	2	1	0	0	0	1	0	0
dog	2	2	0	0	0	0	0	0

Distributional Similarity

Compute similarity of two words using their vector representations

Compute cosine of angle between two vectors

Cosine similarity (or 'cosine distance')

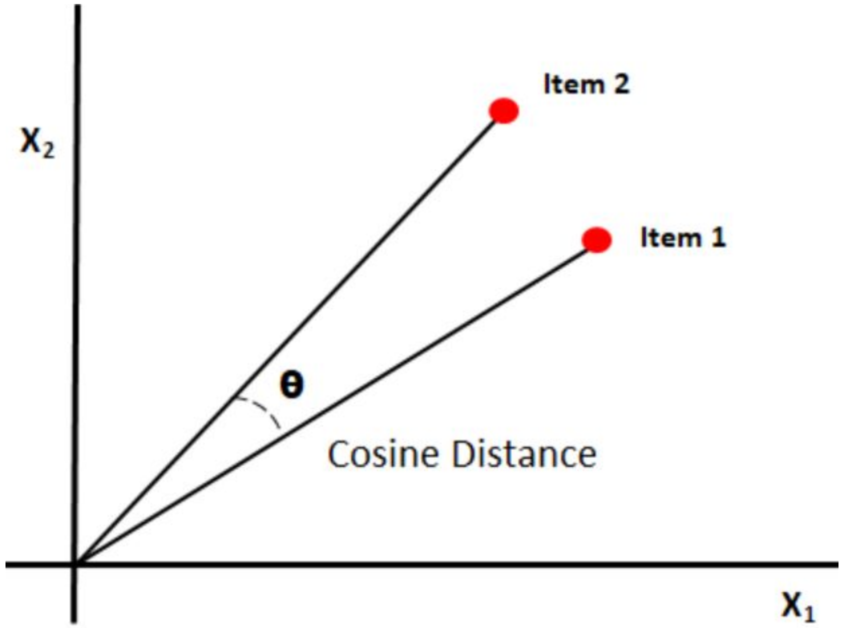


Image from https://www.oreilly.com/library/view/statistics-for-machine

Cosine Similarity

$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$\text{dot-product}(\vec{v},\vec{w}) = \vec{v}\cdot\vec{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

$v_i\, w_i$ *is the count for word v,w in context i*

$$|\vec{v}| = \sqrt{\sum_{i=1}^{N} v_i^2} \qquad \text{vector length}$$

Cosine Similarity - Example

cat	5	1	1	1	0
dog	3	1	1	0	1
child	3	1	1	0	1

cos (cat, dog) *= 15+1+1+0+0 / (√ 25+1+1+1+0 * √ 9+1+1+0+1)*
*= 17 / (√ 28 * √ 12)*
*= 17 / (5.29 * 3.46)*
= 0.93

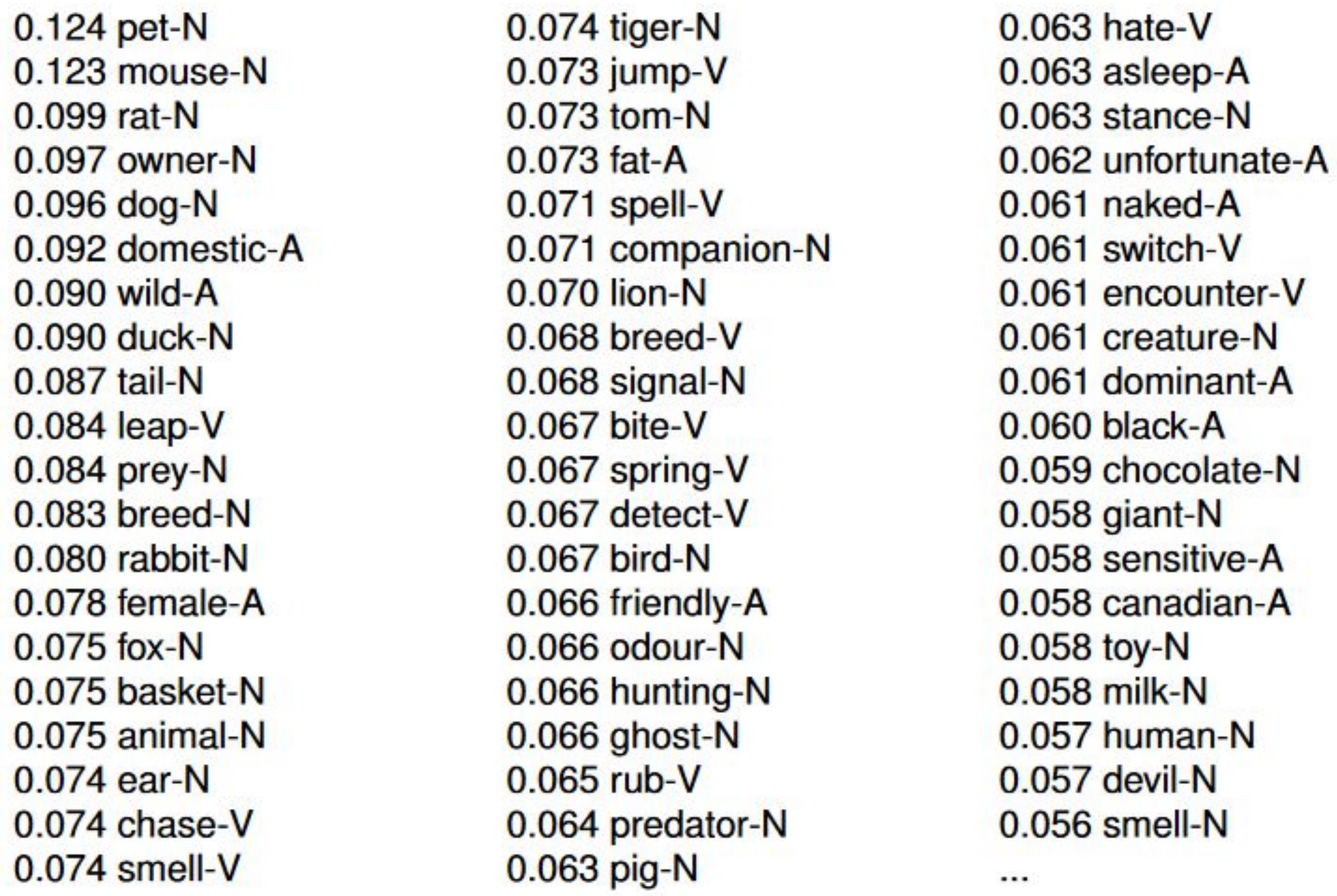
cos (cat, child) *= 15+1+1+0+0 / (√ 25+1+1+1+0 * √ 9+1+1+0+1)*
*= 17 / (√ 28 * √ 12)*
*= 17 / (5.29 * 3.46)*
= 0.93

