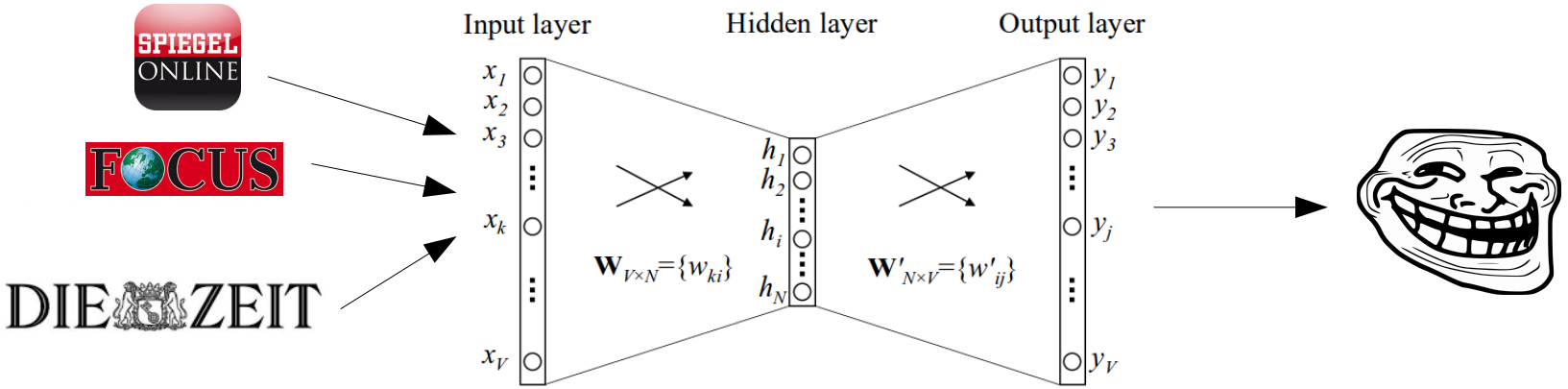
Analysing user comments with Doc2Vec and Machine Learning



Robert Meyer Data Scientist at Flixbus

PyData Berlin 2017 robert.meyer@flixbus.com

What?

- What can we learn from **user comments** on news sites?
  - Scrape Online Comments
  - Doc2Vec (& Word2Vec)
  - Supervised (linear) Machine Learning

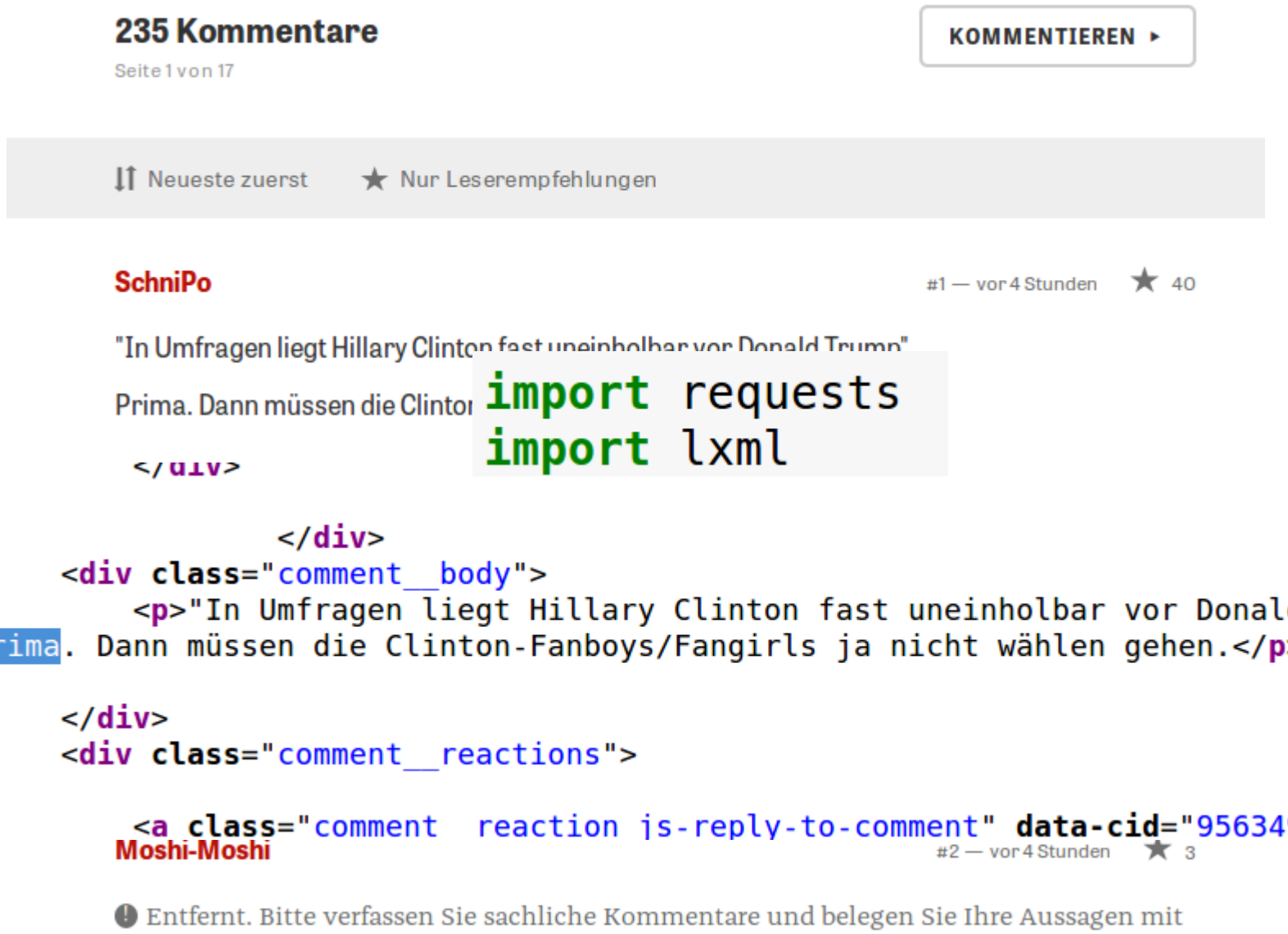
News Sites

Hypothesis

(slightly) smarter

OMG srsly?

Scraping


235 Kommentare
Seite 1 von 17
KOMMENTIEREN
Neueste zuerst
Nur Leserempfehlungen
SchniPo
#1 — vor 4 Stunden
40
"In Umfragen liegt Hillary Clinton fast uneinholbar vor Donald Trump".
Prima. Dann müssen die Clinton
import requests
import lxml

"In Umfragen liegt Hillary Clinton fast uneinholbar vor Donald Trump".
Prima. Dann müssen die Clinton-Fanboys/Fangirls ja nicht wählen gehen.

<a class="comment reaction js-reply-to-comment" data-cid="9563492" href="http:/
Moshi-Moshi
#2 — vor 4 Stunden
3
Entfernt. Bitte verfassen Sie sachliche Kommentare und belegen Sie Ihre Aussagen mit

Scraping

- Comments from January 2014 till June 2016
  - SPON ~280,000
  - ZEIT ~170,000
  - Focus ~50,000
    - tokenized with `import nltk`

Preprocessing

“... an unseren Schulen Einigkeit und Recht und Freiheit für das deutsche Vaterland lehren soll.” /
'*... teach unity and justice and freedom for our German Fatherland at our schools.'* (**Focus**)

“Wenn ich nun aber überzeugter Vegetarier bin, dennoch aber ab und an einen Hamburger essen möchte ...” /
'*If I was a staunch vegetarian, but I would like to eat a hamburger ...'*
(**ZEIT**)

“Trump gewinnt US Wahl, die EU zerbricht nach dem Brexit ...” /
'*Trump will win the US election, the EU will break down after the Brexit ...'*
(**SPON**)

...

Preprocessing

“... an unseren Schulen Einigkeit und Recht und Freiheit für das deutsche Vaterland lehren soll.” /
'*... teach unity and justice and freedom for our German Fatherland at our schools.'* (**Focus**)

[..., an, unseren, schulen, einigkeit, und, recht, und, freiheit, für, das, deutsche, vaterland, lehren, soll]
(**doc 1**)

“Wenn ich nun aber überzeugter Vegetarier bin, dennoch aber ab und an einen Hamburger essen möchte ...” /
'*If I was a staunch vegetarian, but I would like to eat a hamburger ...'*
(**ZEIT**)

[wenn, ich, nun, aber, überzeugter, vegetarier, bin, dennoch, aber, ab, und, an, einen, hamburger, essen, möchte, ...]
(**doc 2**)

“Trump gewinnt US Wahl, die EU zerbricht nach dem Brexit ...” /
'*Trump will win the US election, the EU will break down after the Brexit ...'*
(**SPON**)

[trump, gewinnt, us, wahl, die, eu, zerbricht, nach, dem, brexit ...]
(**doc 3**)

...

...

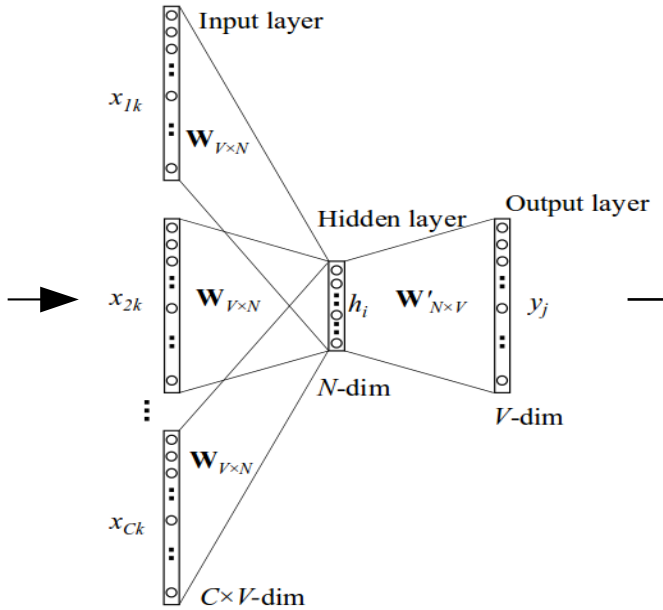
Doc2Vec / Word2Vec

[..., an, unseren, schulen, einigkeit, und, recht, und, freiheit, für, das, deutsche, vaterland, lehren, soll] (**doc 1**)

[wenn, ich, nun, aber, überzeugter, vegetarier, bin, dennoch, aber, ab, und, an, einen, hamburger, essen, möchte, ...] (**doc 2**)

[trump, gewinnt, us, wahl, die, eu, zerbricht, nach, dem, brexit ...] (**doc 3**)

...



