NLP WITH MACHINE LEARNING

DAVID STRÖM

CADEC 2022.02.02 | CALLISTAENTERPRISE.SE

CALLISTA

AGENDA

- What is Natural Language Processing (NLP)
- What is Machine Learning (ML)
- Use ML for NLP
- An example, the Callista chatbot project

WHAT IS NLP?

WHAT IS NLP?

- The automatic manipulation of natural language by software
- Two main flavours:
  - Linguistic methods/models (e.g. WordNet)
  - Machine learning methods/models
- Purpose: *systems that can understand human language*

Alan Turing ★ 1912 † 1954

USE-CASES FOR NLP

- Advanced customer support / QnA (e.g. chatbots)
- Text categorisation / sentiment analysis
- Summarisation
- Translation

Also, make loads of do$h!

NLP CHALLENGES

LANGUAGES ARE VAST

… Companies try to ensure that customers will be happy with their product…

… our company want to ensure that the client is always satisfied with the product…

… Companies like ours need to ensure that the buyer is thoroughly satisfied with their product…

Picture by Minette Lontsie

LANGUAGES ARE AMBIGUOUS

- “The bank closes at 3 p.m.” vs. “The river bank was lined with sea-weed”
- “The passerby helped dog bite victim”
- The prime minister met with Ursula von der Leyen where she expressed a wish for more cooperation

Picture by Regeringskansliet

LANGUAGES ARE VERSATILE

- The cat is on the table
- The table is under the cat
- The table-cloth is on the table and the cat is sleeping on top of the table-cloth

Picture by TripAdvisor.com

LANGUAGES EVOLVE

WHAT IS MACHINE LEARNING?

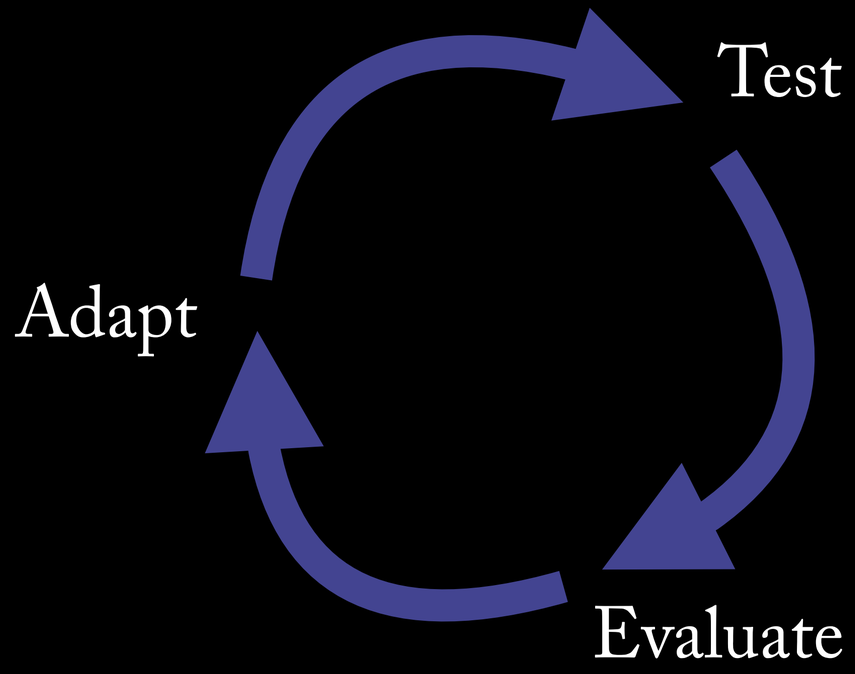
WHAT IS MACHINE LEARNING?

"Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so."

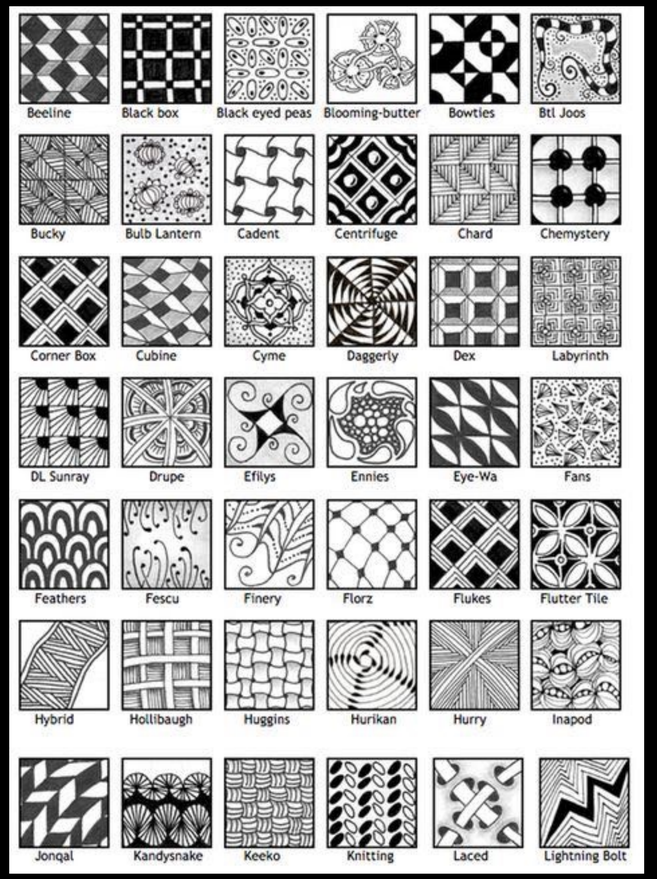
Source: Wikipedia

WHAT IS MACHINE LEARNING?