cos (dog, child) *= 9+1+1+0+1 / (√ 9+1+1+0+1 * √ 9+1+1+0+1)*
*= 12 / (√ 12 * √ 12)*
*= 12 / (3.46 * 3.46)*
= 1.00

Word Similarity & Relatedness

Similarity of *cat* to other words in the British National Corpus

Notice ‘relatedness’ instead of ‘similarity’ in many cases

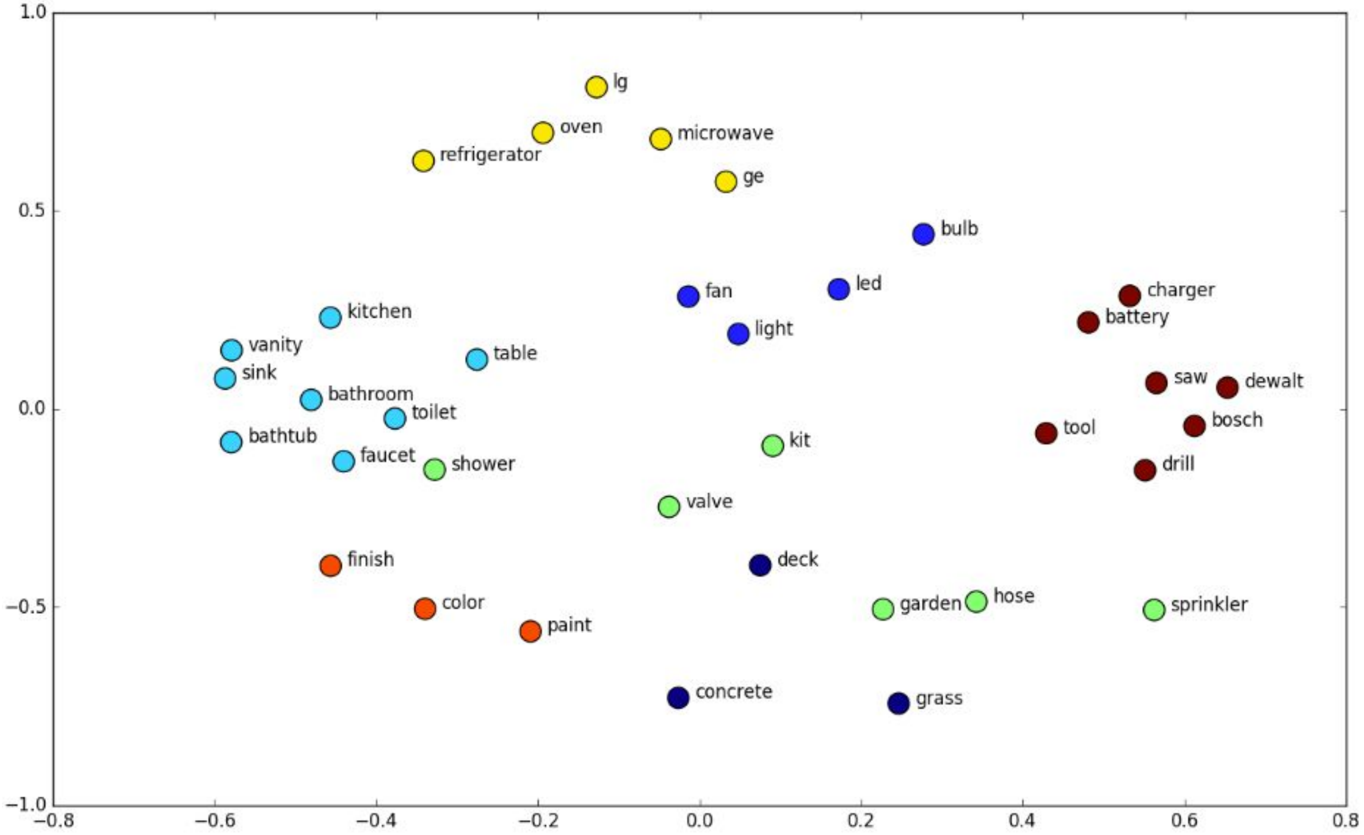


0.124	pet-N
0.123	mouse-N
0.099	rat-N
0.097	owner-N
0.096	dog-N
0.092	domestic-A
0.090	wild-A
0.090	duck-N
0.087	tail-N
0.084	leap-V
0.084	prey-N
0.083	breed-N
0.080	rabbit-N
0.078	female-A
0.075	fox-N
0.075	basket-N
0.075	animal-N
0.074	ear-N
0.074	chase-V
0.074	smell-V
0.074	tiger-N
0.073	jump-V
0.073	tom-N
0.073	fat-A
0.071	spell-V
0.071	companion-N
0.070	lion-N
0.068	breed-V
0.068	signal-N
0.067	bite-V
0.067	spring-V
0.067	detect-V
0.067	bird-N
0.066	friendly-A
0.066	odour-N
0.066	hunting-N
0.066	ghost-N
0.065	rub-V
0.064	predator-N
0.063	pig-N
0.063	hate-V
0.063	asleep-A
0.063	stance-N
0.062	unfortunate-A
0.061	naked-A
0.061	switch-V
0.061	encounter-V
0.061	creature-N
0.061	dominant-A
0.060	black-A
0.059	chocolate-N
0.058	giant-N
0.058	sensitive-A
0.058	canadian-A
0.058	toy-N
0.058	milk-N
0.057	human-N
0.057	devil-N
0.056	smell-N
...	

Example from http://disi.unitn.it/~bernardi/Courses/CL/Slides/ds_semantics.pdf

Visualizing Word Similarity

Principal Component Analysis (PCA) to project into 2D space



https://en.wikipedia.org/wiki/Principal_component_analysis

Overview

Vector Space Model

Distributional Semantics

Context in Distributional Representations

Word Embeddings

Context

Context **window**

document, sentence, phrase, word sequence

Context **content**

stopwords, specific part-of-speech, syntactic dependencies, etc.

Context **weights**

measure of co-occurrence strength between target and context words

Context Window - Document

