

You Are the Company That You Keep

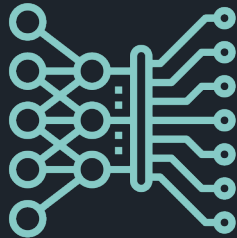
Converting Words to Numbers Using Text Embeddings

Key Takeaways

WHAT WORD
EMBEDDINGS
ARE



WHY THEY'RE
USEFUL IN
MACHINE
LEARNING



WHERE YOU
SHOULD BE
CAREFUL



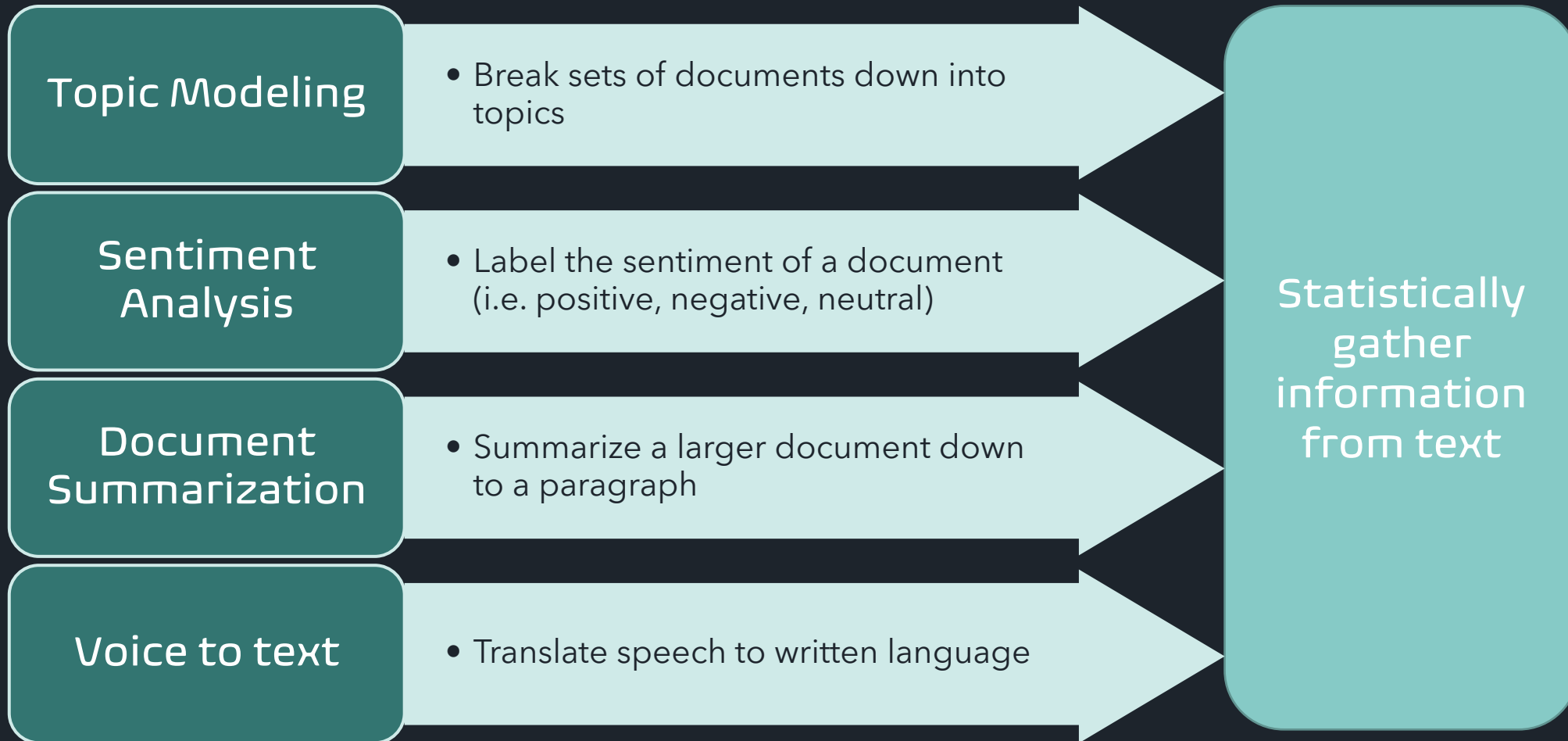
What is Natural Language Processing?

Computational
Linguistic



Machine Learning

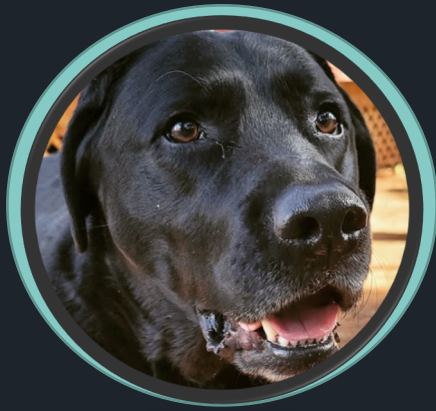
Why is it useful?