SELF-OPTIMISATION

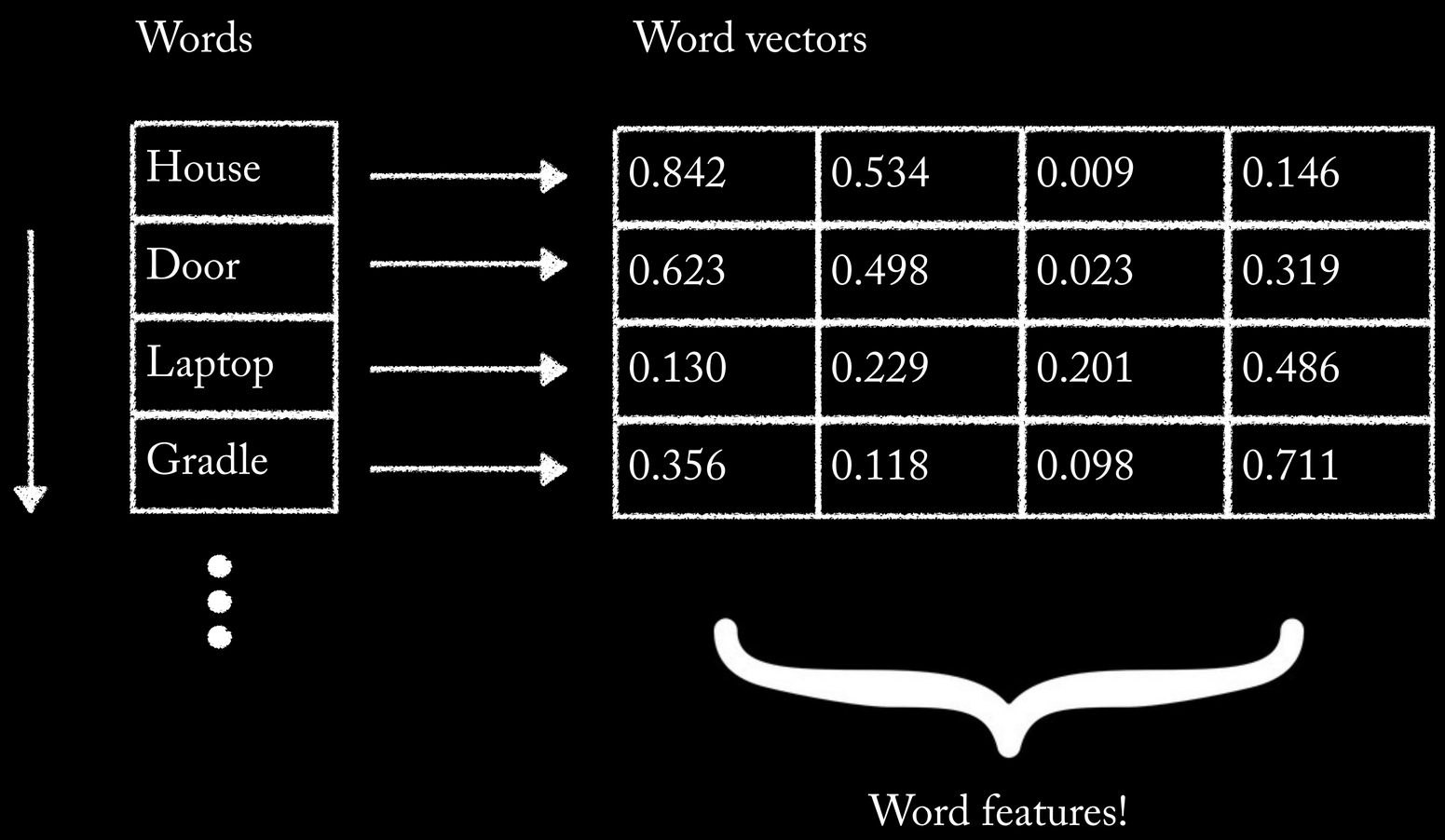


PATTERN RECOGNITION



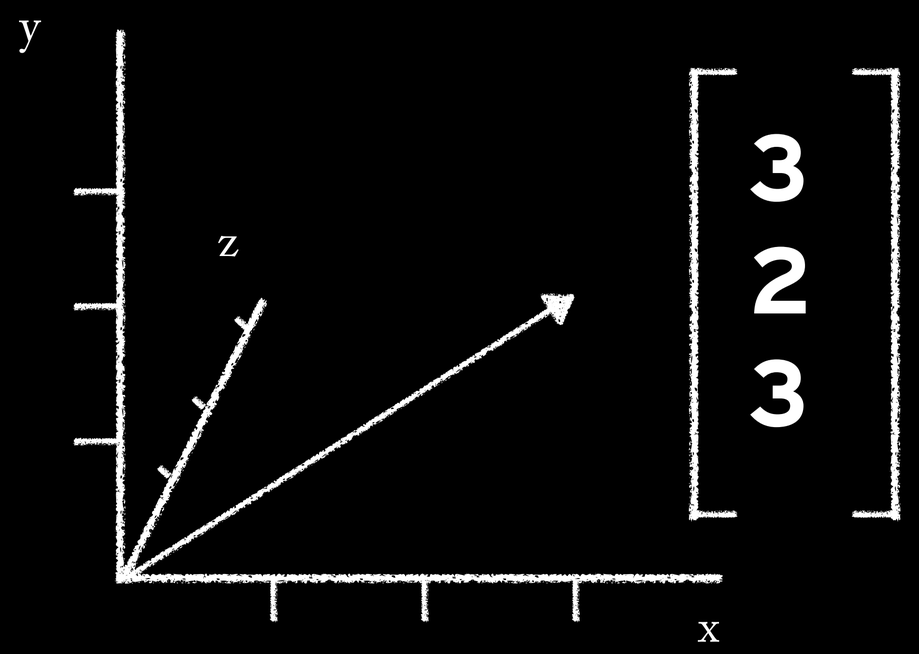
USE MACHINE LEARNING FOR NLP

WORD EMBEDDINGS / WORD VECTORS

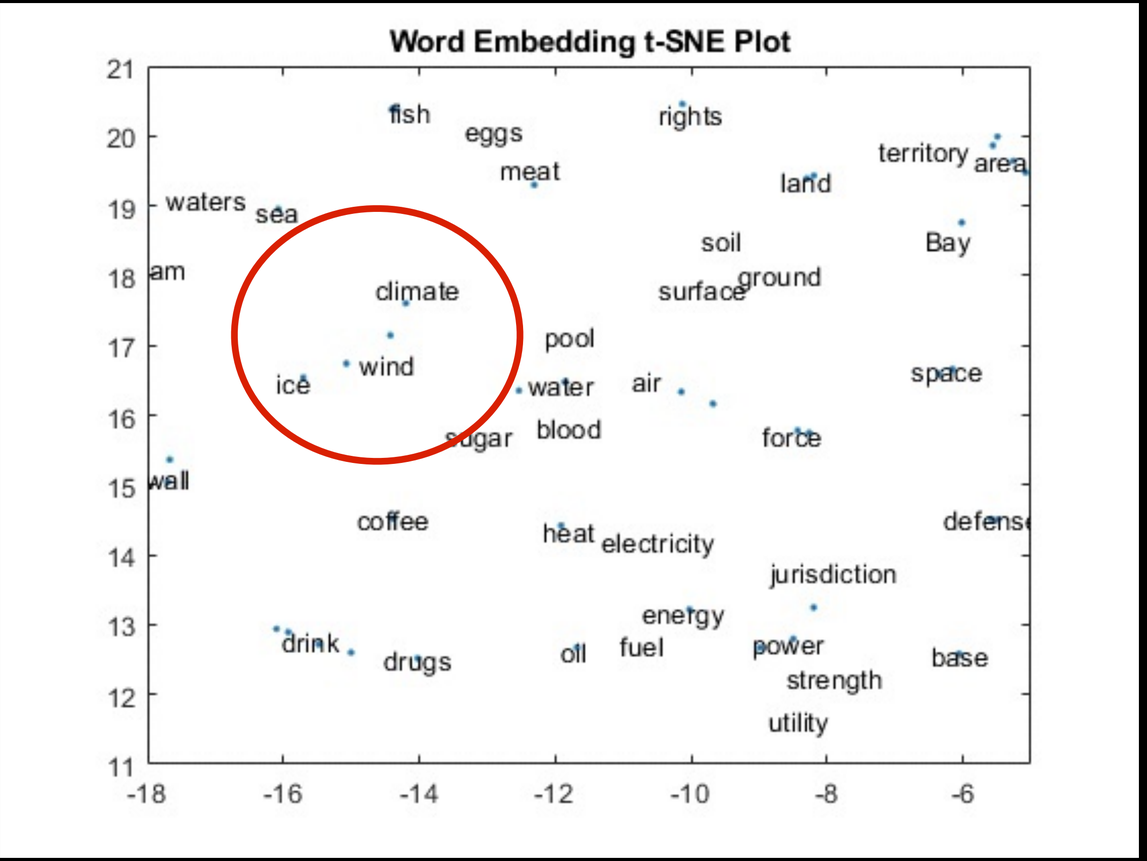


WORD VECTORS

- Being vectors, we can do certain mathematical operations on them
- A vector points to a position in vector space
- In ML we are generally not concerned with the direction of the vector

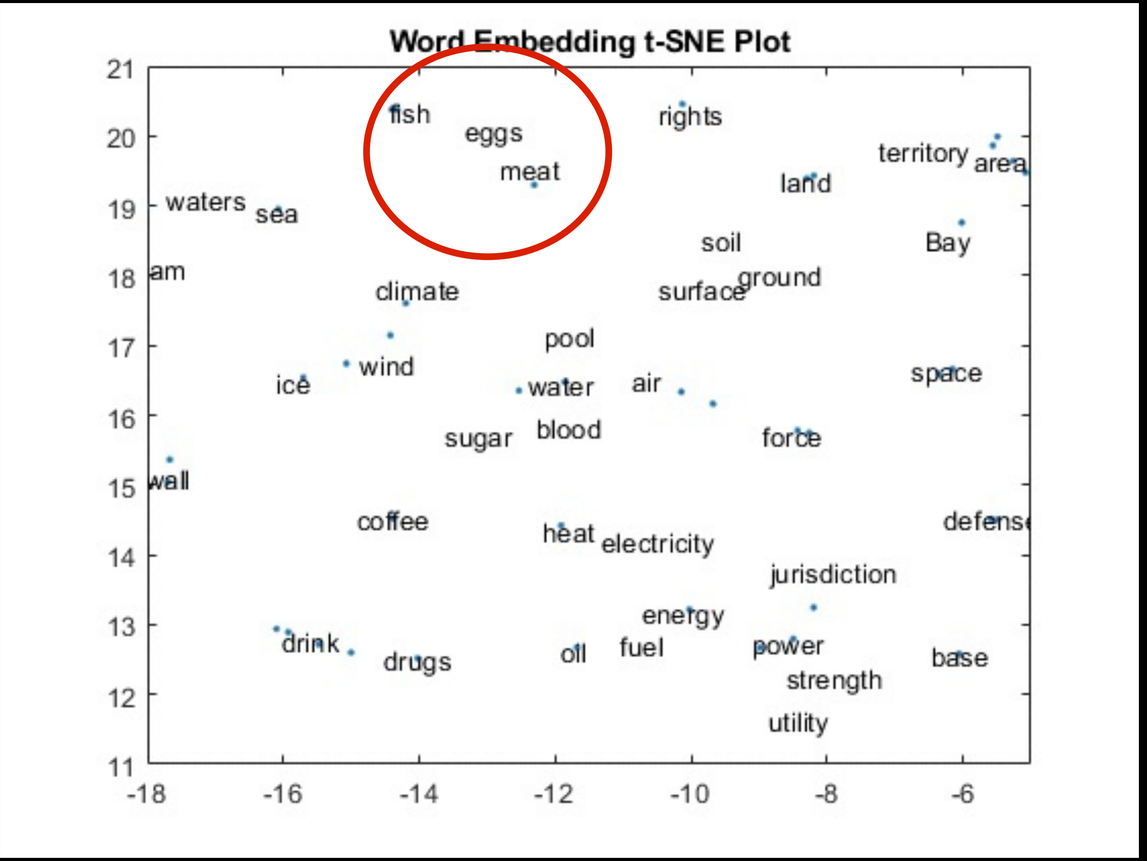


MACHINE LEARNING AND WORD VECTORS



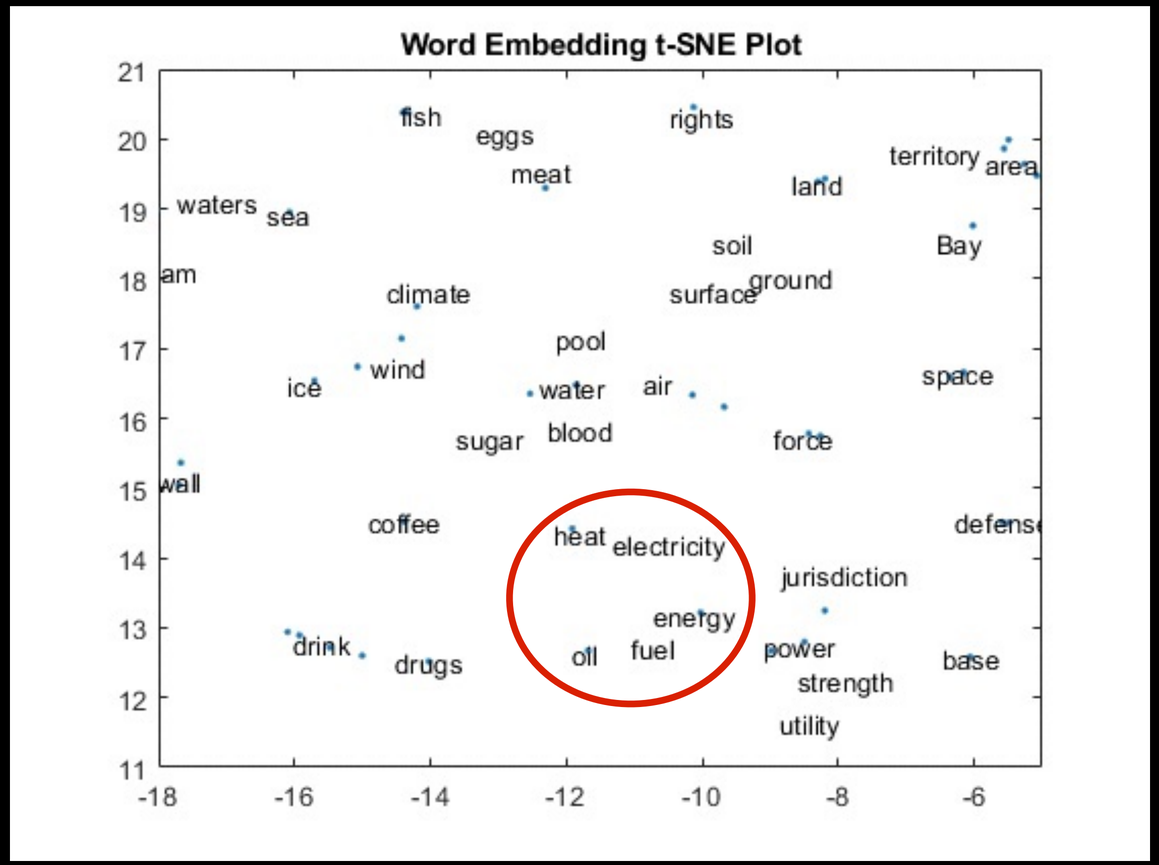
Word similarity in vector space using t-SNE plot