*As carnivores, **cat**s eat herbivores like mice and voles. Since these herbivores eat grass, the free-roaming **cat** ingests any grass inside its prey's stomach and intestines. Grass provides the **cat** with roughage, which helps move hair ingested during grooming along the gastrointestinal tract.*

*A pet **cat** has been diagnosed with Covid-19, the first case of animal infection with coronavirus in the UK, the government has confirmed. The animal, which is said to have only experienced mild symptoms, is not believed to been involved in transmitting the disease to its owners or other humans and animals.*

Context Window - Sentence

*As carnivores, **cat**s eat herbivores like mice and voles.*

*Since these herbivores eat grass, the free-roaming **cat** ingests any grass inside its prey's stomach and intestines.*

*Grass provides the **cat** with roughage, which helps move hair ingested during grooming along the gastrointestinal tract.*

*A pet **cat** has been diagnosed with Covid-19, the first case of animal infection with coronavirus in the UK, the government has confirmed.*

Context Window - Word Sequence (N=2)

as carnivores,	***cats***	*eat herbivores*
the free-roaming	***cat***	*ingests any*
provides the	***cat***	*with roughage*
a pet	***cat***	*has been*

Context Content - Remove Stopwords

as carnivores,	***cat**s*	*eat herbivores*
the free-roaming	***cat***	*ingests any*
provides the	***cat***	*with roughage*
a pet	***cat***	*has been*

Context Content - PoS Tagging

as carnivores-N,	***cat**s*	*eat-V herbivores-N*
the free-roaming-Adj	***cat***	*ingests-V any*
provides-V the	***cat***	*with roughage-N*
a pet-Adj	***cat***	*has been*

Context Content - Syntactic Dependencies

as carnivores-N-NMOD,	***cats***	*eat-V herbivores-N-DOBJ*
the free-roaming-Adj-AMOD	***cat***	*ingests-V any*
provides-V the	***cat***	*with roughage-N-POBJ*
a pet-Adj-AMOD	***cat***	*has been*