Amazing Stuff!

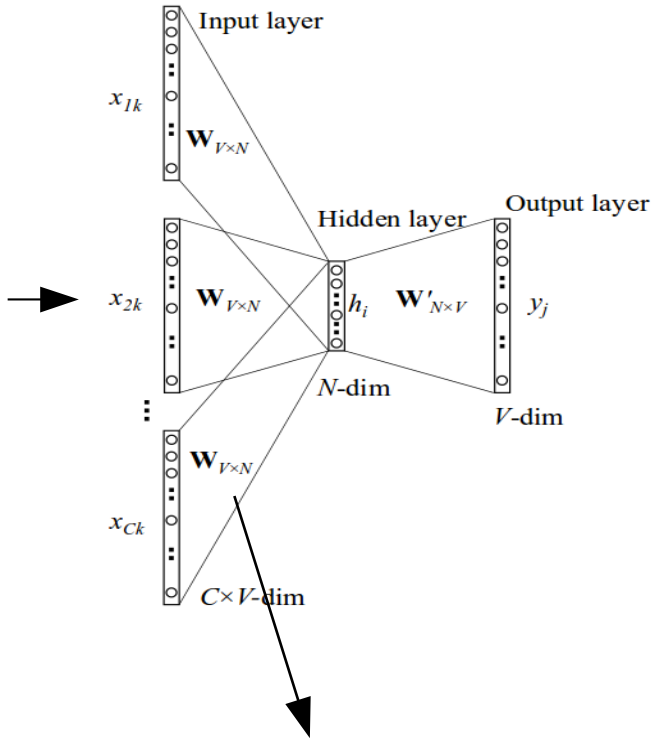
Doc2Vec / Word2Vec

[..., an, unseren, schulen, einigkeit, und, recht, und, freiheit, für, das, deutsche, vaterland, lehren, soll] (**doc 1**)

[wenn, ich, nun, aber, überzeugter, vegetarier, bin, dennoch, aber, ab, und, an, einen, hamburger, essen, möchte, ...] (**doc 2**)

[trump, gewinnt, us, wahl, die, eu, zerbricht, nach, dem, brexit ...] (**doc 3**)

...



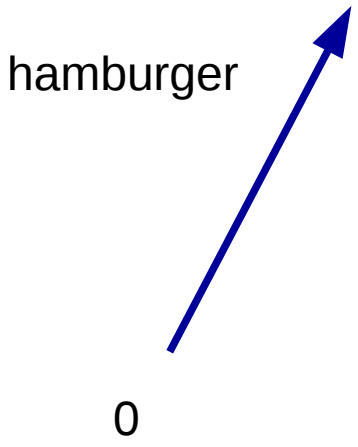
Amazing Stuff!

Word2Vec

hamburger

Word2Vec

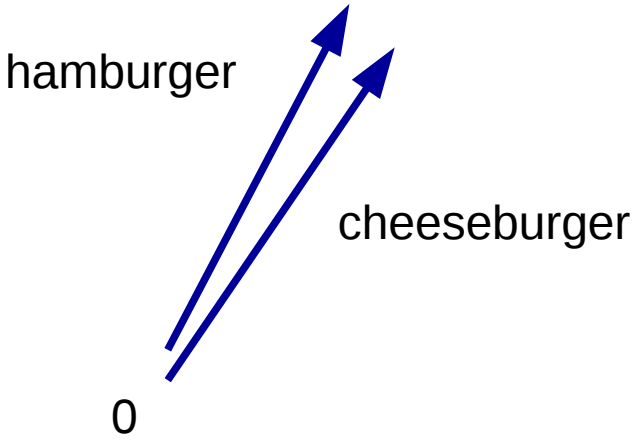
hamburger = $(0.4, 12.1, 0, 10)^T$



Word2Vec

$$\text{hamburger} = (0.4, 12.1, 0, 10)^T$$

$$\approx (0.5, 13, 0, 9.8)^T = \text{cheeseburger}$$

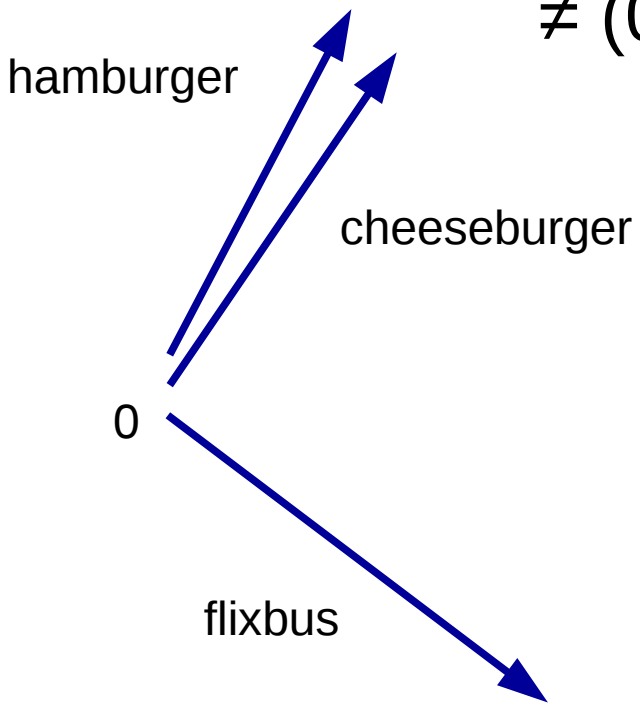


Word2Vec

hamburger = $(0.4, 12.1, 0, 10)^T$

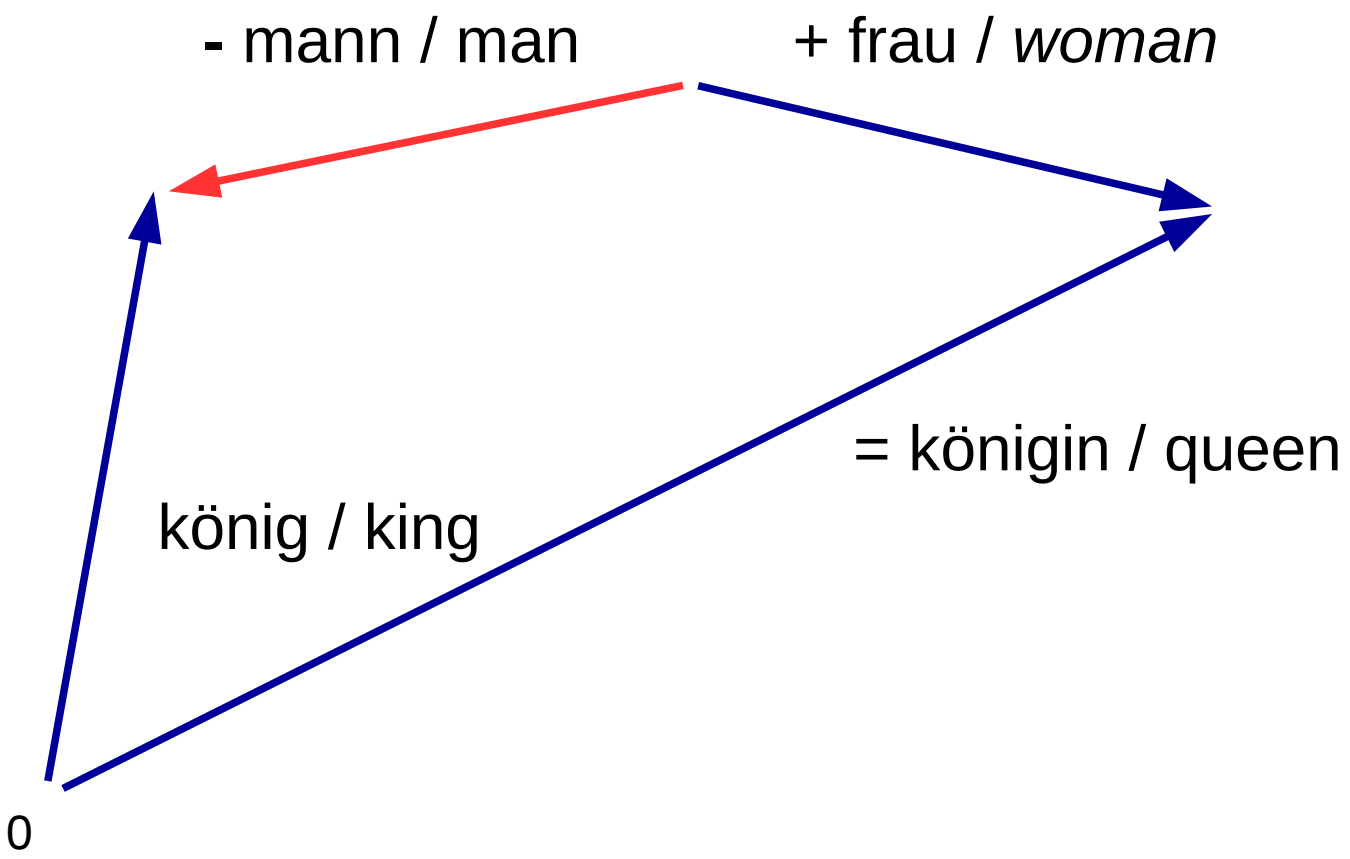
$\approx (0.5, 13, 0, 9.8)^T$ = cheeseburger