MACHINE LEARNING AND WORD VECTORS



Word similarity in vector space using t-SNE plot

MACHINE LEARNING AND WORD VECTORS



Word similarity in vector space using t-SNE plot

HOW CAN WE DO THIS?

“You shall know a word by the company it keeps”

- JOHN RUPERT FIRTH

KNOW A WORD BY THE COMPANY THAT IT KEEPS

… code necessary making our microservices publish monitoring data prometheus…

… requests are passed between microservices storage backends with messages…

… trace id is across microservices using http amqp headers…

Microservices seems to have something to do with: code, data, requests, backend, messages, trace id and http

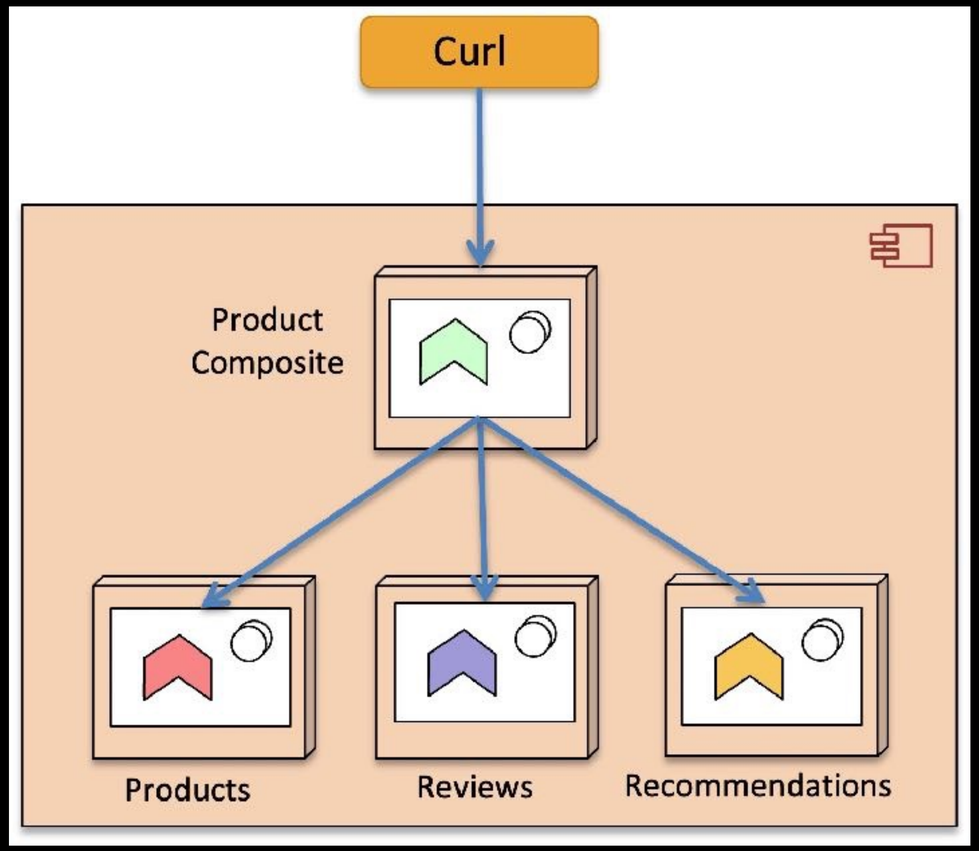


Illustration by Magnus Larsson, Callista Enterprise AB

KNOW A WORD BY THE COMPANY THAT IT KEEPS

… code necessary making our microservices publish monitoring data prometheus…

… requests are passed between microservices storage backends with messages…

… trace id is across microservices using http amqp headers…

Microservices seems to have something to do with: code, data, requests, backend, messages, trace id and http

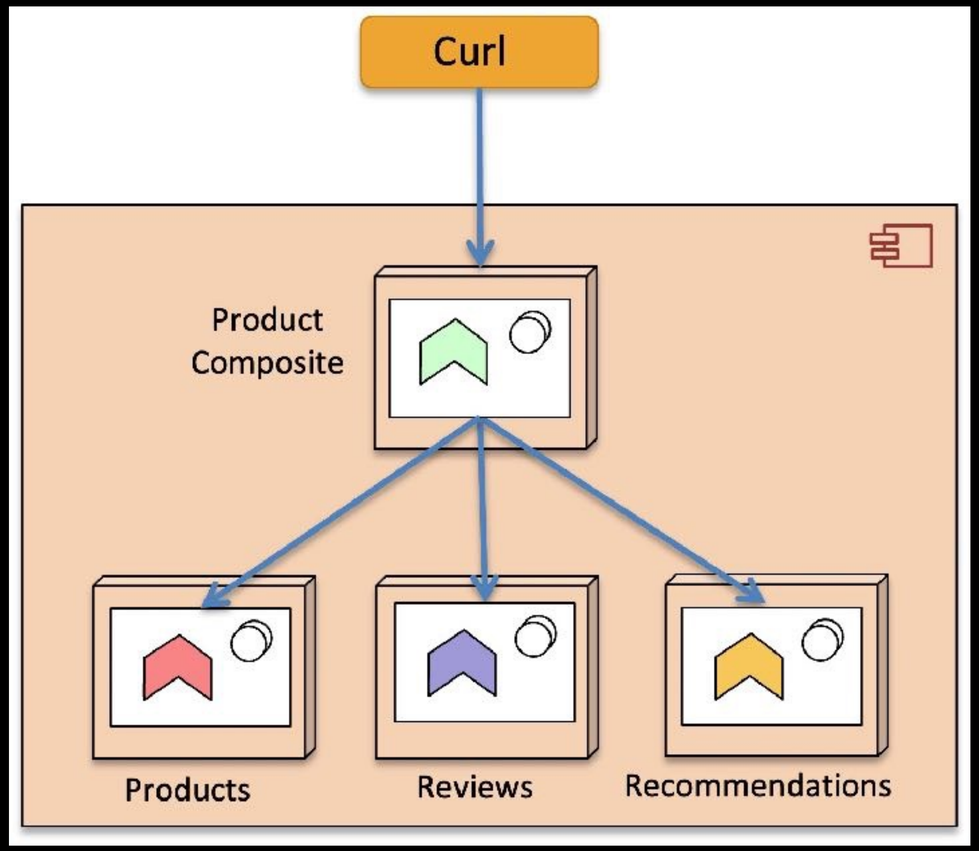


Illustration by Magnus Larsson, Callista Enterprise AB

KNOW A WORD BY THE COMPANY THAT IT KEEPS

… code necessary making our microservices publish monitoring data prometheus…

… requests are passed between microservices storage backends with messages…

… trace id is across microservices using http amqp headers…

Microservices seems to have something to do with: code, data, requests, backend, messages, trace id and http

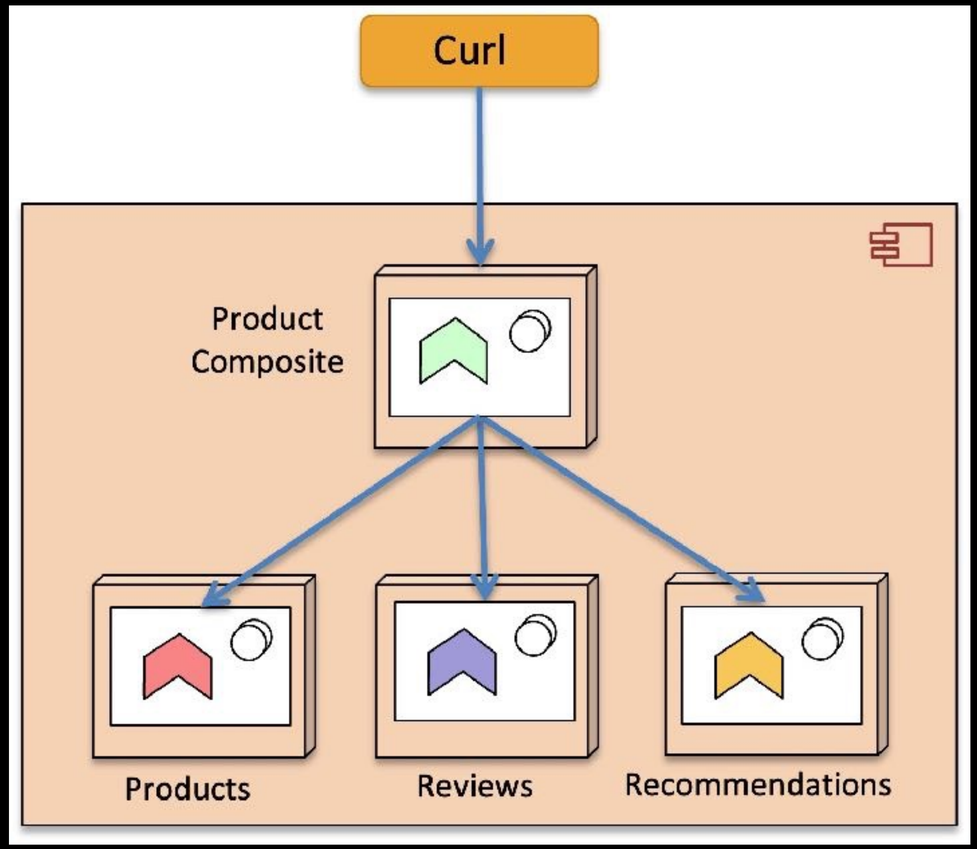


Illustration by Magnus Larsson, Callista Enterprise AB

KNOW A WORD BY THE COMPANY THAT IT KEEPS

… code necessary making our microservices publish monitoring data prometheus…

… requests are passed between microservices storage backends with messages…

… trace id is across microservices using http amqp headers…

Microservices seems to have something to do with: code, data, requests, backend, messages, trace id and http