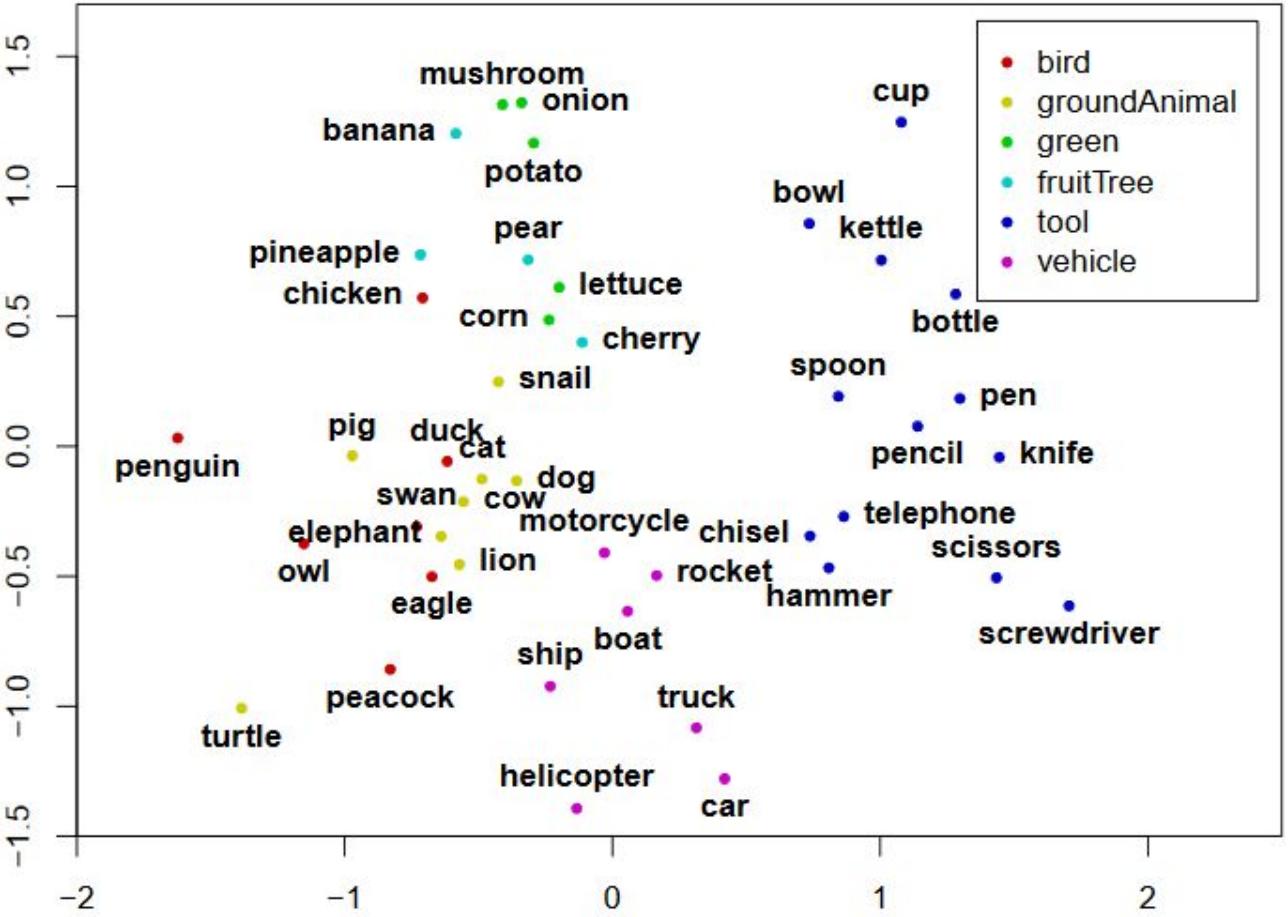
Syntactic Dependencies - Example

Similarity of nouns in syntactic dependency relation DOBJ with verbs in the BNC corpus

For example:

eat-V	*mushrooms-N-DOBJ*
eat-V	*onions-N-DOBJ*
eat-V	*potatoes-N-DOBJ*
use-V	*spoon-N-DOBJ*
use-V	*pen-N-DOBJ*

…



Example from slides by Stefan Evert https://esslli2016.unibz.it/wp-content/uploads/2015/10/dsm_tutorial_part1.slides.pdf

Context - Occurrence, Frequency, Weight

Binary

the value of vector $\vec{w}$ for dimension c is 1 if context *c* occurs with word *w* and 0 otherwise

Frequency

the value of vector $\vec{w}$ for dimension *c* is the number of times that context c occurs with word *w*

Weight (Relevance, Specificity)

the value of vector $\vec{w}$ for dimension *c* is a weight, expressing how relevant or specific context *c* is for word *w*

Examples

*As carnivores, a **cat** eats mice and voles but eats no grass.*

Binary - 1 or 0 depending if the context word occurs in this sentence

$\vec{w}$ {eats:1, mice:1, voles:1, sits:0}

Frequency - count of the context word in this sentence

$\vec{w}$ {eats:2, mice:1, voles:1, sits:0}

Weight - weight of the context word in a corpus of sentences

$\vec{w}$ {eats:0.95, mice:0.56, voles:0.01, sits:0.80}

Context Weight - TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF measures how important a Term is in a Document where it occurs, in comparison to the overall occurrence of this Term in a set of Documents.

TF-IDF is mostly used in IR to put weights on index terms but can be used also in distributional semantics, using words (Terms) and sentences (Docs) here:

TF-IDF(w) = TF(w) * IDF(w)

TF(w) = frequency of word w in sentence s

IDF(w) = log (all sentences / all sentences with w)

Context Weight - PMI

Pointwise Mutual Information (PMI)

PMI is a correlation or association measure that quantifies the likelihood of co-occurrence of two independent events. In distributional semantics we use PMI to measure the likelihood of co-occurrence of two words across a corpus:

PMI (w_1, w_2) = log (P(w_1, w_2) / P(w_1) * P(w_2))

P(w_1) P(w_2) = probability (relative frequency) of w_1 w_2 occurring in C

= frequency of w_1 w_2 occurring in C / all words W in C

P(w_1, w_2) = probability of w_1 co-occurring with w_2 in corpus C

PMI - Example

Corpus *C* with size (all words) = 24

*The **cat lies** on the mat. The **cat sits** near the door.*
*The **dog lies** on the floor. The **dog lies** near the door*

