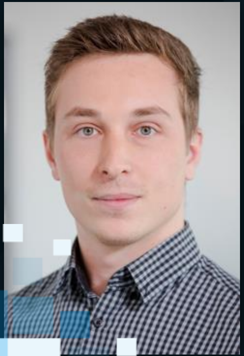
We develop Individual Software &
Artificial Intelligence Solutions



Simon Stiebellehner

Data Scientist @ craftworks
Lecturer @ WU Wien & FH Wien

simon.stiebellehner@craftworks.at

craftworks

Our **algorithms** and **solutions** have been **awarded multiple times**

craftworks.at

Literally Recommendable

Using Text Embedding Algorithms in Recommendation Systems

1. **Natural Language Processing (NLP)**
2. Recommendation Systems (RecSys)
3. Hybrid Filtering RecSys powered by word2vec

craftworks

1 Natural Language Processing (NLP)

craftworks

Big Leaps Forward

How come **machines** have become decent at **understanding humans** recently?

amazon alexa

Hey Siri

Hi. I'm Cortana.

"Ok Google"

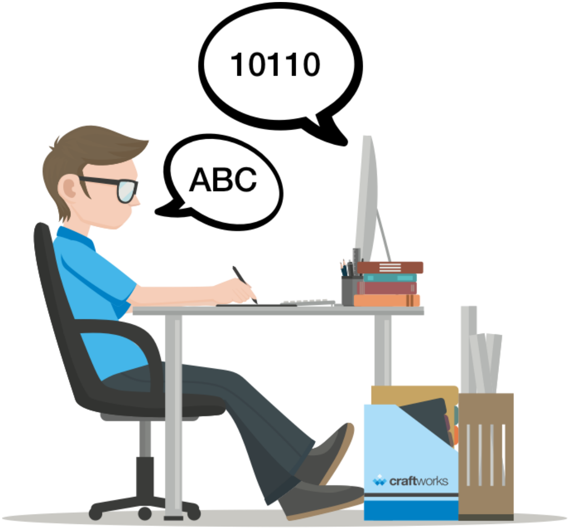
It's Complicated

"Communication is key to a healthy relationship"

… well, it's complicated.

Humans
communicate using
natural languages
(German, English, …)

Nondeterministic



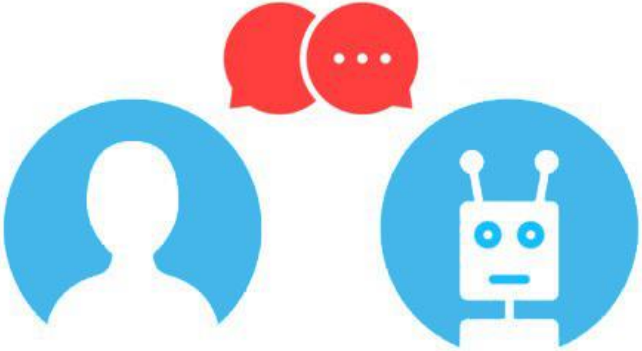
Machines
communicate using
artificial languages
(binary, C++, …)

Deterministic

The Translator

Natural Language Processing is the **translator**.

NLP is a collection of **methods** from **linguistics, statistics** and **computer science** that aims to make *computers “understand” natural language.*



craftworks

The Translation Process

Computers excel at **Maths**. So let's turn **text into numbers**:

1. Intelligent humans control machines.
2. Intelligent machines control humans.
3. Be careful with machines.

"Bag of Words"

	be	careful	control	human	intelligent	machine	with
Sentence 1	0	0	1	1	1	1	0
Sentence 2	0	0	1	1	1	1	0
Sentence 3	1	1	0	0	0	1	1

Natural Language is Complex

The **bag of words** method is **ignorant**. For instance, it fails to model:

- word order
- context
- grammar

We lost any information about what truly matters for humans:

The relationship between words/phrases and symbols, constituting *meaning*. **We lack Semantics.**

Company is Key

We need to **preserve semantics** when converting text to a numerical, **machine-readable** form.

HOW?

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

J. R. Firth, linguist, 1957

An Old Hypothesis in New Clothes

Distributional Hypothesis:

*Words that appear in **similar contexts** are **similar in meaning**.*

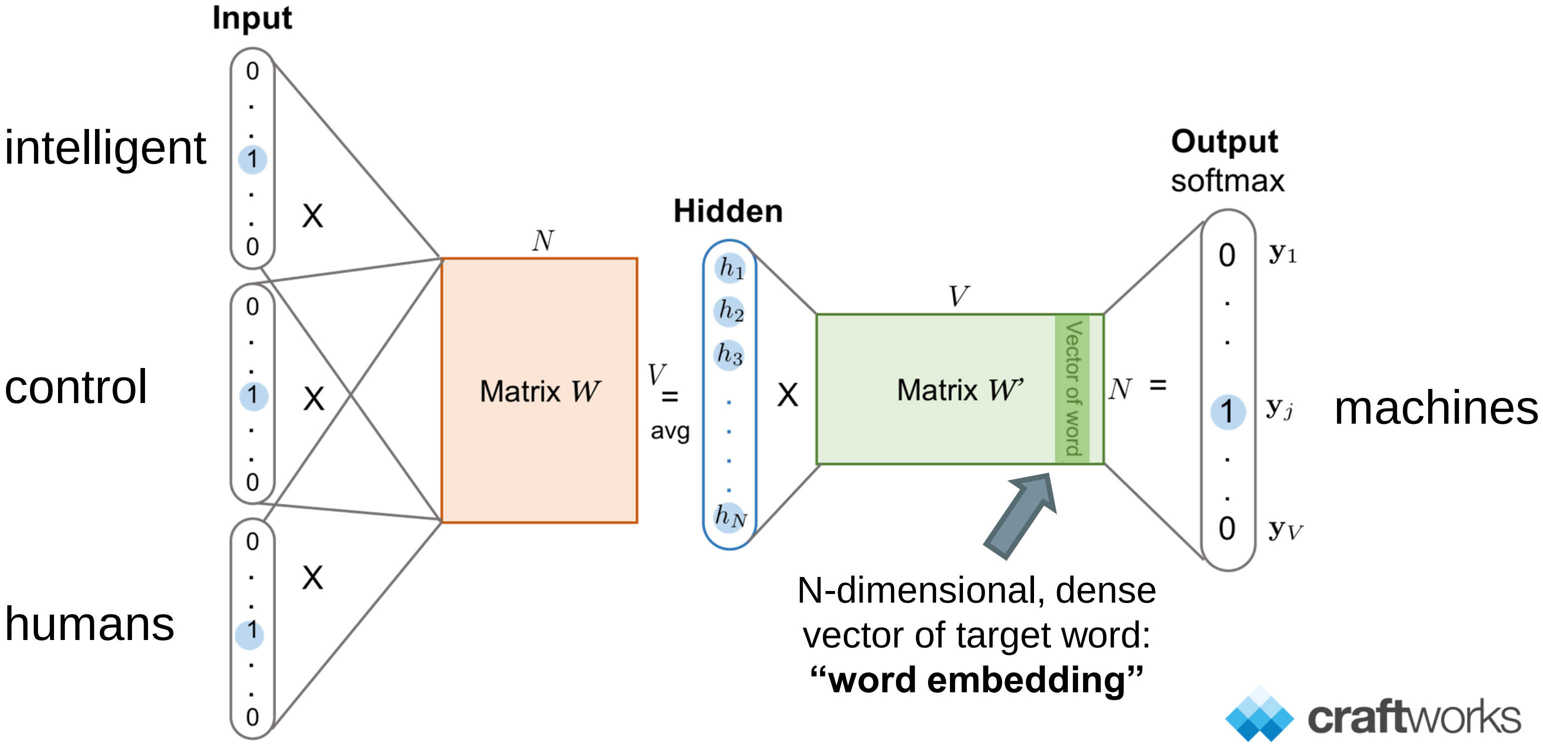
e.g. father ↔ dad

Modern **Text Embedding Algorithms** make use of this idea.

Most notably: **word2vec** by Mikolov et al.

word2vec – CBOW Model in Action

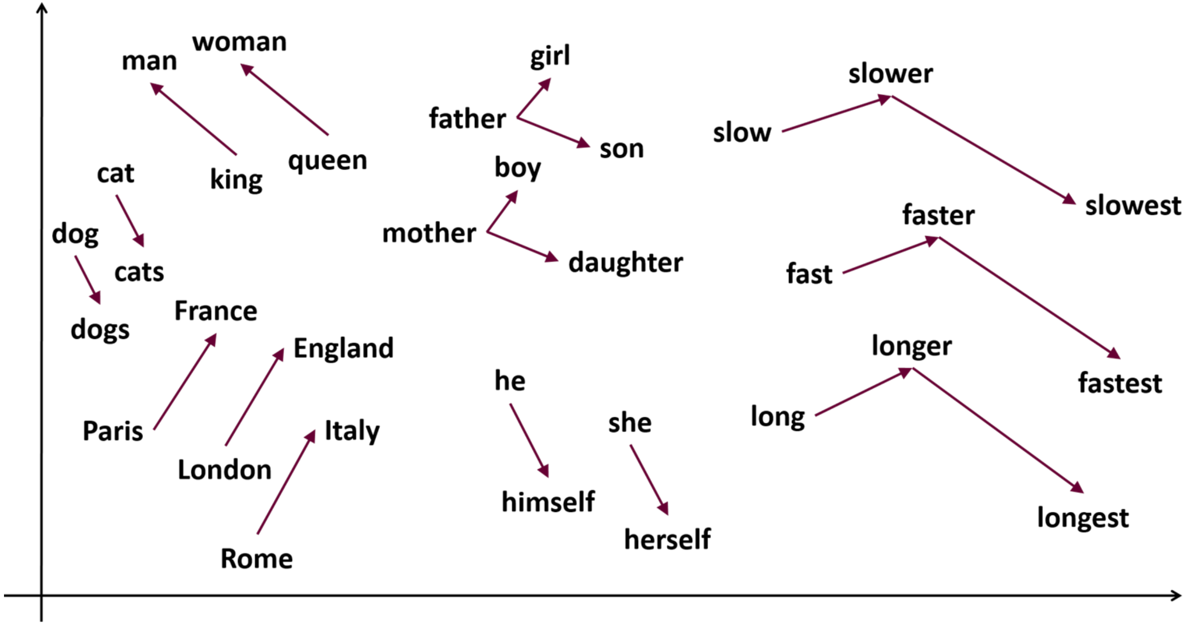
“intelligent machines control humans”