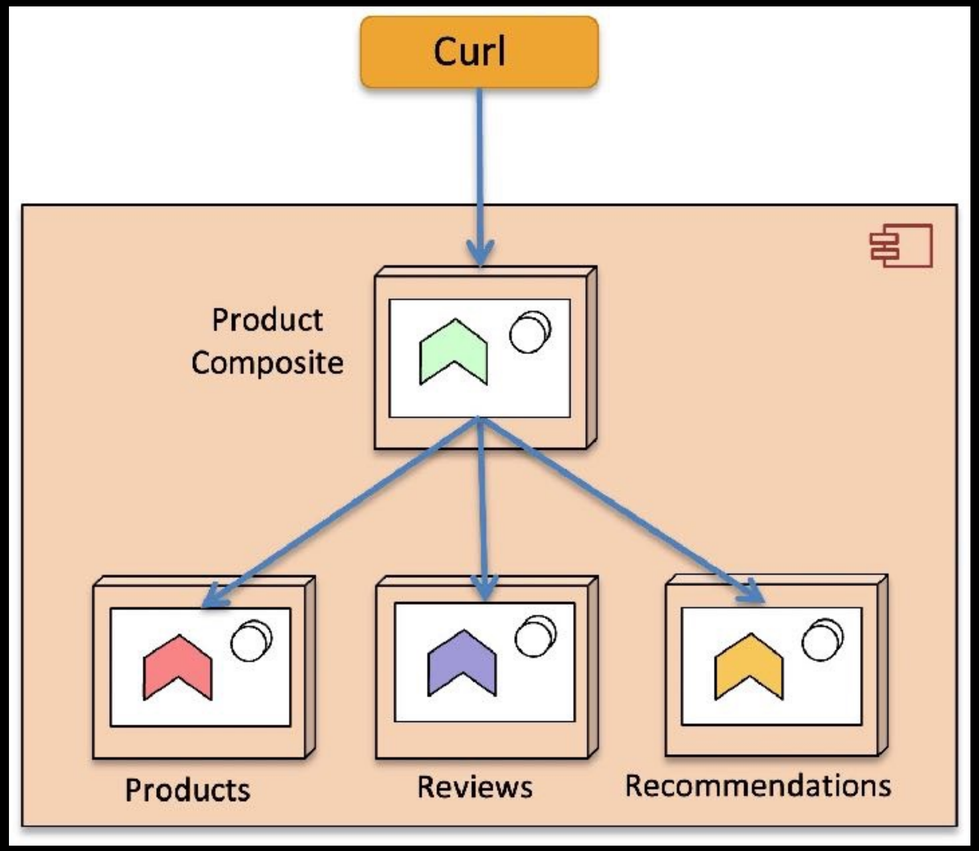


Illustration by Magnus Larsson, Callista Enterprise AB

KNOW A WORD BY THE COMPANY THAT IT KEEPS

… code necessary making our microservices publish monitoring data prometheus…

… requests are passed between microservices storage backends with messages…

… trace id is across microservices using http amqp headers…

Microservices seems to have something to do with: code, data, requests, backend, messages, trace id and http

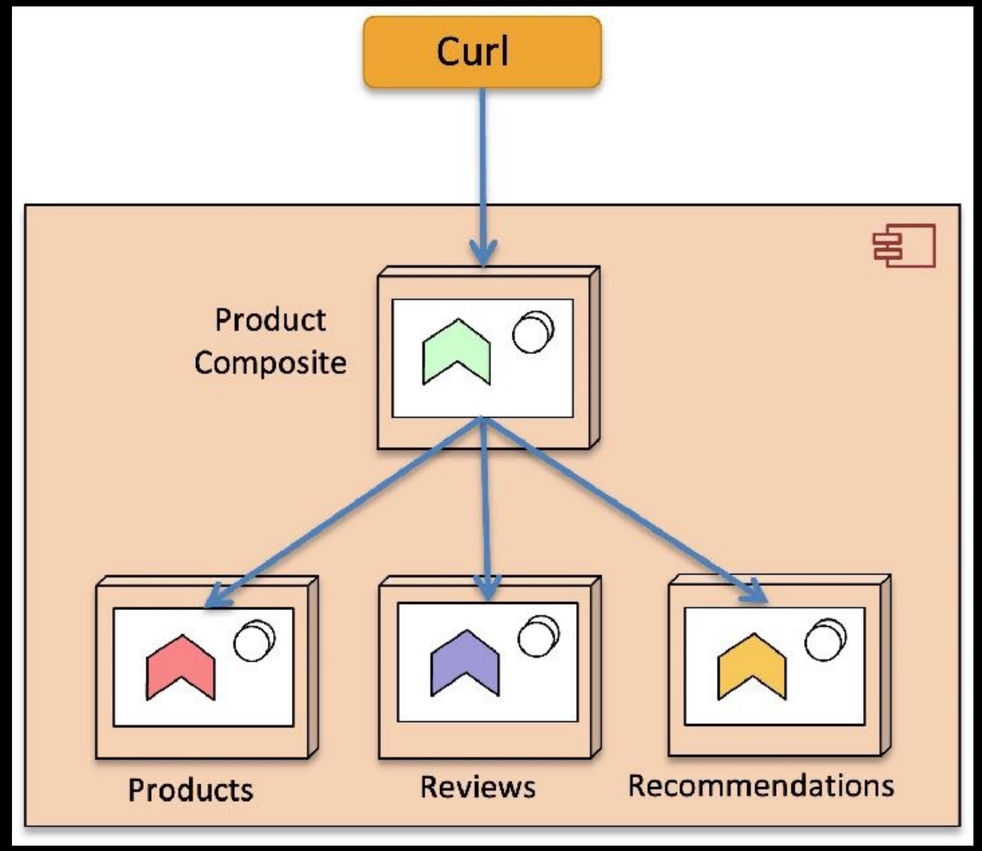
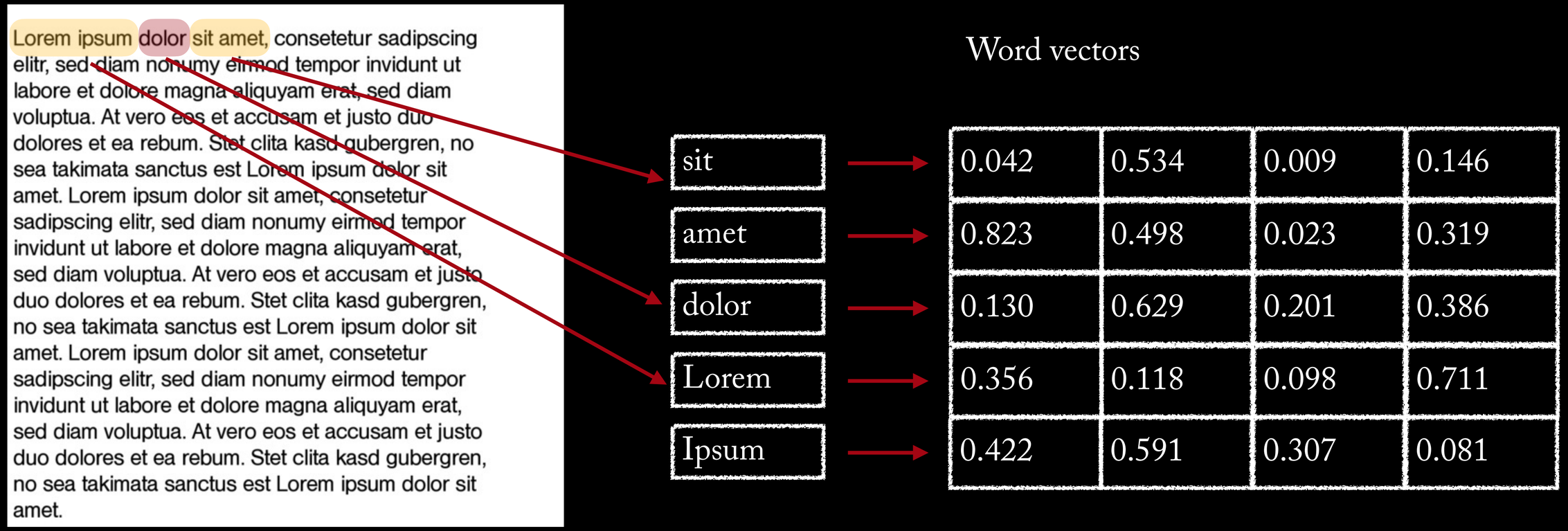