Frequency of target words (w_1)

cat:2, dog:2

Frequency of context words (w_2)

lies:3, sits:1

Frequency of w_1 and w_2 co-occurring

cat lies:1, cat sits:1, dog sits:0, dog lies:2

PMI - Example

$\boldsymbol{PMI\ (w_1, w_2) \quad = \log_2 (\ P(w_1, w_2)\ /\ P(w_1)\ *\ P(w_2))}$

$PMI\ (cat, sits) \quad = \log_2 (P(cat, sits) / (P(cat) * P(sits)))$

$= \log_2 (1/24 / (2/24 * 1/24))$

$= \log_2 (0.042 / (0.083 * 0.042))$

$= \log_2 (0.042 / 0.004)$

$= \log_2 (10.5) = \mathbf{1.02}$

	lies	***sits***
cat		1.02
dog		

PMI - Example

$$PMI\ (w_1, w_2) \quad = \log_2 (\ P(w_1, w_2)\ /\ P(w_1)\ *\ P(w_2))$$

	lies	***sits***
cat	0.61	1.02
dog	0.90	0

Distributional Semantic Model

	get	see	use	hear	eat	kill
knife	0.027	-0.024	0.206	-0.022	-0.044	-0.042
cat	0.031	0.143	-0.243	-0.015	-0.009	0.131
dog	-0.026	0.021	-0.212	0.064	0.013	0.014
boat	-0.022	0.009	-0.044	-0.040	-0.074	-0.042
cup	-0.014	-0.173	-0.249	-0.099	-0.119	-0.042

Adapted from slides by Stefan Evert https://esslli2016.unibz.it/wp-content/uploads/2015/10/dsm_tutorial_part1.slides.pdf

Overview

Vector Space Model

Distributional Semantics

Context in Distributional Representations

Word Embeddings

Word Embeddings

Generalization

Capture similarity on the level of word vectors rather than words

Prediction

Predict words and their contexts by use of language modeling

Dimensionality reduction

Optimize sparse vectors (with lots of zeros) into shorter dense vectors

Generalization

He drove his car on the road.

He drove his car on the street.

He drove his automobile on the street.

He steered his automobile on the street.

She steered his automobile across the street.

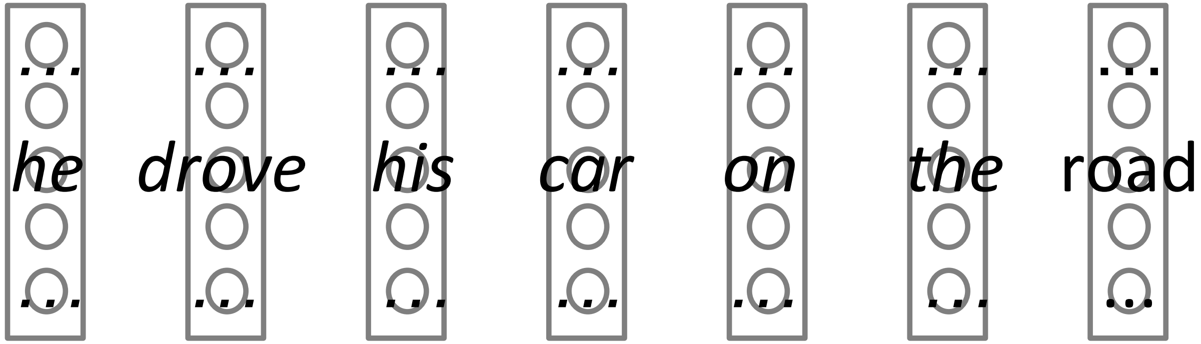
She steered her automobile across the street.

She steered her automobile across a street.

Generalization

…	…	…	…	…	…	…
he	*drove*	*his*	*car*	*on*	*the*	road
…	…	…	…	…	…	…

Word Embeddings (Vectors) instead of Words



From DSM to Word Embeddings

Co-occurrence matrix over vectors instead of words

	drive	***park***
car		
road		

	V(drive)	***V(park)***
V(car)		
V(road)		

Prediction

Predict co-occurrence contexts instead of counting as with DSMs

Use **language models** to assign a co-occurrence probability to a sequence of words (bigrams here)

$$P(w^{(1)}, w^{(2)}, \cdots, w^{(n)}) = \prod_{i=2}^{n} P(w^{(i)}|w^{(i-1)})$$

Formula from https://cs224d.stanford.edu/lecture_notes/LectureNotes1.pdf

Dimensionality Reduction

Reduce a high-dimensional vector space with long sparse vectors of many dimensions into **low-dimensional embeddings** of short dense vectors of limited dimensions that are expressive for a given task

One-Hot Vectors

Initialize high-dimensional vector space with **one-hot vectors**: sparse word representations over a vocabulary V with length |V|

For example:

<table>
<tr><th>V</th><th colspan="13">|V|</th></tr>
<tr><td><b><i>cat</i></b></td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td><b><i>dog</i></b></td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td><b>…</b></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