$\neq (0.1, 7, 42.1, 0)^T$ = flixbus



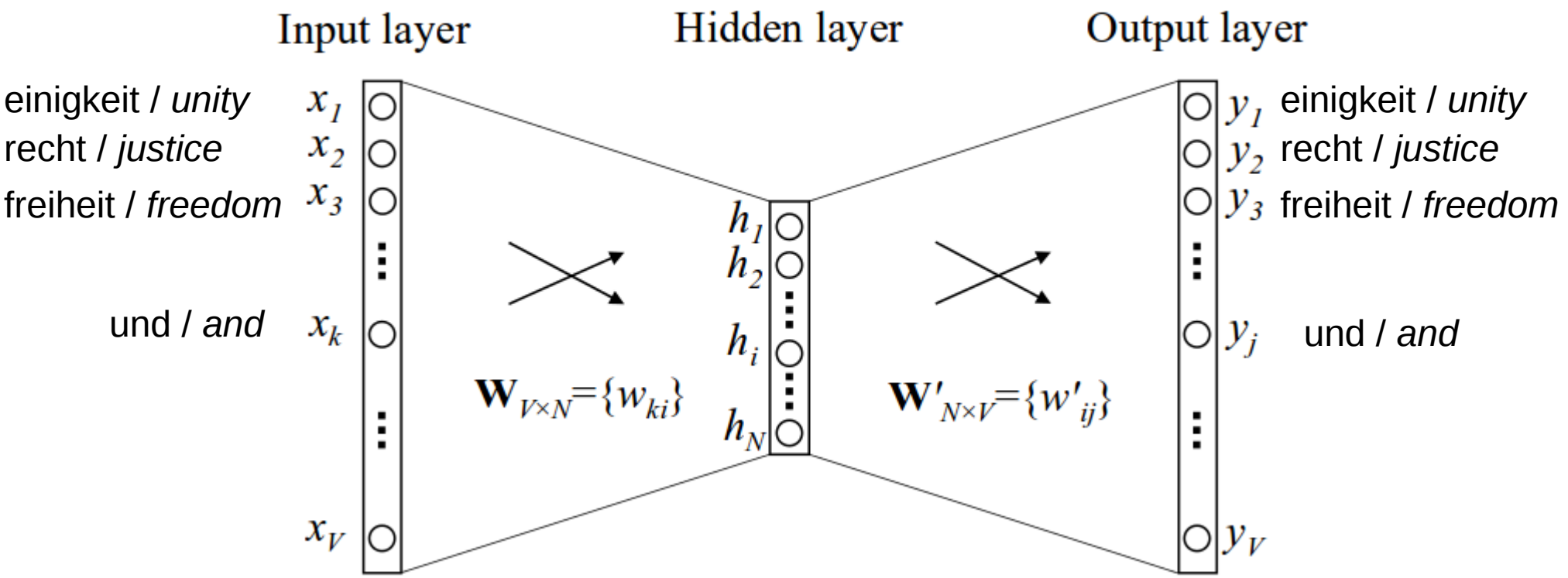
$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\ ||\mathbf{v}||}$$

Word2Vec



Word2Vec

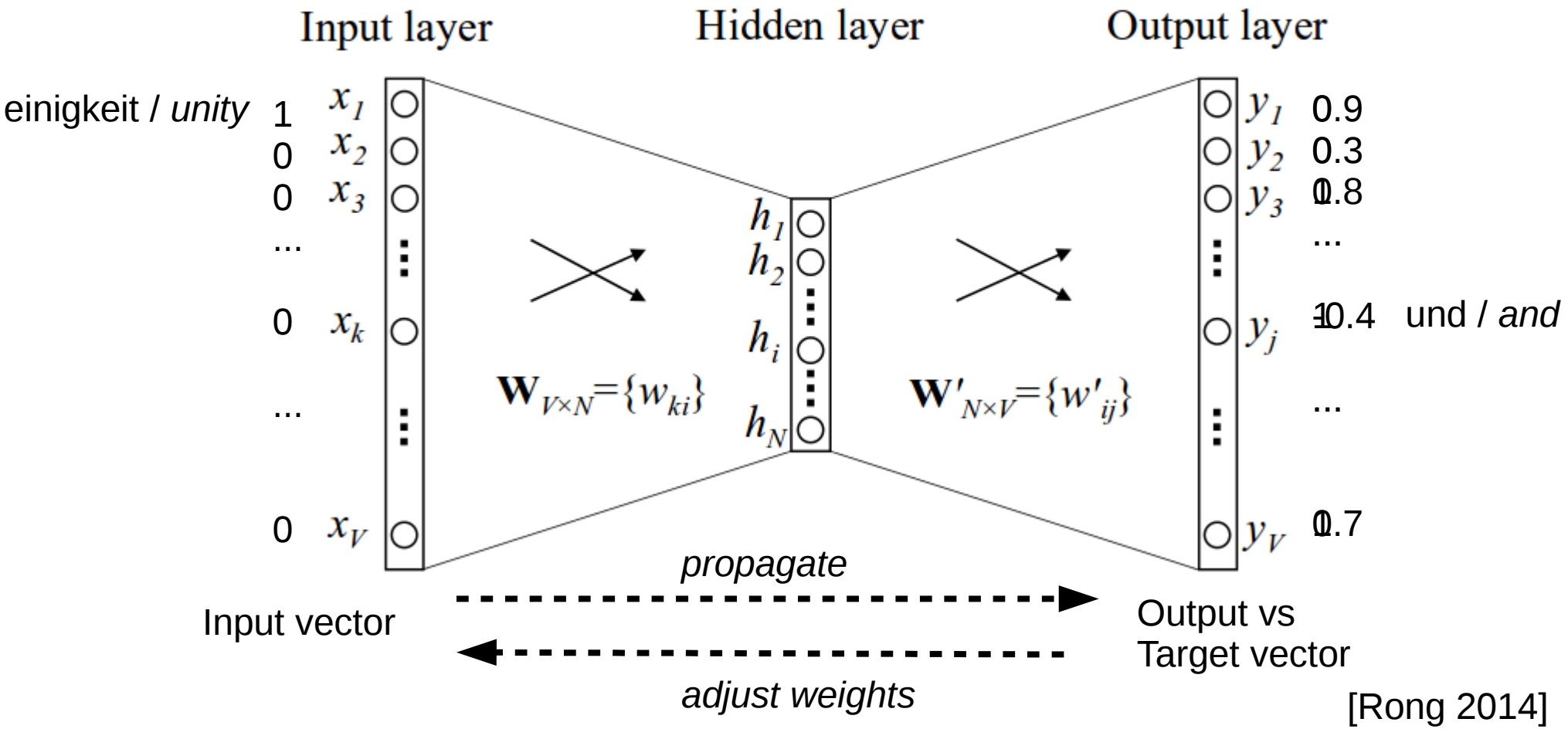
1st word **2nd word**



[Rong 2014]

Word2Vec

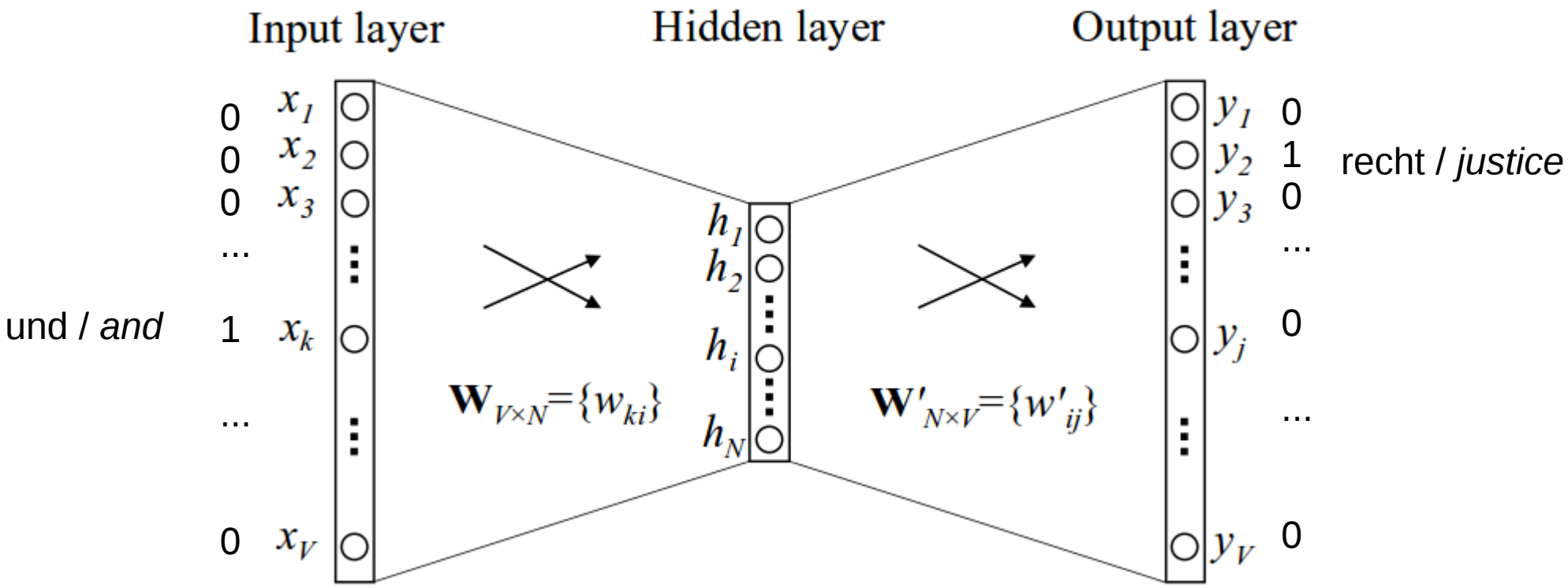
[... einigkeit, und, recht, und, freiheit, für, das, ...]
[... unity, and, justice, and, freedom, for, the, ...]