Illustration by Magnus Larsson, Callista Enterprise AB

BUILD WORD VECTORS

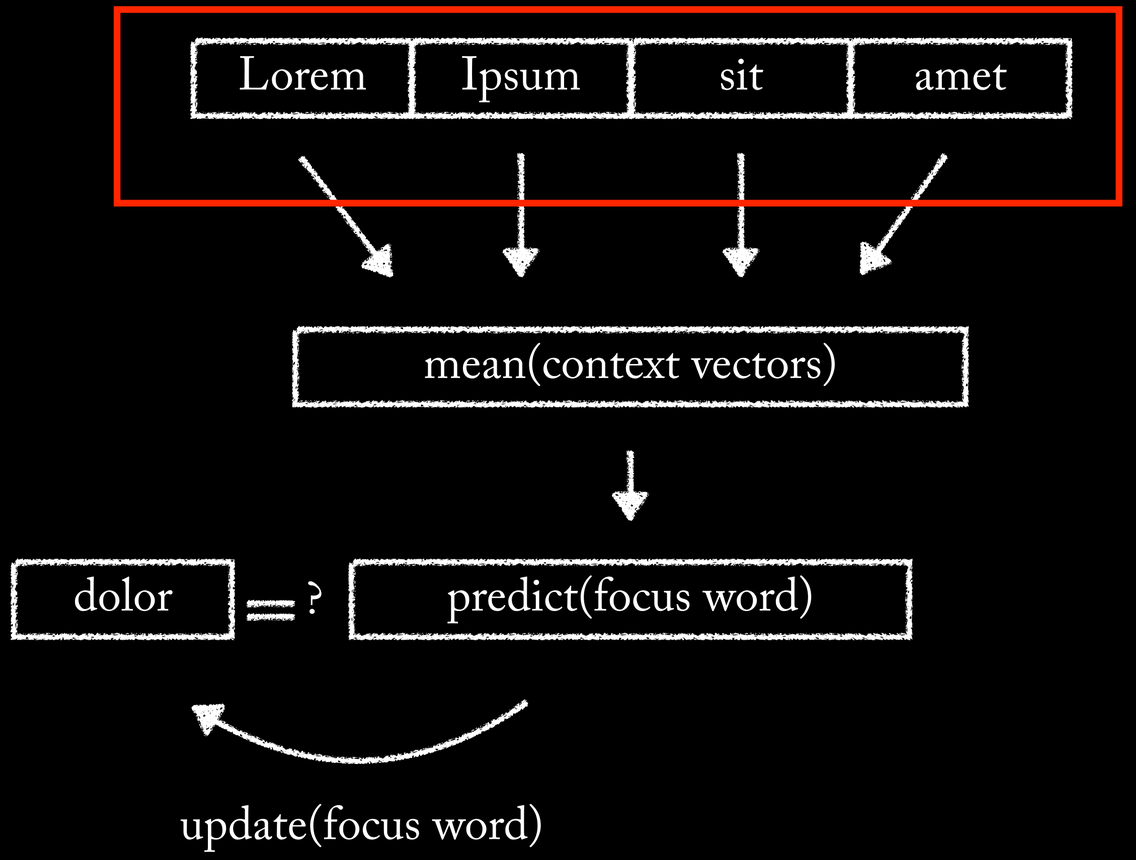
WORD2VEC - CONTINUOUS BAG-OF-WORDS (CBOW)



WORD2VEC - CBOW

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

(Word vectors!)



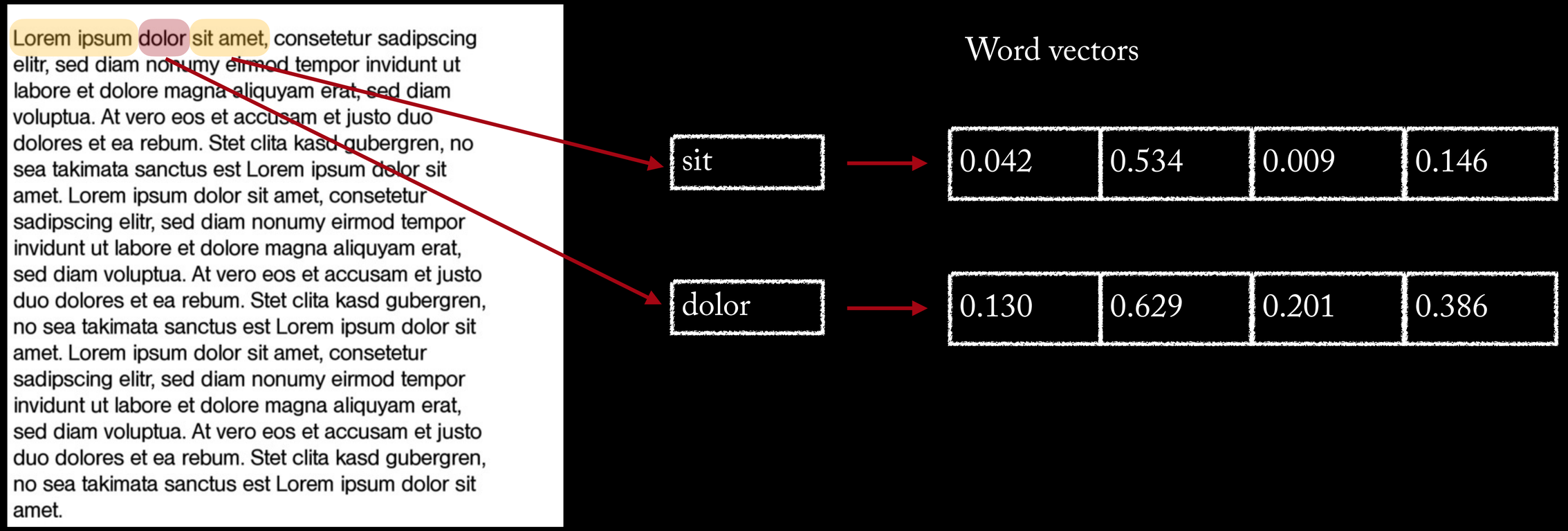
WORD2VEC - CBOW

Training:

- ~50k vocabulary
- 50 dimensional word vectors
- Window size: 5 (i.e. 10 context words)
- Avg. training time/epoch: 48-50 min