Word Embeddings

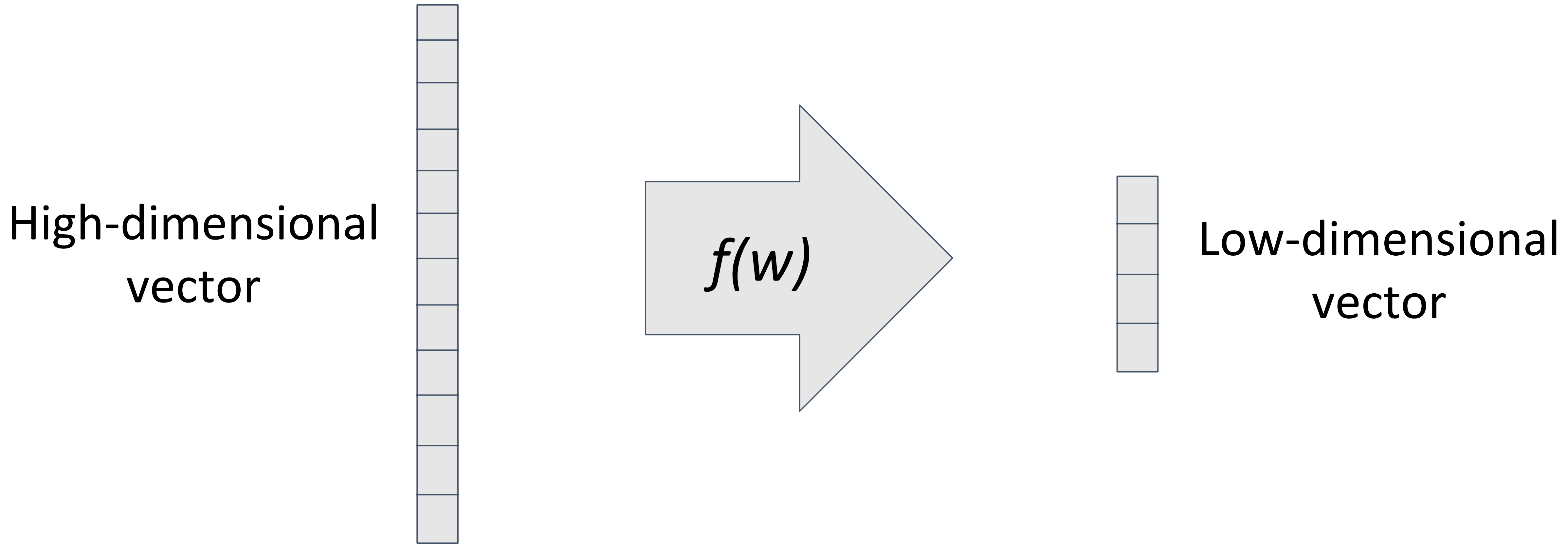
Initialize high-dimensional vector space with one-hot vectors over a vocabulary V derived from a **training corpus**

Count or predict (**language model**) co-occurrence for all words in V within a given context (e.g. bigrams) with co-occurrence taken as positive instance and as negative instance otherwise

Train a low-dimensional embedding space on a **classification task** that correctly classifies positive and negative co-occurrence instances

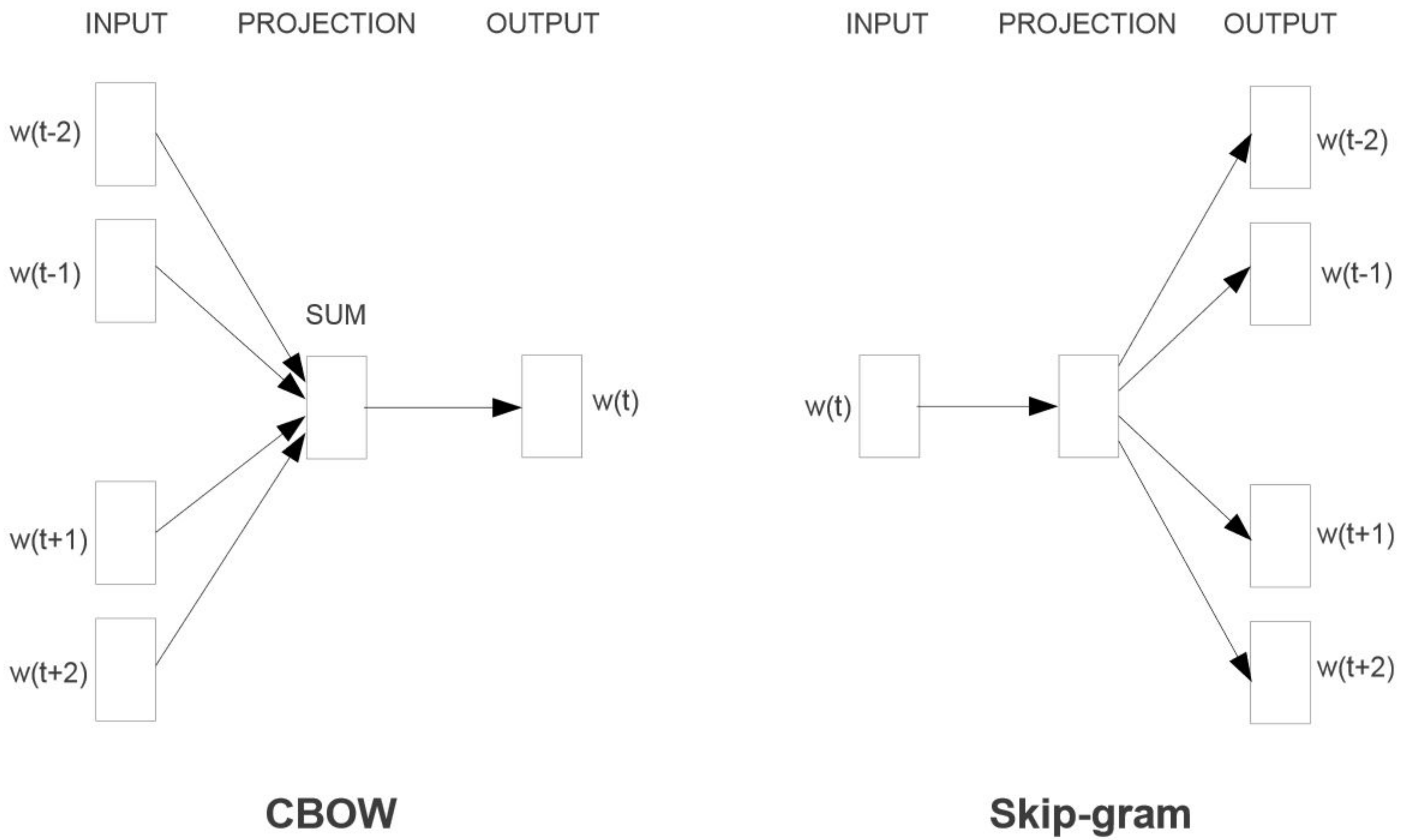
Results in a **word embedding** for this corpus, model and task