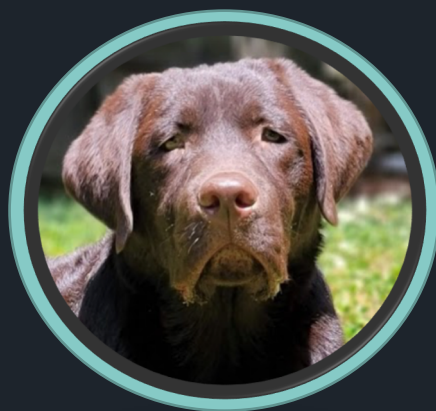
Meet the Team



Kara



Dash



Wally



Reggie

From words to math

Text Frequency via Count vectorization



	a	another	dash	dog	got	have	I	Kara	named	new	puppy	wally	we
I have a dog named Kara	1	0	0	1	0	1	1	1	1	0	0	0	0
I have another dog I named Dash	0	1	1	1	0	1	2	0	1	0	0	0	0
We got a new puppy named Wally	1	0	0	0	1	0	0	0	1	1	1	1	1

So what's the problem?

missing words in training set



	a	another	dash	dog	got	have	I	Kara	named	new	puppy	wally	we
I have a dog named Kara	1	0	0	1	0	1	1	1	1	0	0	0	0
I have another dog I named Dash	0	1	1	1	0	1	2	0	1	0	0	0	0
We got a new puppy named Wally	1	0	0	0	1	0	0	0	1	1	1	1	1

testing set

I also have a cat named Reggie	1	0	0	0	0	1	1	0	1	0	0	0	0
--	---	---	---	---	---	---	---	---	---	---	---	---	---

So what's the problem?

Limitations with labelled data



sentence label

I have a **dog** named Kara

positive

I have another **dog** I named Dash

positive

We got a new **puppy** named Wally

neutral

I also have a **cat** named Reggie

?

So what now? Word embeddings!



Article **Talk**

Cat

From Wikipedia, the free encyclopedia

This article is about the species that is commonly kept as a pet. For the cat family, see [Felidae](#). For other uses, see [Cat \(disambiguation\)](#) and [Cats \(disambiguation\)](#).

For technical reasons, "Cat #1" redirects here. For the album, see [Cat 1 \(album\)](#).

The **cat** (*Felis catus*) is a domestic species of small carnivorous mammal.^{[1][2]} It is the only domesticated species in the family [Felidae](#) and is often referred to as the domestic cat.^[4] A cat can either be a **house cat**, a **farm cat** or a **feral cat**; the latter ranges freely and avoids human contact.^[5] Domestic cats are valued by humans for their companionship and the [cat breeds](#) are recognized by various [cat registries](#).^[6]

The cat is similar in [anatomy](#) to the other felid species: it has a strong flexible body, quick [reflexes](#), sharp teeth and [retractable claws](#) adapted to killing small prey. Its [communication](#) includes [vocalizations](#) like meowing, purring, trilling, hissing, growling and grunting as well as [cat-specific body language](#). A [predator](#) that hunts small prey, it is a [social species](#). It can hear sounds too faint or too high in [frequency](#) for human ears, such as those made by [mice](#) and other small mammals.^[7] It secretes

Word embeddings infer word meaning
based on **context**, so we don't need to rely
on **labels**

What are embeddings?

Dog

[0.995, -0.833, 0.327, 0.945, -0.547, 0.421, 0.333]

What? How? What does that mean?

Mike's presentation is truly **vapid**. I've never been so uninterested in my life.

Humans learn a word without a definition
through **Context** clues

Types of Embeddings

DETERMINISTIC

Given the information I have available, I can deduce that the probability of x is $y\%$

PROBABILISTIC

I'm learning as I go and continuously improving my accuracy. Given what I've learned in my training, the probability of x is $y\%$

Probabilistic: Word2Vec

Learn to predict a **word** given **its context** (CBOW)

Mike's	presentation	is	truly	vapid	I've	never
been	so	uninterested	in	my	life	

Learn to predict a **context** given a **word** (Skip gram)

Mike's	presentation	is	truly	vapid	I've	never
been	so	uninterested	in	my	life	

Probabilistic: Word2Vec

Learn to predict a **word** given **its context** (CBOW)

Mike's	presentation	is	truly	vapid	I've	never
been	so	uninterested	in	my	life	

Learn to predict a **context** given a **word** (Skip gram)

Mike's	presentation	is	truly	vapid	I've	never
been	so	uninterested	in	my	life	

Deterministic: GloVe

The probability of **co-occurrence** of two words has the potential to encode some form of meaning

	Mike's	presentation	is	truly	vapid	I've	never	been	so	uninterested	in	my	life
Mike's		1											
presentation			1										
is		1		1									
truly			1		1								
vapid						1							
I've					1	1	1						
never						1		1					
been									1				
so										1			
uninterested											1		
in												1	
my													1
Life													

Benefits of Embeddings



UNSUPERVISED
ALGORITHM



DON'T NEED
LABELLED
DATA



TRAIN ON
MASSIVE
CORPUS OF
TEXT

So how can we tell these vectors have meaning?

A is to **B** as **C** is to **D**

So how can we tell these vectors have meaning?

$$A - B = C - D$$

So how can we tell these vectors have meaning?

$$D = C - A + B$$

Cosine Similarity and Analogies

Paris is to **France** as **Tokyo** is to ?

0.879 : 'Japan'

0.732 : 'Tokyo'

0.726 : 'Korea'

0.682 : 'Germany'

0.679 : 'Japanese'

Cosine Similarity and Analogies

Cow is to **Mammal** as **Snake** is to ?

```
0.710 : 'mammal'  
0.696 : 'reptile'  
0.649 : 'lizard'  
0.648 : 'fishes'  
0.640 : 'reptiles'
```

Proceed With Caution

The blind application of machine learning runs the **risk** of amplifying **biases** present in data.

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings, Bolukbasi et al. 2016

Thank you

Questions?

Mike Stackhouse

Mike.Stackhouse@atorusresearch.com