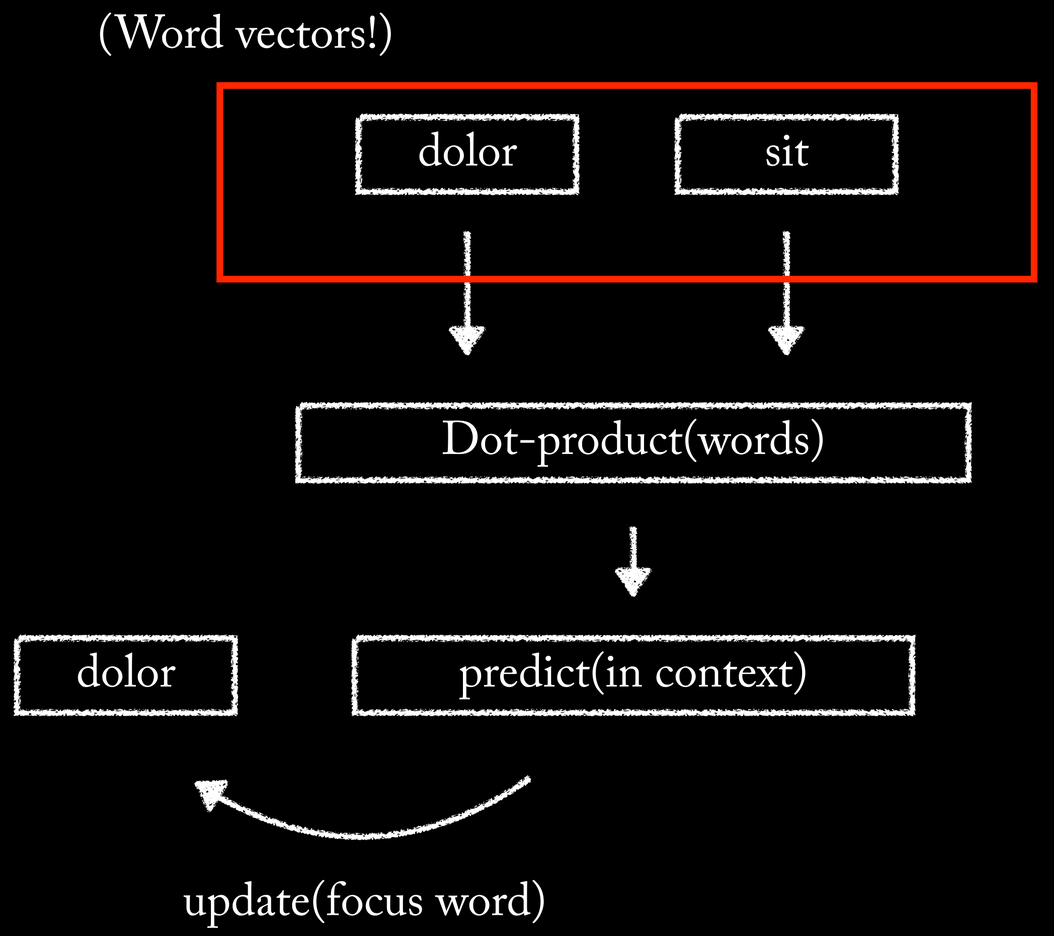
James van der Beek - Dawson's Creek

WORD2VEC - SKIP-GRAM



WORD2VEC - SKIP-GRAM

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



WORD2VEC - SKIP-GRAM

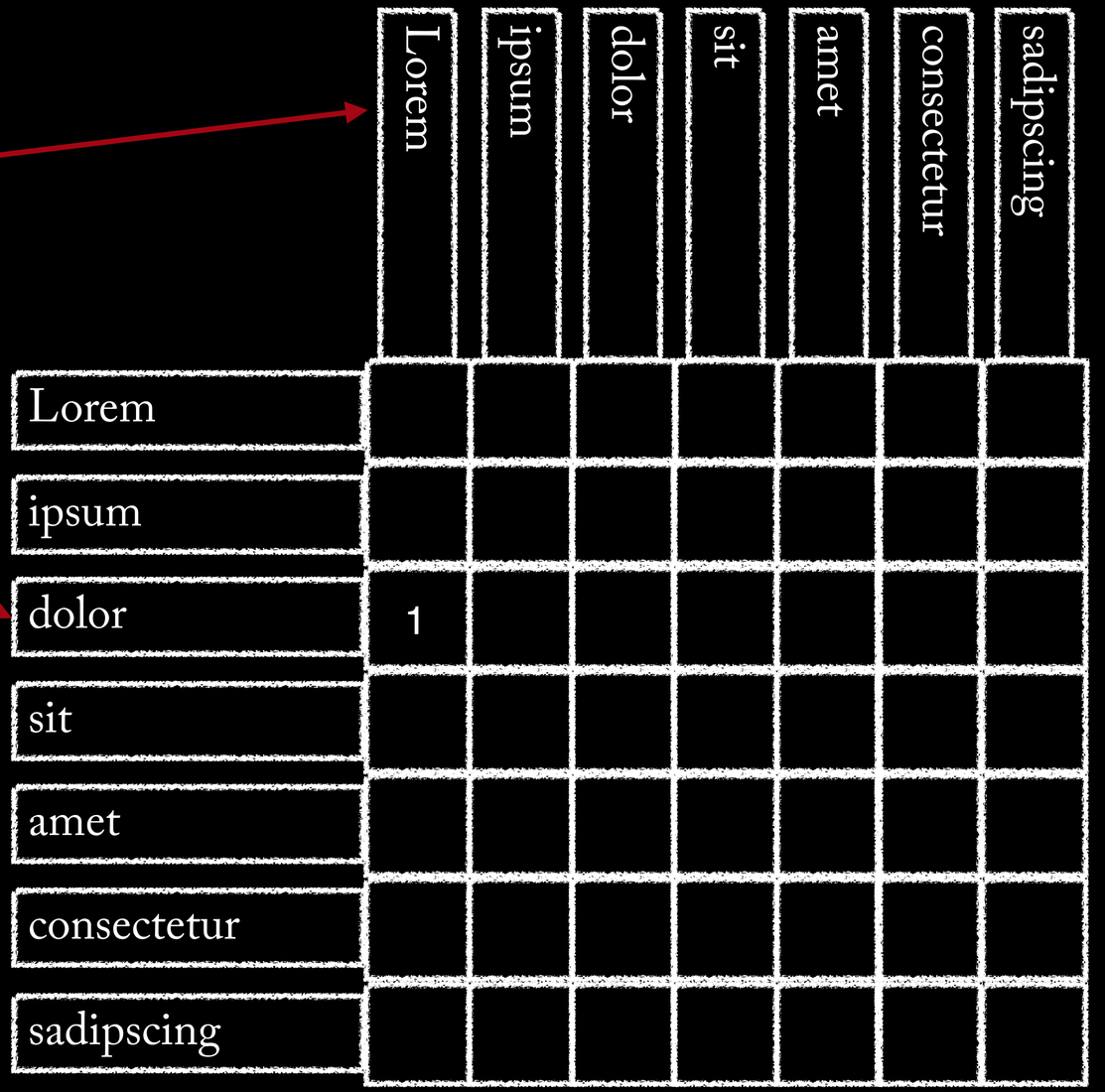
Training:

- ~50k vocabulary
- 50 dimensional word vectors
- Window size: 5 (i.e. 10 context words)
- Avg. training time/epoch: 27-28 min