Word Embedding



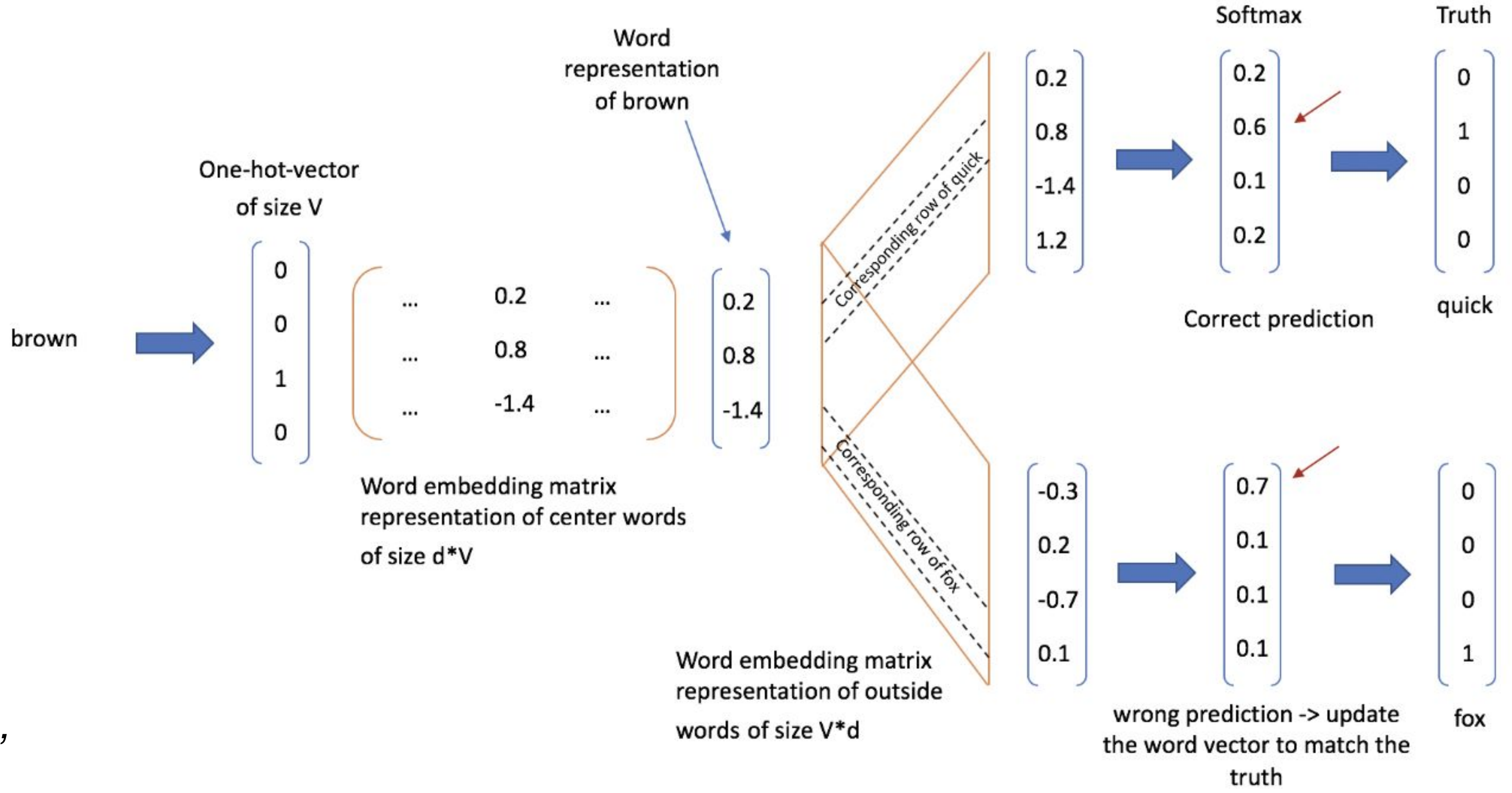
Adapted from a slide by John McCrae & Mihael Arcan - Advanced NLP, CT5121, NUI Galway