NLP

Semantics in a Nutshell: Dense Vectors

- w2v generates **dense N-dimensional numeric word vectors**
- **similar words ↔ similar vectors** (cosine similarity)



Projection of vectors into 2-dimensional space using t-SNE

craftworks

Literally Recommendable

Using Text Embedding Algorithms in Recommendation Systems

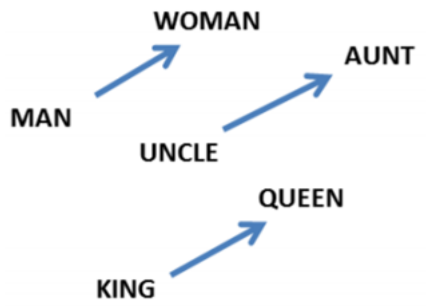
✓ 1. Natural Language Processing (NLP)
2. **Recommendation Systems (RecSys)**
3. Hybrid Filtering RecSys powered by word2vec

craftworks

Similarity

Remember:
word2vec is excellent at extracting the **meaning of words.**

similar context → similar meaning → similar vectors



Similarity is not only important when it comes to words ...

RecSys

You are the Average of the Five People you Spend the Most Time with

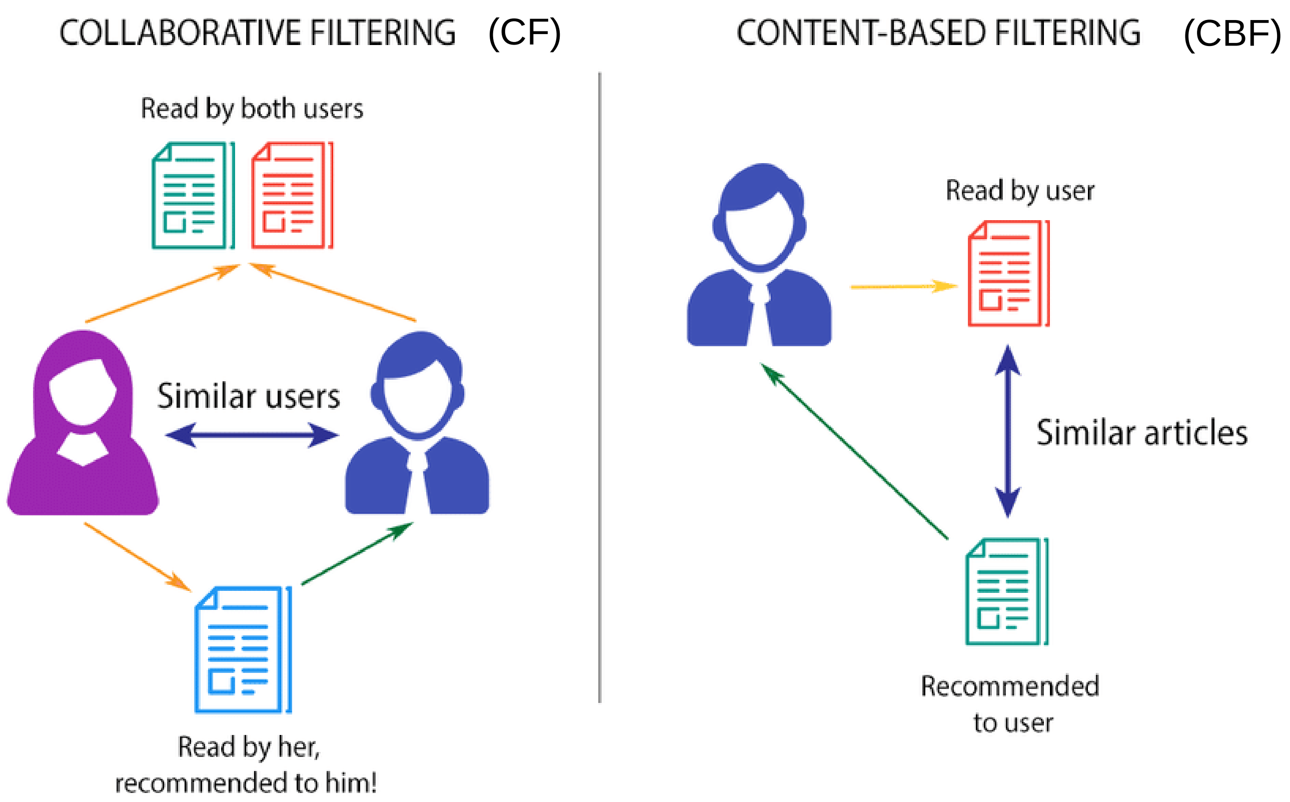
- **Similarity** also plays a key role in our social relationships