[Rong 2014]

Word2Vec

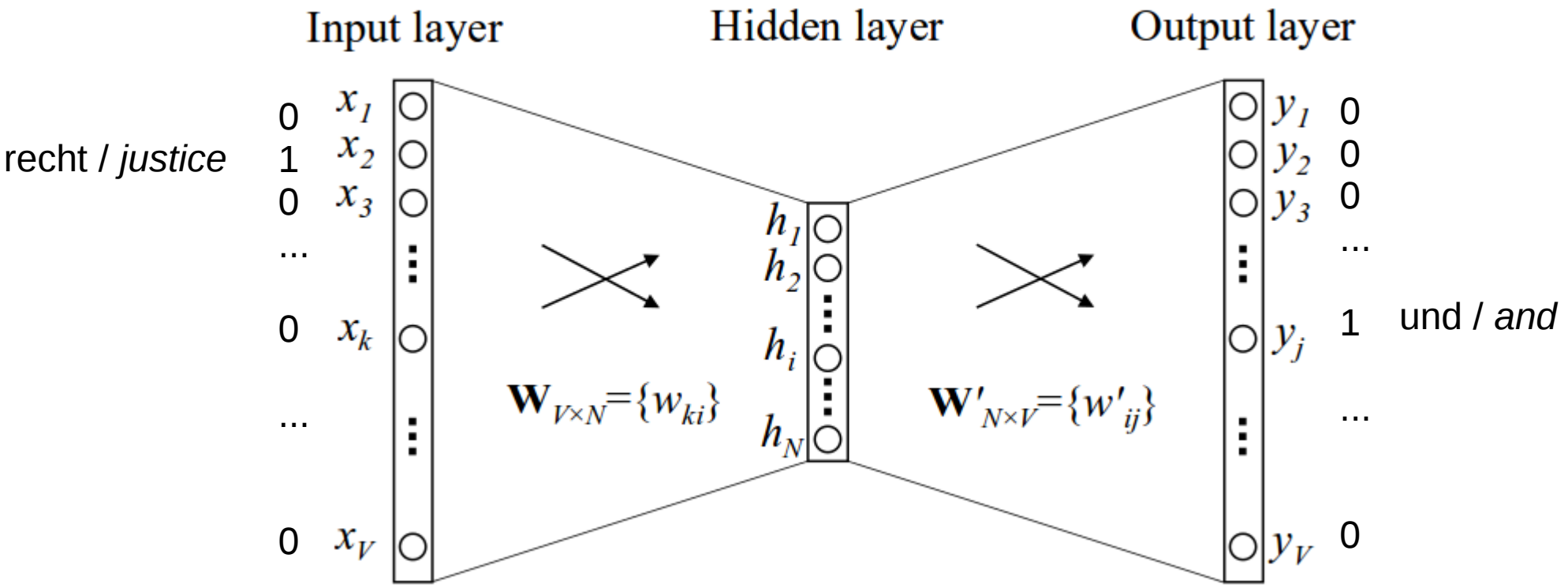
[... einigkeit, und, recht, und, freiheit, für, das, ...]
[... unity, and, justice, and, freedom, for, the, ...]



[Rong 2014]

Word2Vec

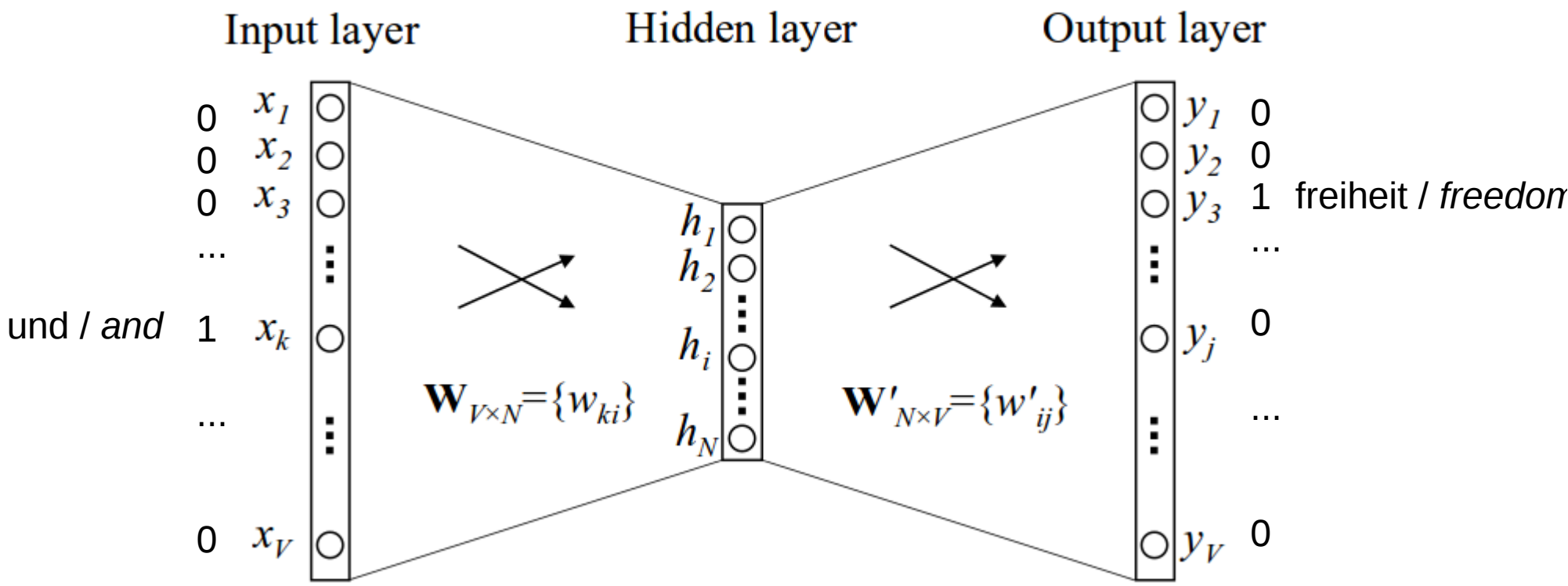
[... einigkeit, und, recht, und, freiheit, für, das, ...]
[... unity, and, justice, and, freedom, for, the, ...]



[Rong 2014]

Word2Vec

[... einigkeit, und, recht, und, freiheit, für, das, ...]
[... unity, and, justice, and, freedom, for, the, ...]

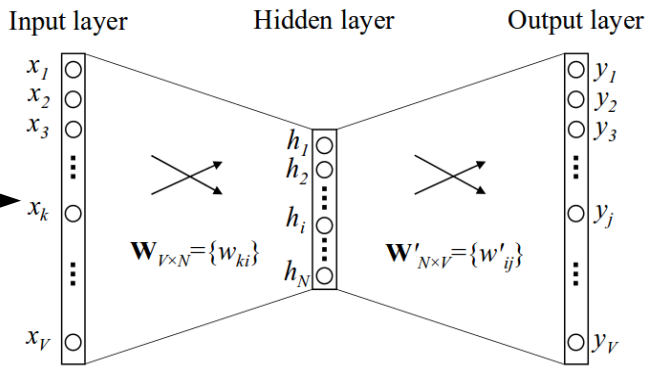


[Rong 2014]

Word2Vec

$$p(\text{2nd word} | \text{1st word} = \text{und})$$

0
0
0
...
und / *and* 1
...
0



0.1 recht / *justice*

0.1 freiheit / freedom

0.05 hand / *hand*

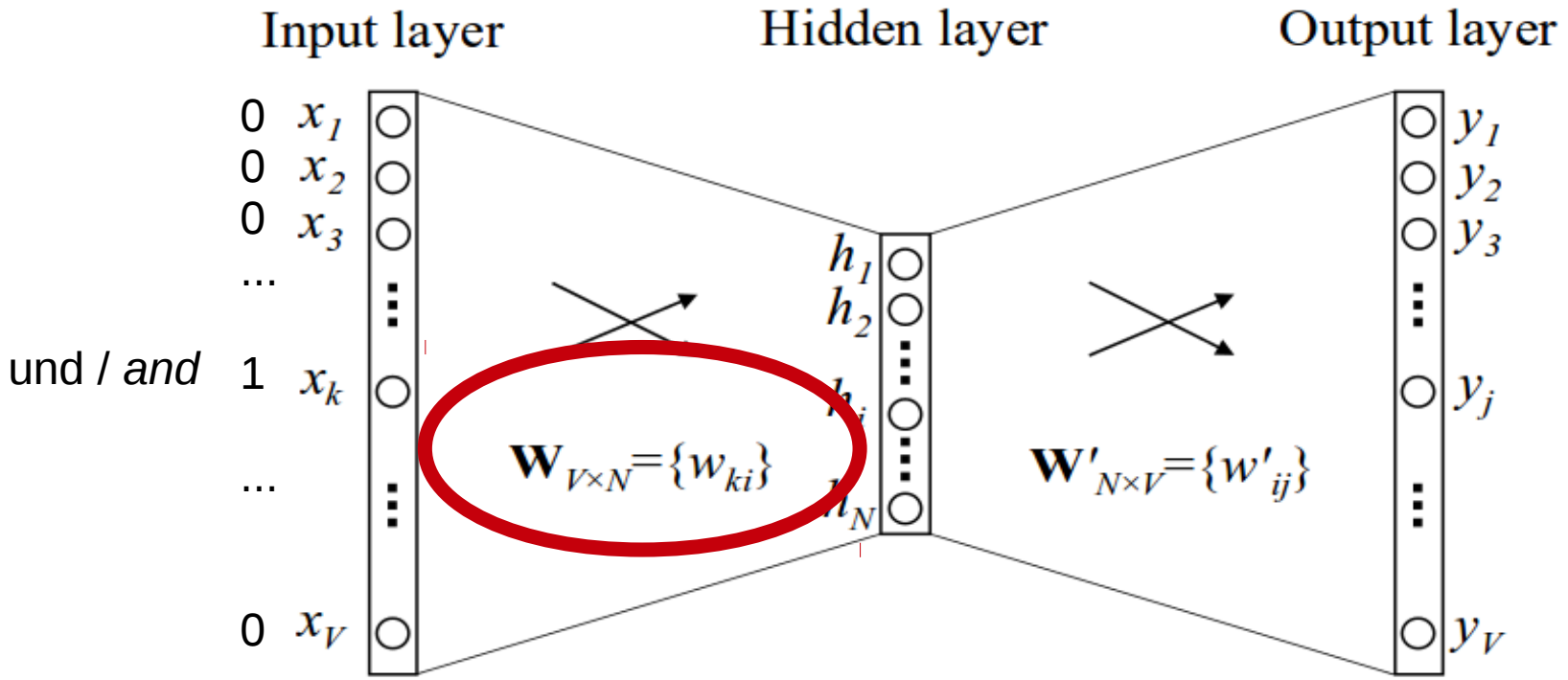
...