Implementation with Word2Vec



Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Implementation with Word2Vec - Skip-gram

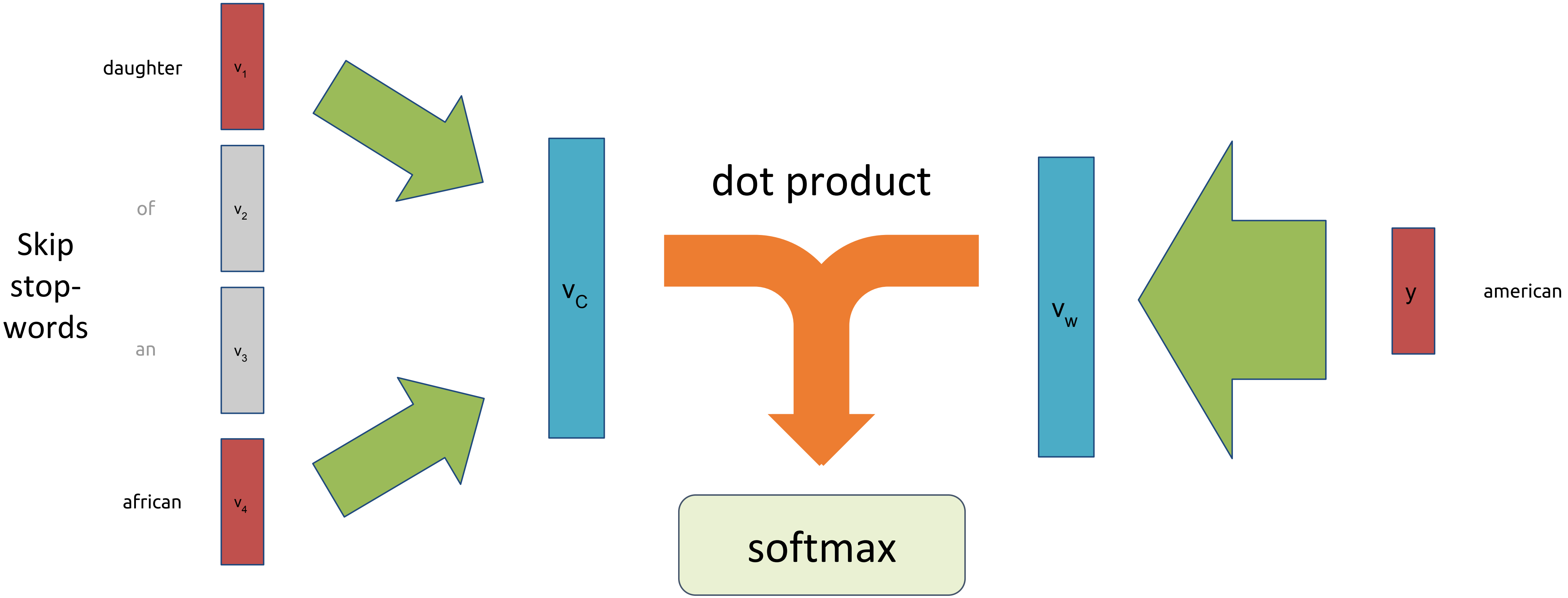


brown
One-hot-vector of size V
0
0
1
0
...
0.2
...
...
0.8
...
...
-1.4
...
Word embedding matrix representation of center words of size d*V
Word representation of brown
0.2
0.8
-1.4
Corresponding row of quick
Corresponding row of fox
Word embedding matrix representation of outside words of size V*d
0.2
0.8
-1.4
1.2
Softmax
0.2
0.6
0.1
0.2
Correct prediction
Truth
0
1
0
0
quick
-0.3
0.2
-0.7
0.1
0.7
0.1
0.1
0.1
wrong prediction -> update the word vector to match the truth
0
0
0
1
fox

*"the quick **brown** fox"*

Image from https://towardsdatascience.com/word-embeddings-intuition-and-some-maths-to-understand-end-to-end-skip-gram-model-cab57760c745

Implementation with Word2Vec - Skip-gram



*“she was the daughter of an african **american**”*

Adapted from a slide by John McCrae & Mihael Arcan - Advanced NLP, CT5121, NUI Galway

Pre-trained Models

A word embedding can be used as a pre-trained unsupervised model

Current best-performing NLP approaches combine pre-trained models with task-specific, supervised fine-tuning, aka as **‘transfer learning’**

Originally, widely used pre-trained models were based on Word2Vec

Many other pre-trained models have now been developed

Brief Overview of Pre-trained Models

Word2Vec, GloVe - local (Word2Vec) and global (GloVe) word-word contexts with single-layer neural network architecture (word embedding)

ULMFiT - 3-layer LSTM neural network architecture

ELMo - bidirectional LSTMs to model word-word contexts from left-to-right and right-to-left

Transformer - attention mechanisms instead of LSTM (RNN) to model long-range interaction between words across a sentence

BERT - transformer, bidirectional, next sentence prediction

T5 - transformer, text as input and output

For more details see https://www.paperswithcode.com/methods/category/language-models

Lab of this Week

Exercises in vector space models for NLP

NUI Galway
OÉ Gaillimh

QA