Will Smith - TV Interview

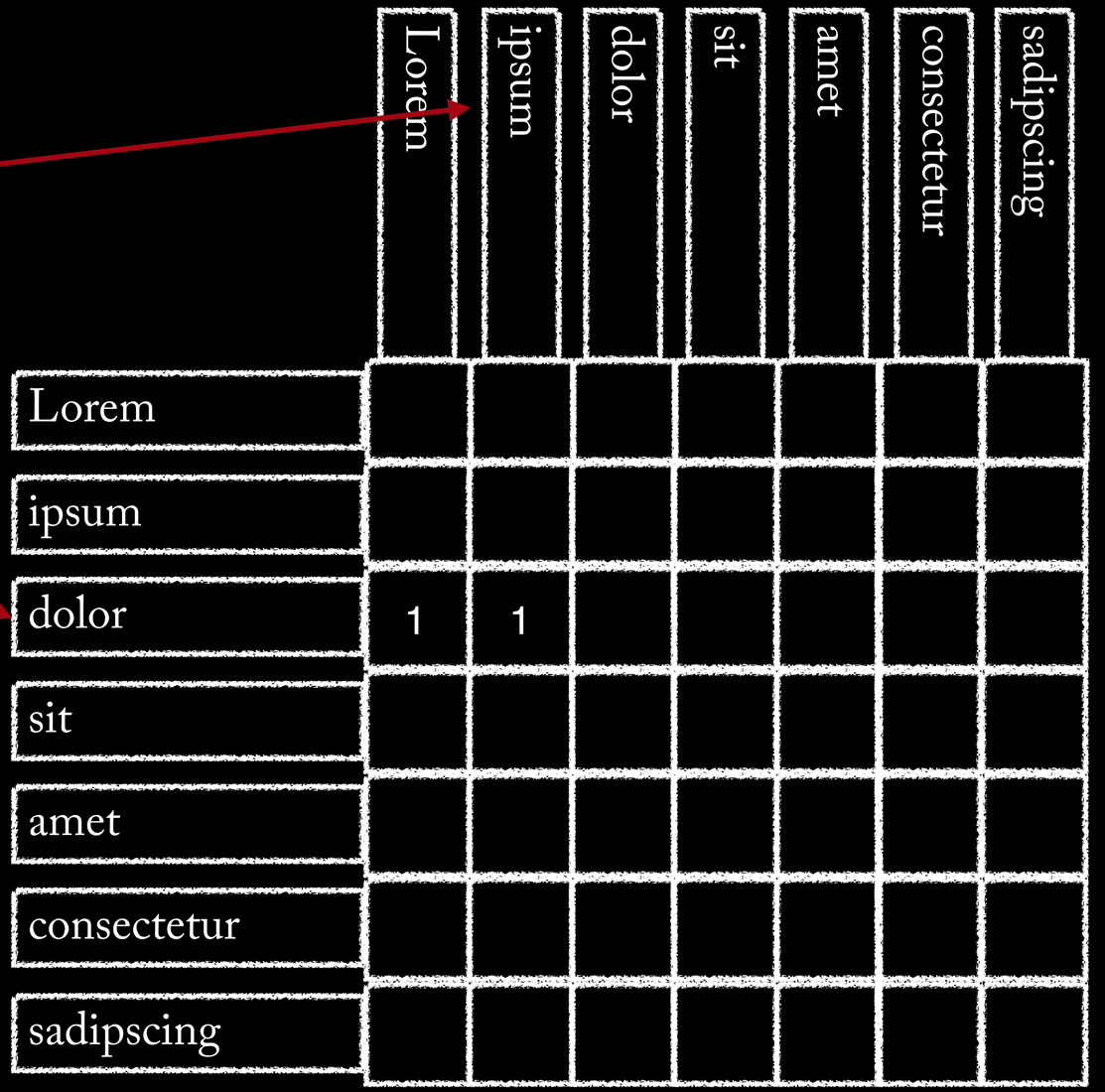
GLOVE

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



GLOVE

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

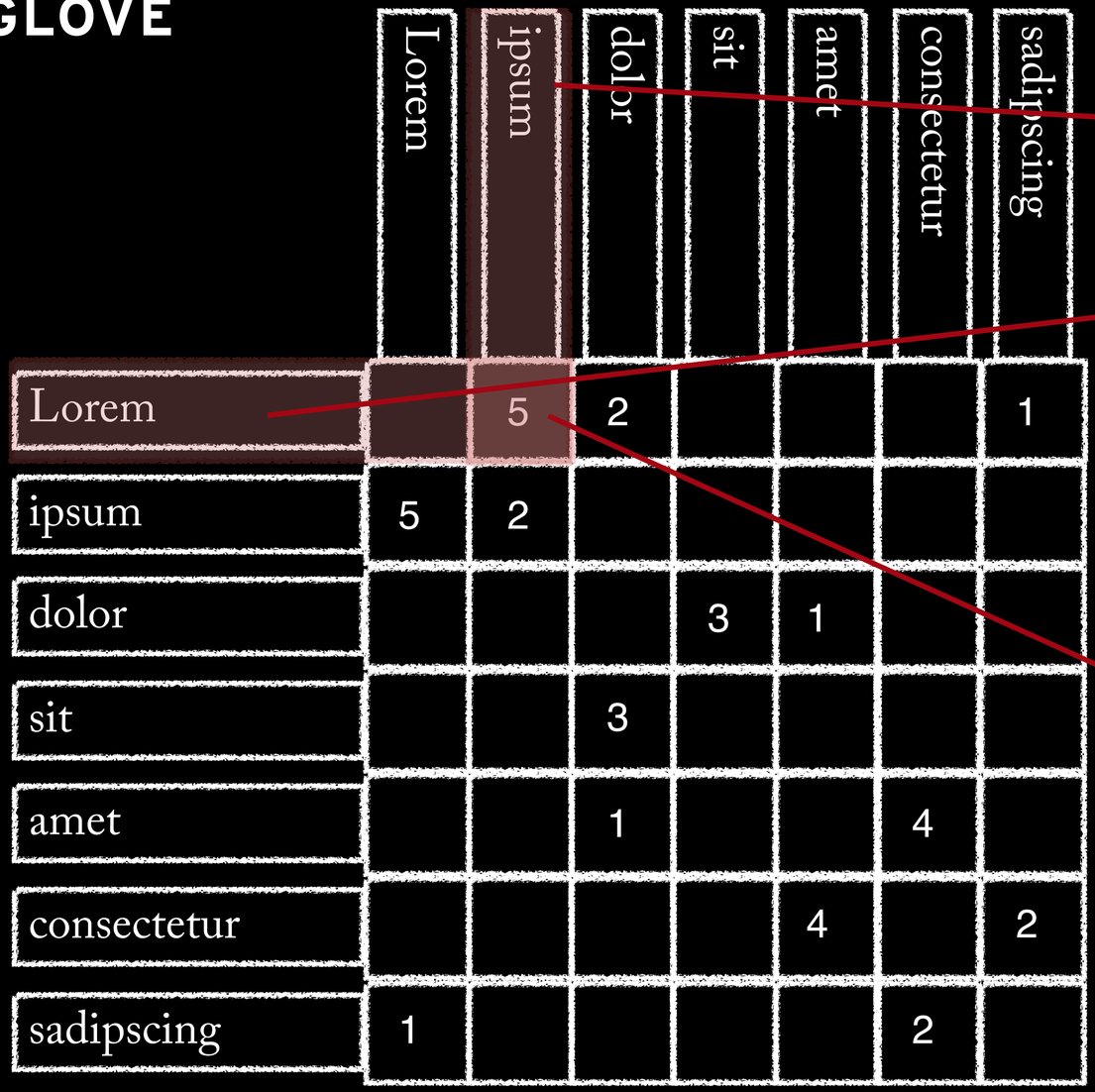


GLOVE

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

	Lorem	ipsum	dolor	sit	amet	consectetur	sadipscing
Lorem							
ipsum							
dolor	1	1		1	1		
sit		1	1		1	1	
amet			1	1		1	1
consectetur							
sadipscing							

GLOVE



(Word Vectors!)

Lorem ipsum

Dot-product(word-pair)

log(5) = ? predict(log(co-occurrence))

update(focus word)

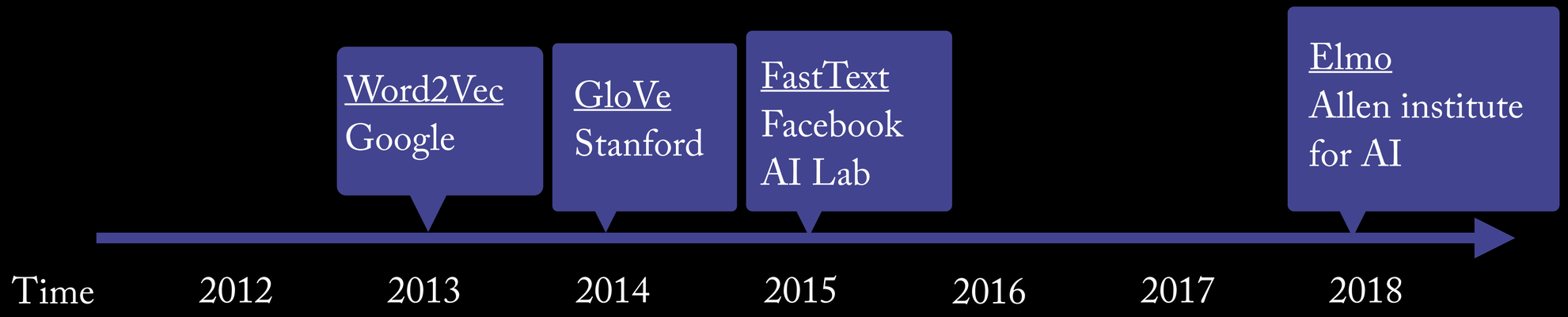
GLOVE

Training:

- ~50k vocabulary
- 50 dimensional word vectors
- Window size: 5 (i.e. 10 context words)
- Avg. training time/epoch: 8-9 min

CAN WE GET EVEN BETTER?

TIMELINE

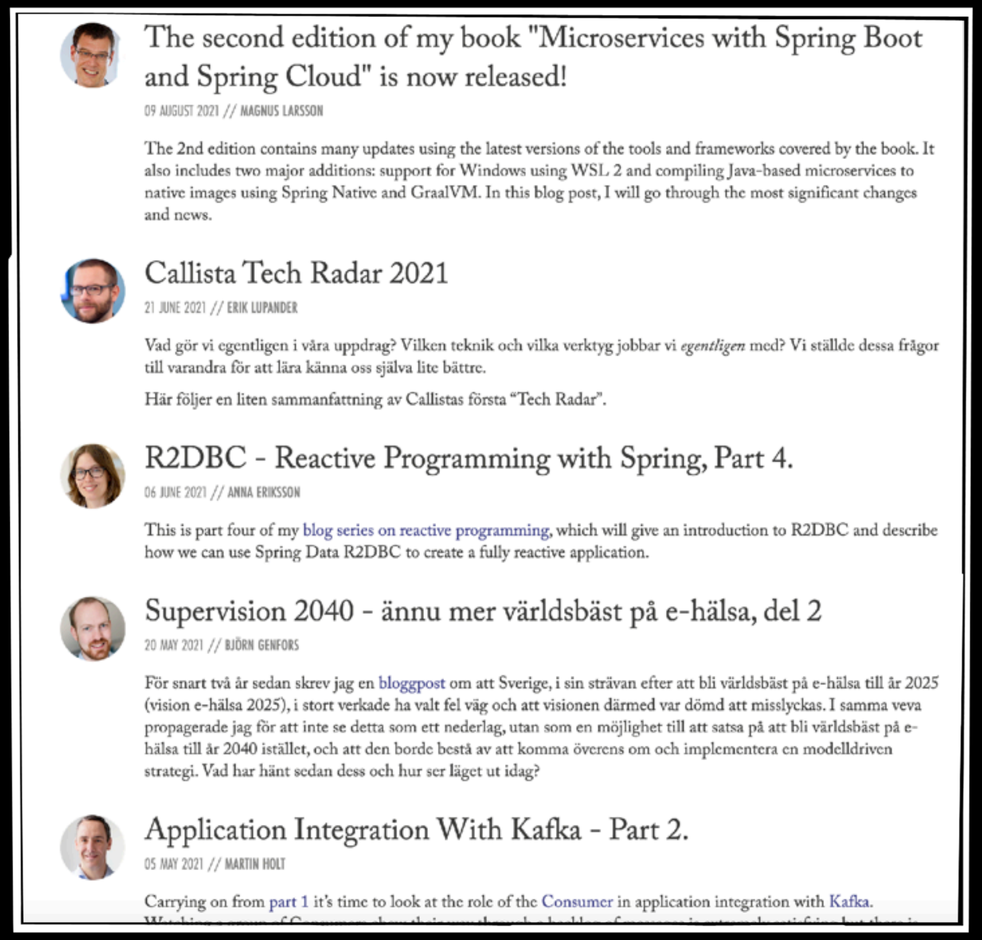

Word2Vec
Google
GloVe
Stanford
FastText
Facebook
AI Lab
Elmo
Allen institute
for AI
Time
2012
2013
2014
2015
2016
2017
2018

HOW CAN WE USE THIS?

CALLISTA CHATBOT PROJECT

CALLISTA CHATBOT PROJECT

https://callistaenterprise.se/blogg/



The second edition of my book "Microservices with Spring Boot and Spring Cloud" is now released!

09 AUGUST 2021 // MAGNUS LARSSON

The 2nd edition contains many updates using the latest versions of the tools and frameworks covered by the book. It also includes two major additions: support for Windows using WSL 2 and compiling Java-based microservices to native images using Spring Native and GraalVM. In this blog post, I will go through the most significant changes and news.

Callista Tech Radar 2021

21 JUNE 2021 // ERIK LUPANDER

Vad gör vi egentligen i våra uppdrag? Vilken teknik och vilka verktyg jobbar vi *egentligen* med? Vi ställde dessa frågor till varandra för att lära känna oss själva lite bättre.

Här följer en liten sammanfattning av Callistas första "Tech Radar".

R2DBC - Reactive Programming with Spring, Part 4.

06 JUNE 2021 // ANNA ERIKSSON

This is part four of my blog series on reactive programming, which will give an introduction to R2DBC and describe how we can use Spring Data R2DBC to create a fully reactive application.

Supervision 2040 - ännu mer världsbäst på e-hälsa, del 2

20 MAY 2021 // BJÖRN GENFORS

För snart två år sedan skrev jag en bloggpost om att Sverige, i sin strävan efter att bli världsbäst på e-hälsa till år 2025 (vision e-hälsa 2025), i stort verkade ha valt fel väg och att visionen därmed var dömd att misslyckas. I samma veva propagerade jag för att inte se detta som ett nederlag, utan som en möjlighet till att satsa på att bli världsbäst på e-hälsa till år 2040 istället, och att den borde bestå av att komma överens om och implementera en modelldriven strategi. Vad har hänt sedan dess och hur ser läget ut idag?

Application Integration With Kafka - Part 2.

05 MAY 2021 // MARTIN HOLT

Carrying on from part 1 it's time to look at the role of the Consumer in application integration with Kafka.