- Typically, our **friends are highly similar** to ourselves and we often **like the same things**

Also **Recommendation Systems** often use **similarity measures** to find out what we like.

craftworks

Traditional Filtering Methods in Recommendation Systems



Traditional Filtering Methods have significant Weaknesses

Problems of Collaborative Filtering

- Cold Start problem
- Synonyms
- Gray Sheep

Problems of Content-based Filtering

- Entirely relies on quality of metadata
- “boring” recommendations

Hybrid Filtering

Both Collaborative and Content-based Filtering suffer from **significant weaknesses**.
→ Overcome by **combining both techniques**.

$$f\left(\text{Collaborative Filtering}, \text{Content-based Filtering}\right)$$

$$= \textbf{Hybrid Filtering}$$

Literally Recommendable

Using Text Embedding Algorithms in Recommendation Systems

1. Natural Language Processing (NLP)
2. Recommendation Systems (RecSys)
3. **Hybrid Filtering RecSys powered by word2vec**

craftworks

An Enlightening Analogy

“What do **news articles** and **website visitors** have in common?”

Both are **sequences**!

③ Hybrid

A Matter of Perspective

For a website visitor, online news articles are **sequences of words**.

↔

For a news website, visitors are **sequences of articles** they read.

news article = [“intelligent”, “machines”, “control”, “humans”]

↔

news website visitor = [“ArticleID_1”, “ArticleID_2”, “ArticleID_3”]

craftworks

Putting it together: word2vec in Hybrid Filtering

For a news website …

- users are sequences of articles

For a news website visitor (user) ...

- articles are sequences of words

users are sequences of sequences of words

User A = [

[“new”, “tensorflow”, “version”, “released”, ...],
[“learn”, “python”, “for”, “machine”, “learning”, ...],
[“next”, “bitcoin”, “hype”, “is”, “coming”, ...]

]

craftworks

Putting it together: word2vec in Hybrid Filtering

Training word2vec on our corpus of articles (sequences of words) provides us with N-dimensional **vectors for each word i**.

Aggregation on article level:

$$article\_vector_j = \frac{1}{N} \sum_{i \in K=1}^{N} word\_vector_i$$

Aggregation on user level:

$$user\_vector_c = \frac{1}{N} \sum_{j \in P=1}^{N} article\_vector_j$$

Practical Example

You can find the Jupyter Notebook on the craftworks github account:

https://github.com/craftworksgmbh/wad

3 Hybrid

Putting it together: word2vec in Hybrid Filtering

What do we get from this?

- Mapping of items and users into **shared vector space** brings flexibility

- Computation of similarities and making **recommendations**:
  - User-to-User
  - Item-to-Item
  - Item-to-User

- **Overcoming problems** of traditional methods

craftworks

Text Embedding Algorithms in Hybrid Filtering are powerful

Problems of Collaborative Filtering

- Cold Start problem
- Synonyms
- Gray Sheep

Solved!

Problems of Content-based Filtering

- Entirely relies on quality of metadata
- “boring” recommendations

Literally Recommendable

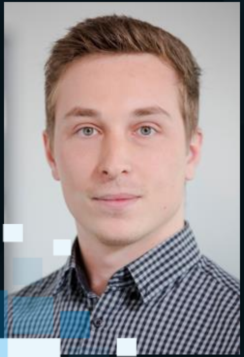
Using Text Embedding Algorithms in Recommendation Systems

1. Natural Language Processing (NLP)
2. Recommendation Systems (RecSys)
3. Hybrid Filtering RecSys powered by word2vec

craftworks

craftworks

We develop Individual Software &
Artificial Intelligence Solutions



Simon Stiebellehner

Data Scientist @ craftworks
Lecturer @ WU Wien & FH Wien

simon.stiebellehner@craftworks.at