$$\mathbf{h} = \mathbf{W}^T\mathbf{x} = \mathbf{w}_k^T$$

$$y_j = \frac{\exp(\mathbf{w_j'}^T\mathbf{h})}{\sum_n \exp(\mathbf{w_n'}^T\mathbf{h})}$$

[Rong 2014]

Word2Vec



und / *and*

Fixed length (N=100-300) weight vector:

und / *and* => $(0.9, 42.42, 3333, 1.01, ...)^T$

[Rong 2014]

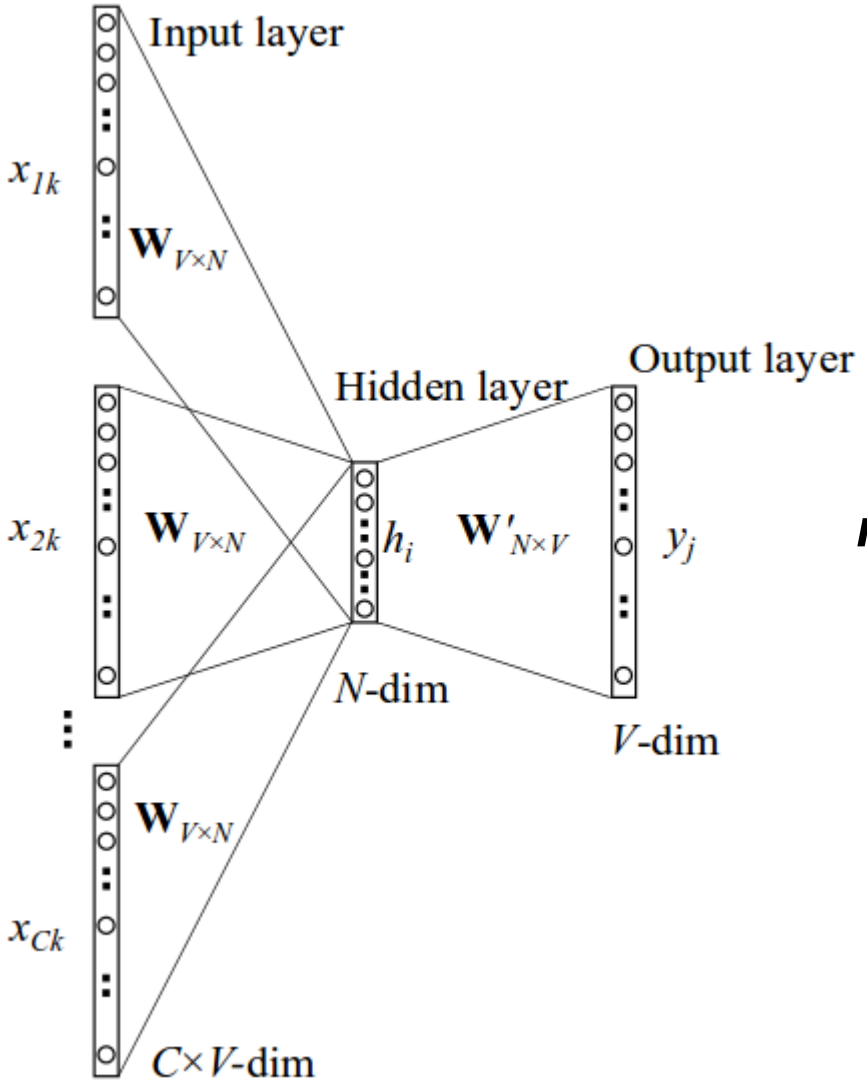
Word2Vec

***n-k*th word**

***n-k+1*th word**

...

***n+k*th word**



***n*th word**

[... einigkeit, und, recht, und, freiheit, für, das, ...]

[... unity, and, justice, and, freedom, for, the, ...]

[Rong 2014]

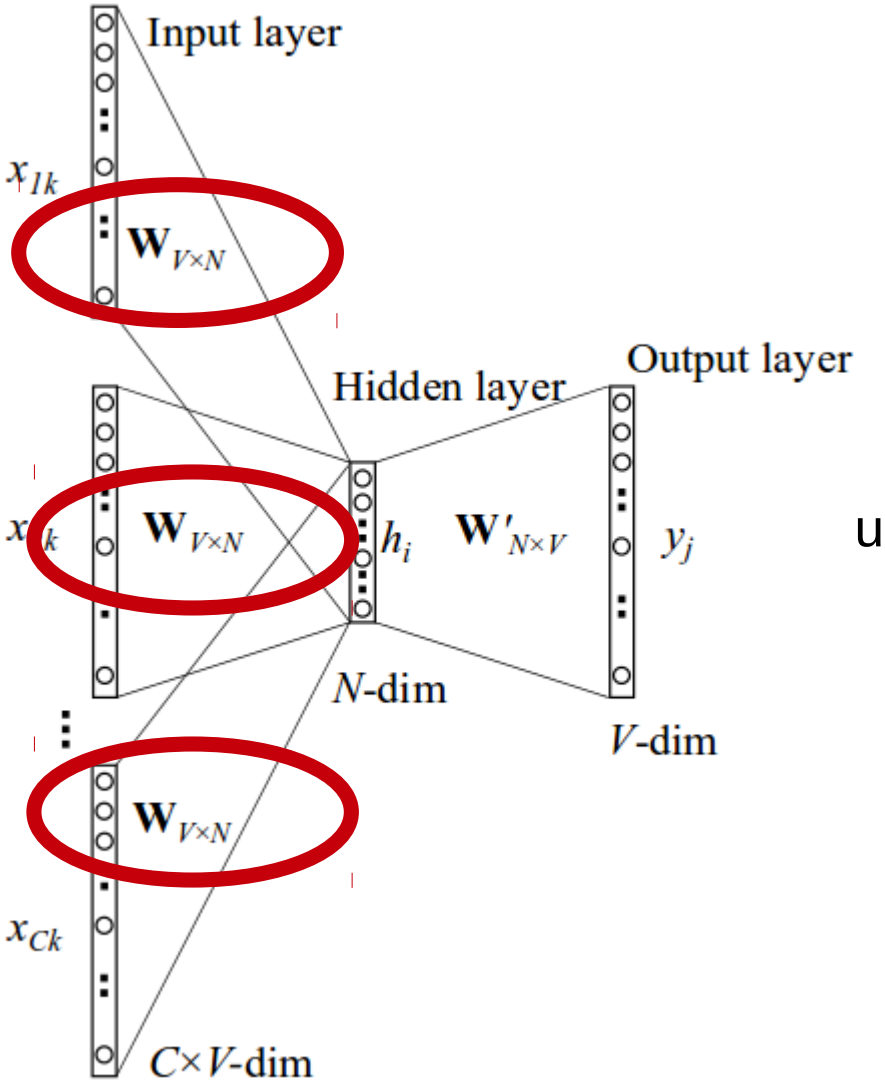
Word2Vec

und / and

recht / justice

...

für / for



und / *and*

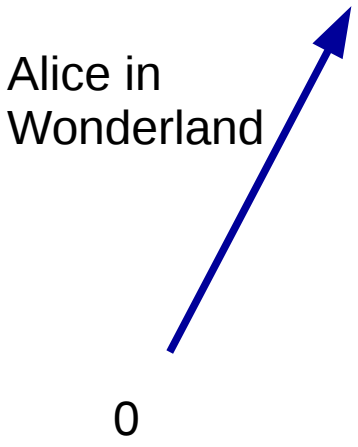
[... einigkeit, und, recht, und, freiheit, für, das, ...]

[... unity, and, justice, and, freedom, for, the, ...]