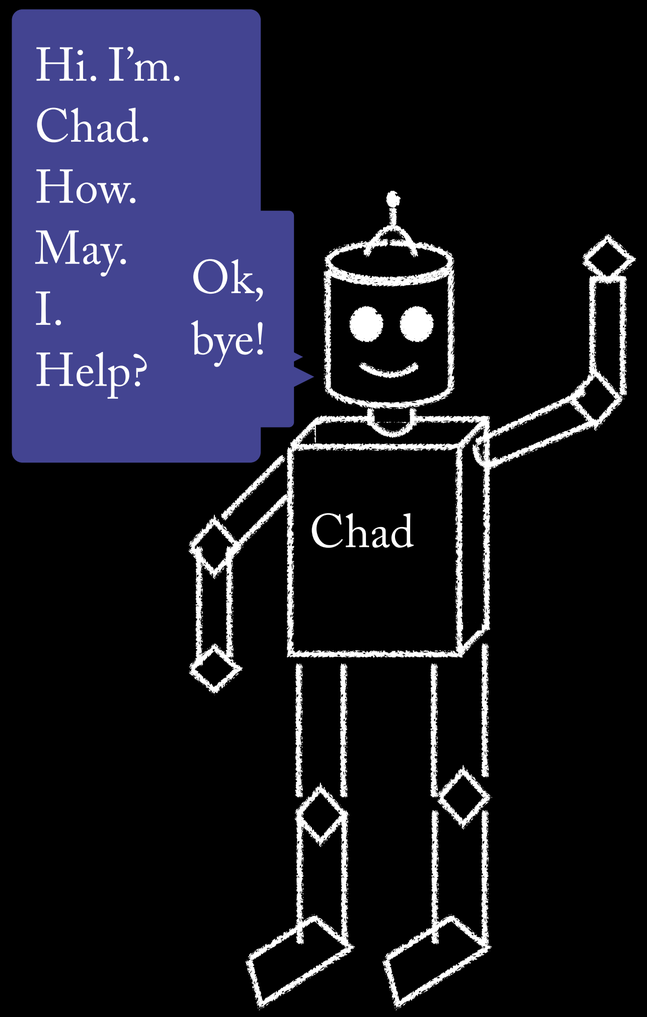
CALLISTA CHATBOT TEAM

Peter Hernfalk

Sara Adenbrant

CALLISTA CHATBOT PROJECT

- Interpret user input
- Identify relevant blog posts

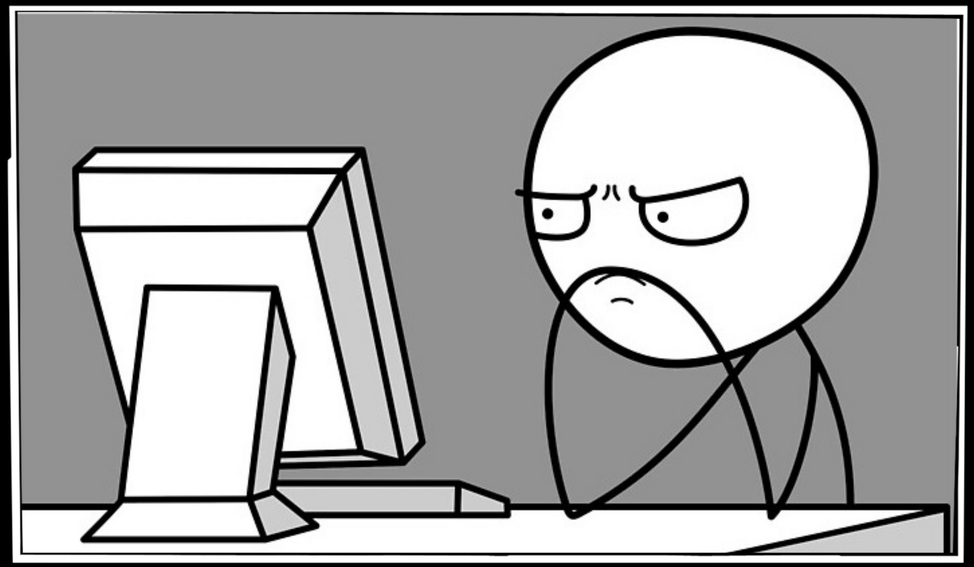


INTERPRET USER INPUT

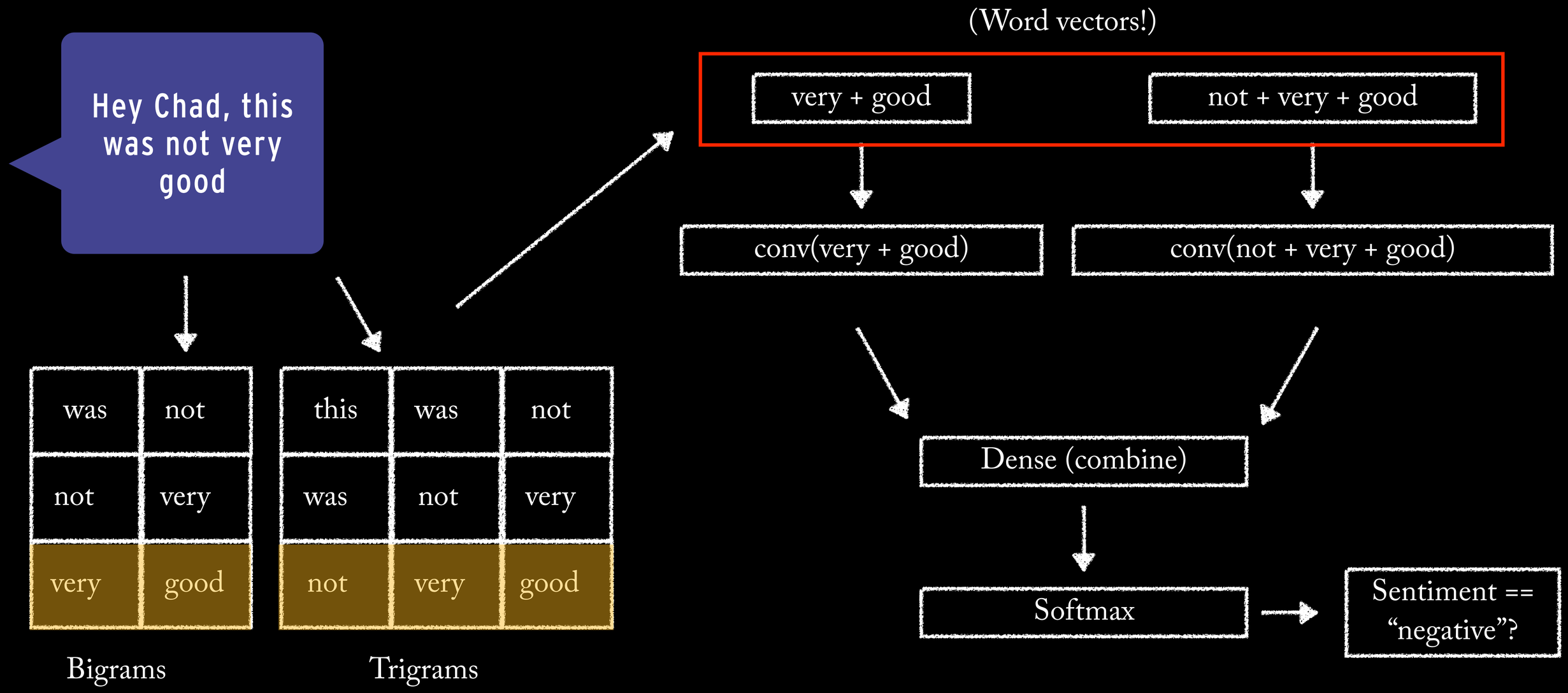
SENTIMENT ANALYSIS

Problem:

We understand words, but how do we distinguish between:
“This was very good” vs.
“This was not very good”



CHATBOT: SENTIMENT ANALYSIS

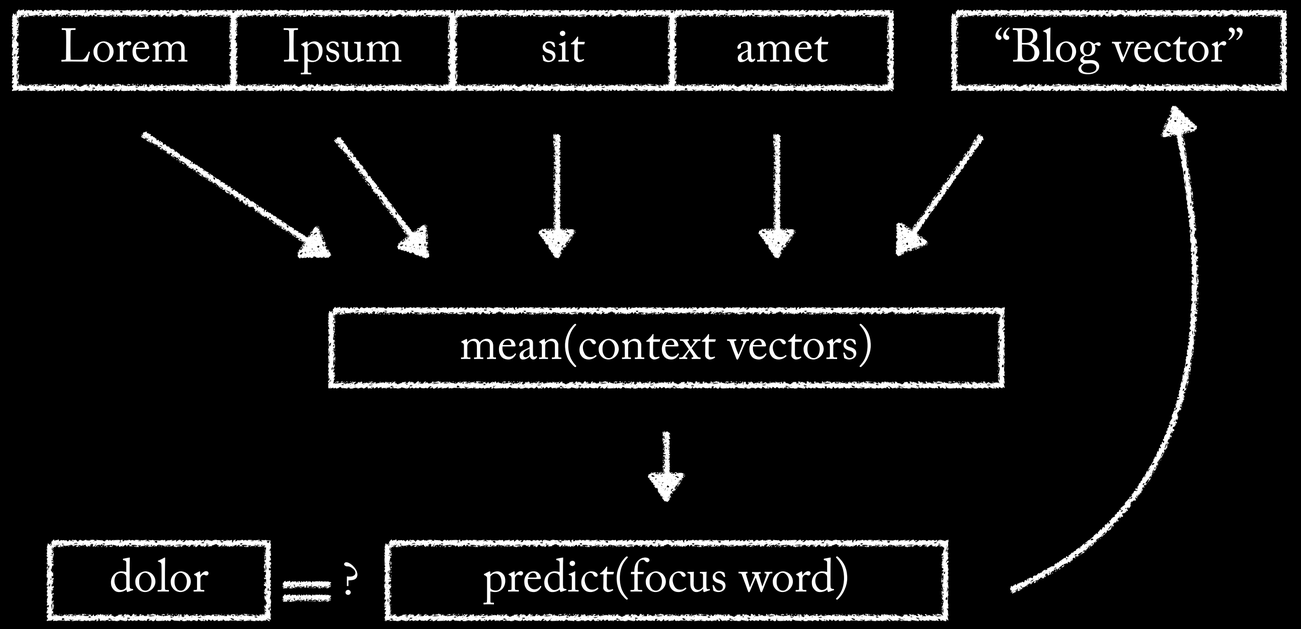


IDENTIFY RELEVANT BLOG POSTS

CHATBOT: IDENTIFY INTERESTING BLOGS

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

It's a blog!



USEFUL RESOURCES

- https://towardsdatascience.com/
- https://keras.io/
- https://nlp.stanford.edu/projects/glove/
- https://fasttext.cc/
- https://allenai.org/allennlp/software/elmo
- https://youtu.be/OQQ-W_63UgQ

THANK YOU!

THE END

CALLISTA