[Rong 2014]

Doc2Vec

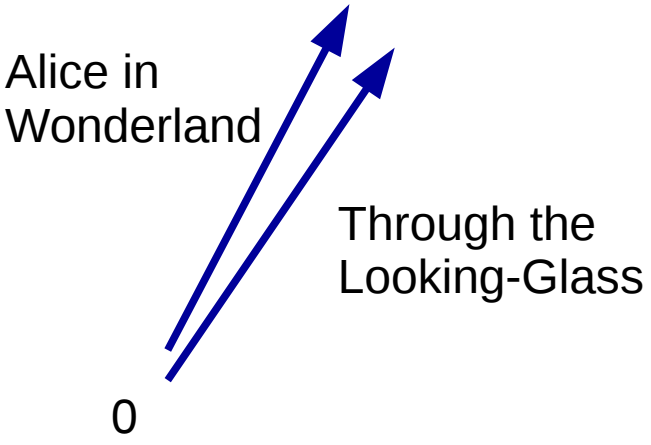
'Alice in Wonderland' = $(1.4, 2.1, 0)^T$



Doc2Vec

'Alice in Wonderland' = $(1.4, 2.1, 0)^T$

$\approx$ $(1.5, 3, 0)^T$ = 'Through the Looking-Glass'

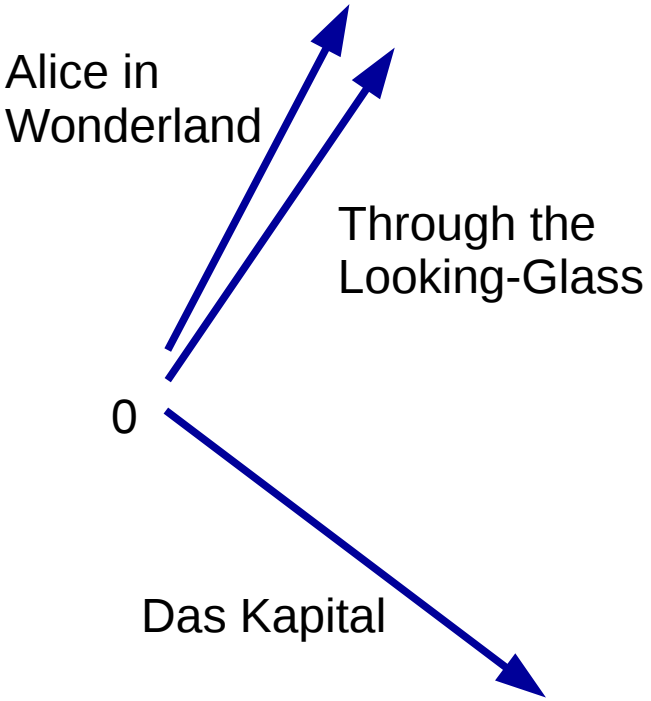


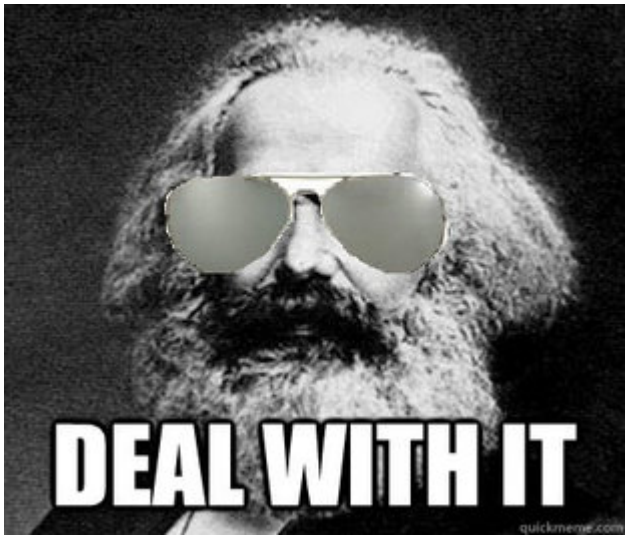
Doc2Vec

'Alice in Wonderland' = $(1.4, 2.1, 0)^T$

$\approx (1.5, 3, 0)^T$ = 'Through the Looking-Glass'

$\neq (33, 7, 1.2)^T$ = 'Das Kapital'





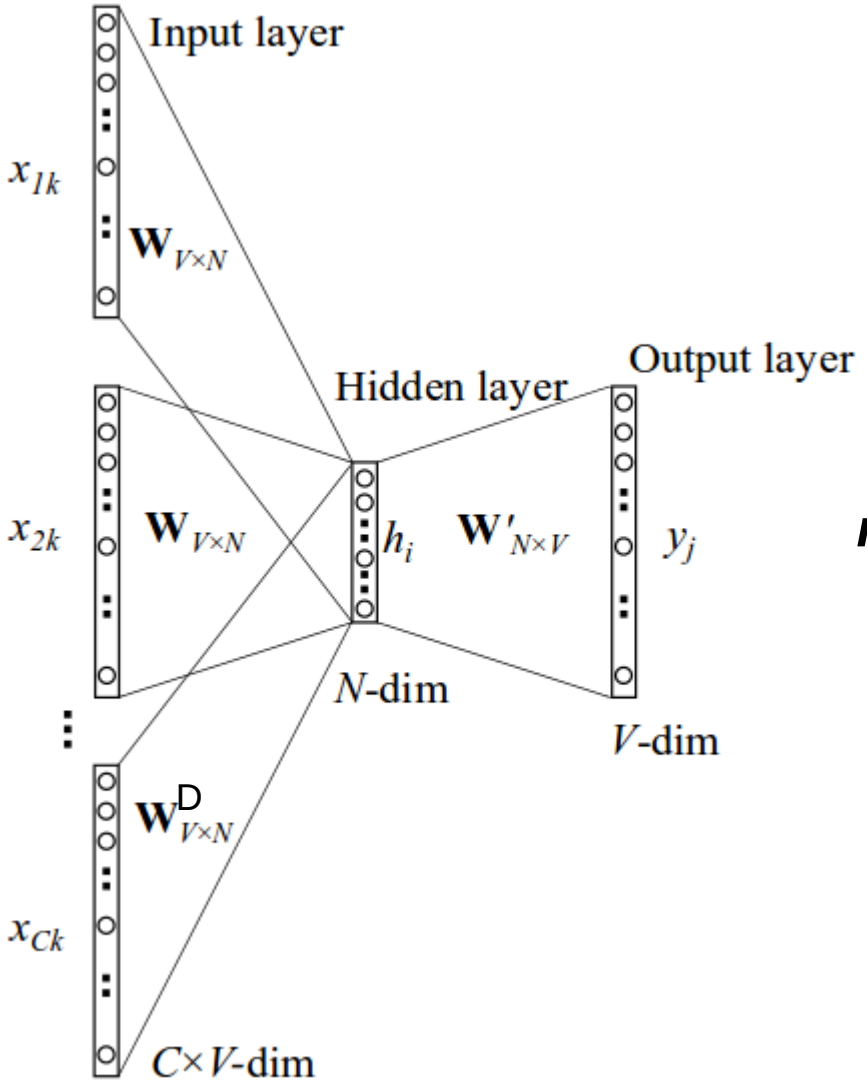
Doc2Vec

***n-k*th word**

***n-k+1*th word**

...

document



***n*th word**

[Rong 2014]

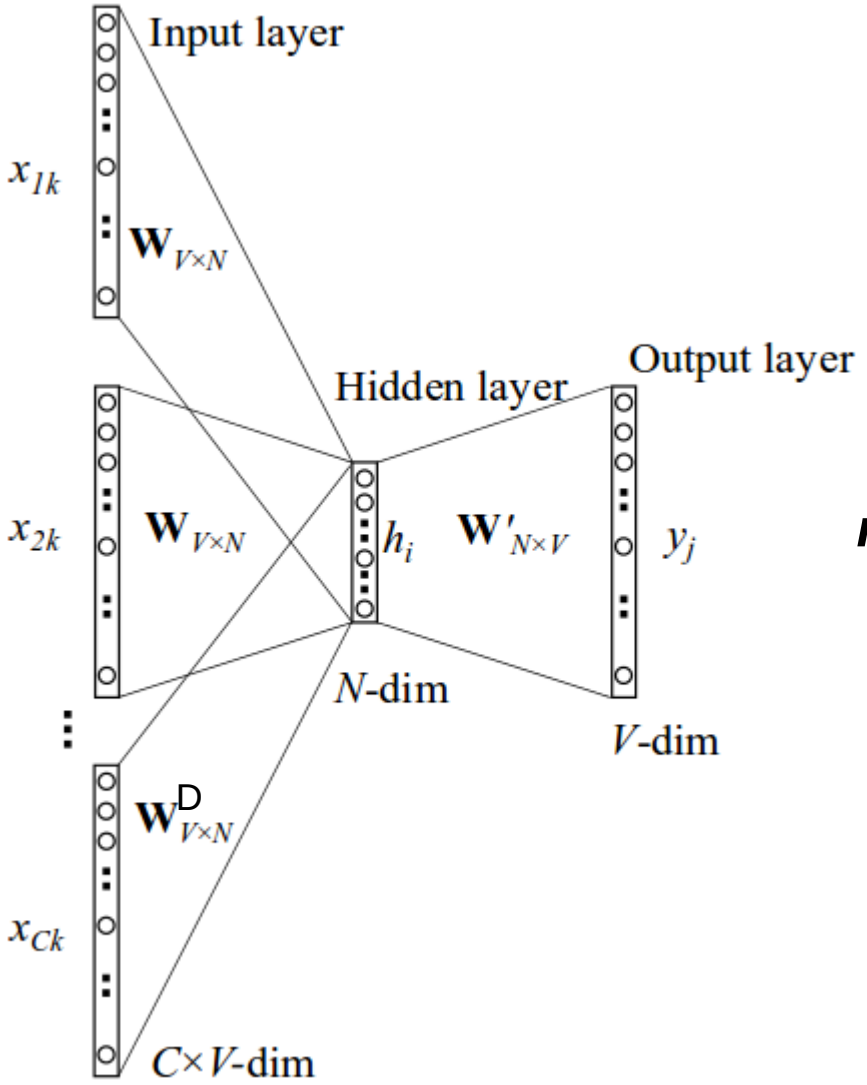
Doc2Vec

***n-k*th word**

***n-k+1*th word**

...

document



***n*th word**

[... einigkeit, und, recht, und, freiheit, für, das, ...]
[... unity, and, justice, and, freedom, for, the, ...]

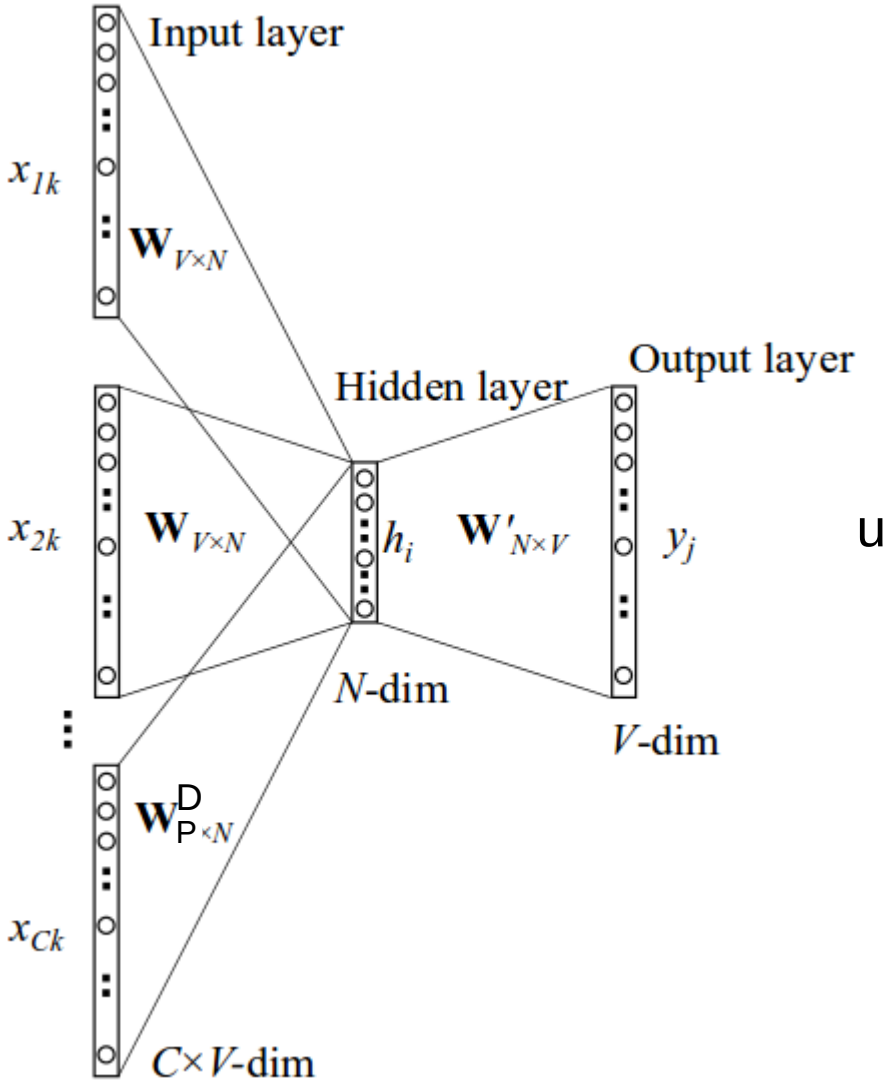
[Rong 2014]

Doc2Vec

recht / justice

freiheit / freedom

doc 1



und / *and*

[... einigkeit, und, recht, und, freiheit, für, das, ...]
[... unity, and, justice, and, freedom, for, the, ...]

[Rong 2014]

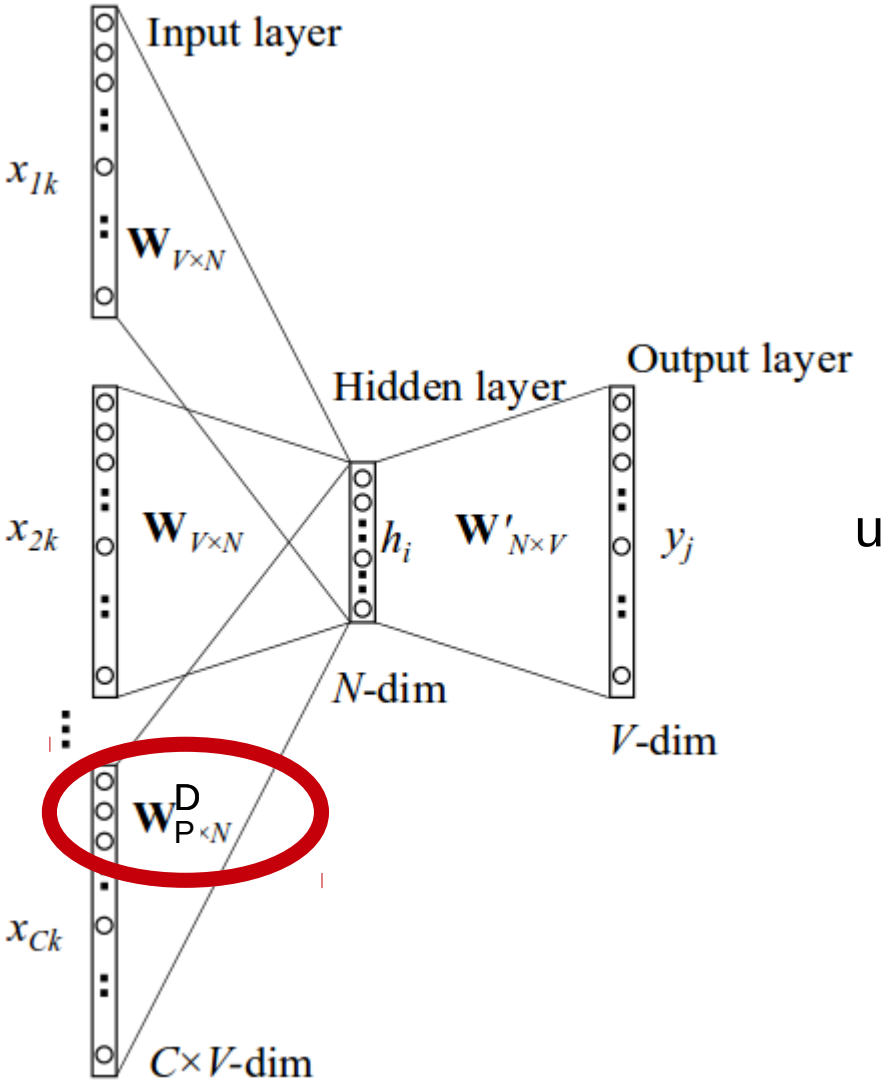
Doc2Vec

recht / justice

freiheit / freedom

und / *and*

doc 1



Fixed length (N=100-300) weight vector:

doc 1 => $(0.9, 42.42, 3333, 1.01, ...)^T$

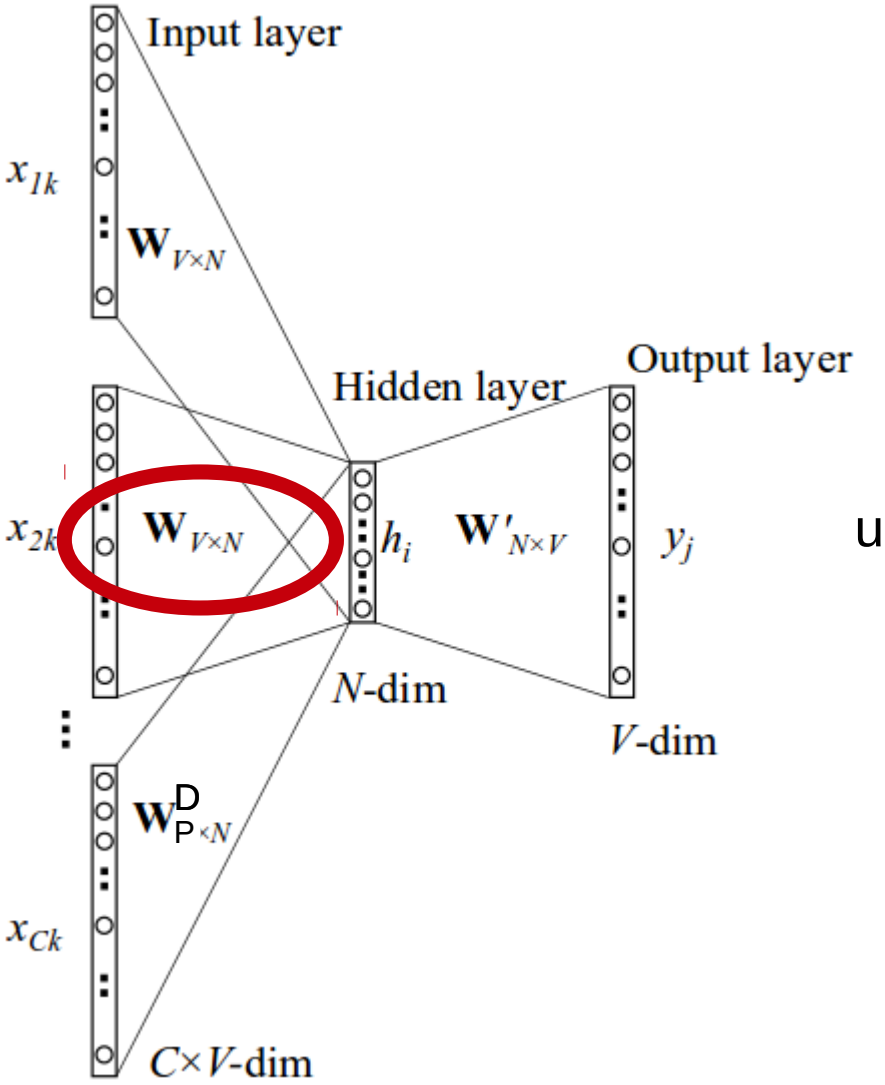
[Rong 2014]

Doc2Vec

recht / justice

freiheit / freedom

doc 1



und / *and*

Fixed length (N=100-300) weight vector:

freiheit / freedom => $(0.4, 13.13, 0.1, 19, ...)^T$

[Rong 2014]

Now what?

- Training scraped online comments on Doc2Vec...

```
import gensim.models.doc2vec as d2v
```

What do we get?

- Similarity
  - Auto / *car*
    - -> Fahrzeug / *automobile* (0.72)
  - Lügenpresse / *fake news*
    - -> Gutmensch / *do-gooder* (0.4)
    - -> Putin-Versteher / *Putin's disciple* (0.36)
    - -> Verschwörungstheorie / *conspiracy theory* (0.33)
  - NPD
    - -> CDU (0.4)
    - -> CSU (0.37)
    - -> FIFA (0.34)

What do we get?

- Arithmetic
  - Brexit – England + Griechenland / *Greece* =
    - Schuldenschnitt / *haircut* (0.36)
    - Grexit (0.34)
  - Hitler + Putin =
    - Erdogan (0.57)
  - König / *king* – Mann / *man* + Frau / *woman* =
    - Angela (0.59)

Supervised Learning

[trump, gewinnt, us, wahl, ...] → Doc2Vec Embedding → $(33, 1, 0, ..)^T$

$(33, 1, 0, ..)^T$ → Machine Learning Classifier →

0.2 ZEIT

0.7 SPON

0.1 FOCUS

Can we make predictions?

- It's (very) difficult

"Es war klar, dass solche Dinge kommen würden."

'It's clear that these things would come.'

DIE ZEIT

ML Classifier

- Linear SGD Classifier with *elasticnet* penalty

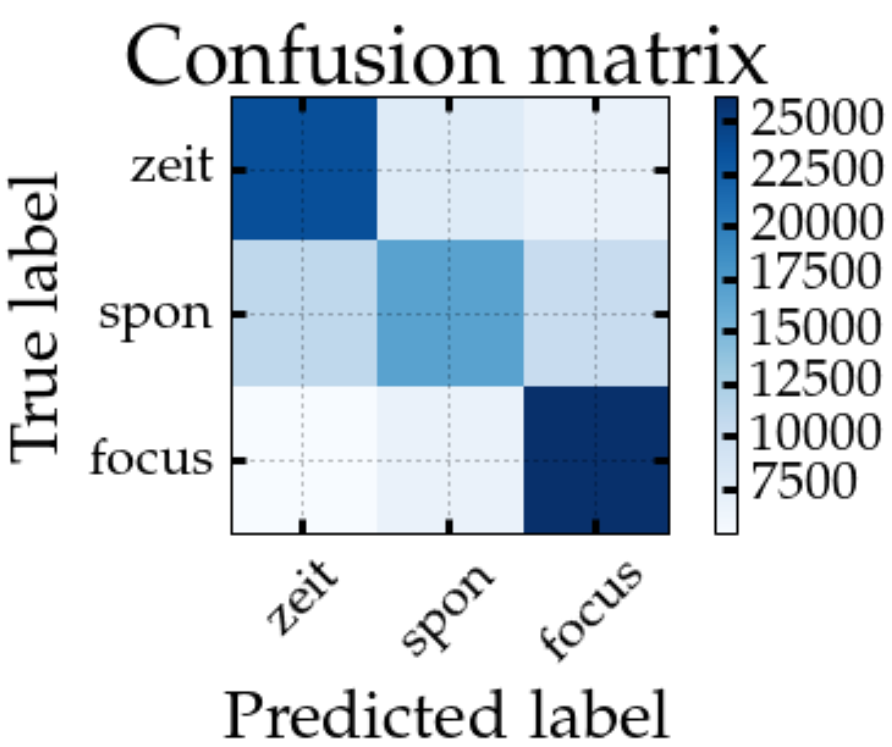
```
from sklearn.linear_model import SGDClassifier
```

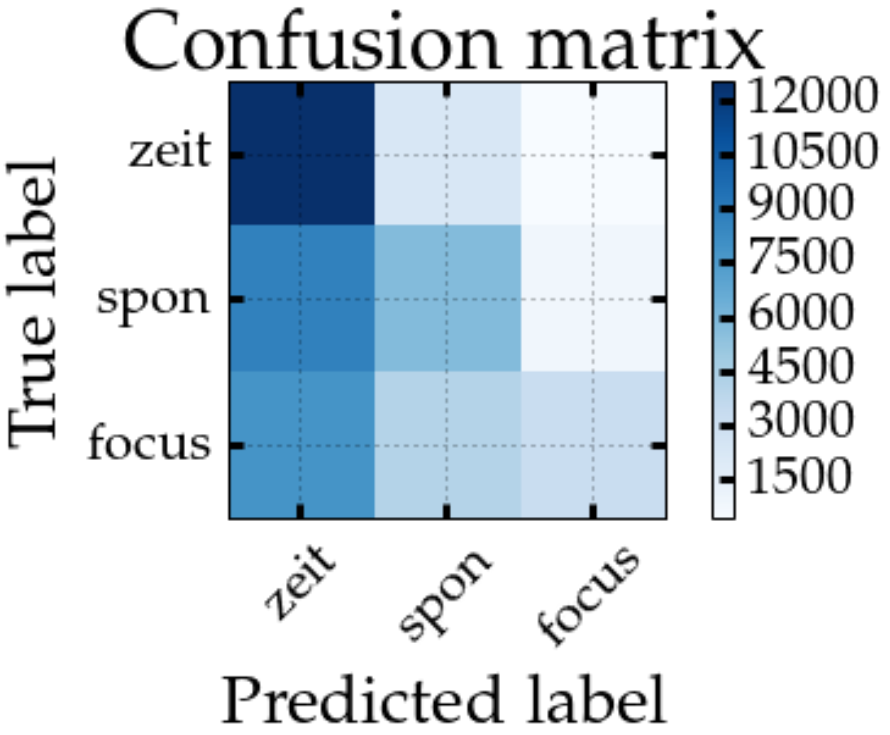
  - Doc2Vec embeddings as input
    - ZEIT, SPON, Focus class labels as output
  - Stratified training and test sets
    - ~35,000 and ~15,000 comments per news site

Can we make predictions?

- Linear SGDClassifier
  - Training Accuracy: **0.58**

Test Accuracy: **0.47**





Who is who?

Best representing:

“Es ist auch nicht nur zu eng für Lebewesen jenseits des Menschen, sondern letztendlich schon für uns selbst, weil es weniger anthroprozentrisch ist, als dass es eine sehr schmale Idee des Lebens impliziert...”

‘*Nor is it just too tight for living things beyond man, but ultimately even for ourselves, because it is less anthropocentric, as it implies a very narrow idea of life...’*

DIE ZEIT

Who is who?

“Das gesparte Geld landet dann aber nicht beim Hersteller des alten Toasters, weswegen dieser eben doch die Sollbruchstelle dort setzt, wo selbst der Fachmann nur mit der Flex hinkommt.”

‘*The manufacturer doesn't get the money saved on an old toaster, therefore the breaking point is placed where even the expert needs a Flex saw.’*

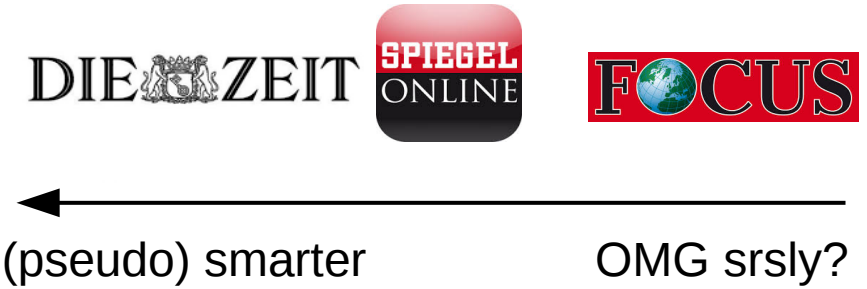
Who is who?

“Diese frauenverachtenden Muslime verstehen nur eine harte Hand und gehören sofort ausgewiesen, sofern sie sich als Asylanten hier aufhalten. Was Frau Merkel uns mit ihrer Politik der offenen Grenzen eingebrockt hat...”

‘*These misogynous Muslims understand only a hard hand and have to be deported immediately, if they reside as refugees here. This is what Mrs. Merkel got us with her open borders policy*...’

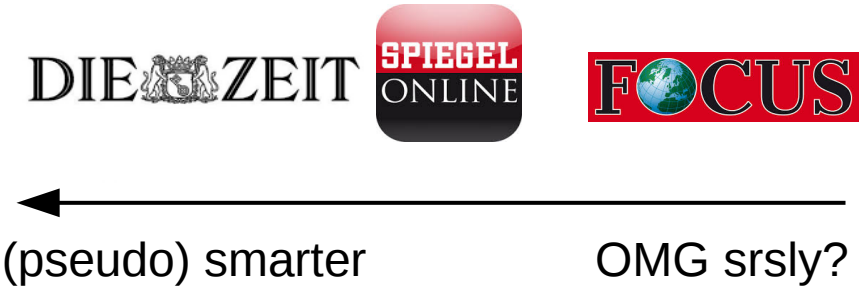
Recap

- Scrape comments from HTML source code
- Train Doc2Vec on user comments
- Interesting semantic relations
  - Hitler + Putin = Erdogan
- Reasonable performance
- Hypothesis supported



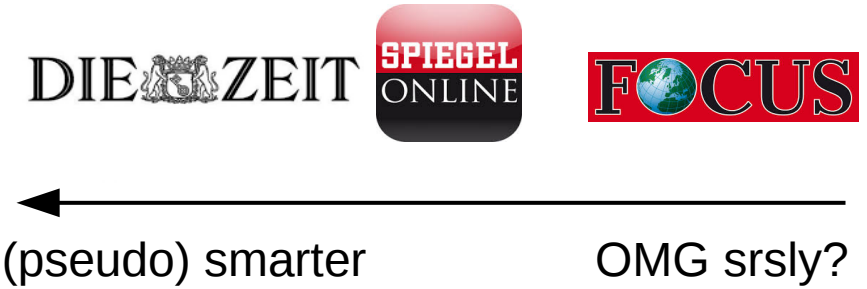
Thank You! Any Questions?

- Scrape comments from HTML source code
- Train Doc2Vec on user comments
- Interesting semantic relations
  - Hitler + Putin = Erdogan
- Reasonable performance
- Hypothesis supported



robert.meyer@flixbus.com

- Scrape comments from HTML source code
- Train Doc2Vec on user comments
- Interesting semantic relations
  - Hitler + Putin = Erdogan
- Reasonable performance
- Hypothesis supported



References

- Rong 2014: "W*ord2Vec Parameter Learning Explained*" https://arxiv.org/pdf/1411.2738v4.pdf

- Quoc and Nikolov 2014: "*Distributed Representation of Sentences and Documents*" https://arxiv.org/pdf/1405.4053v2.pdf