Gruppe 2
Word2Vec

Rosanna
Konstantin
Laura

Problem 2

Trainieren eines Word2Vec Modells und Darstellung von Wort- und Dokumentenvektoren

Word-Embeddings

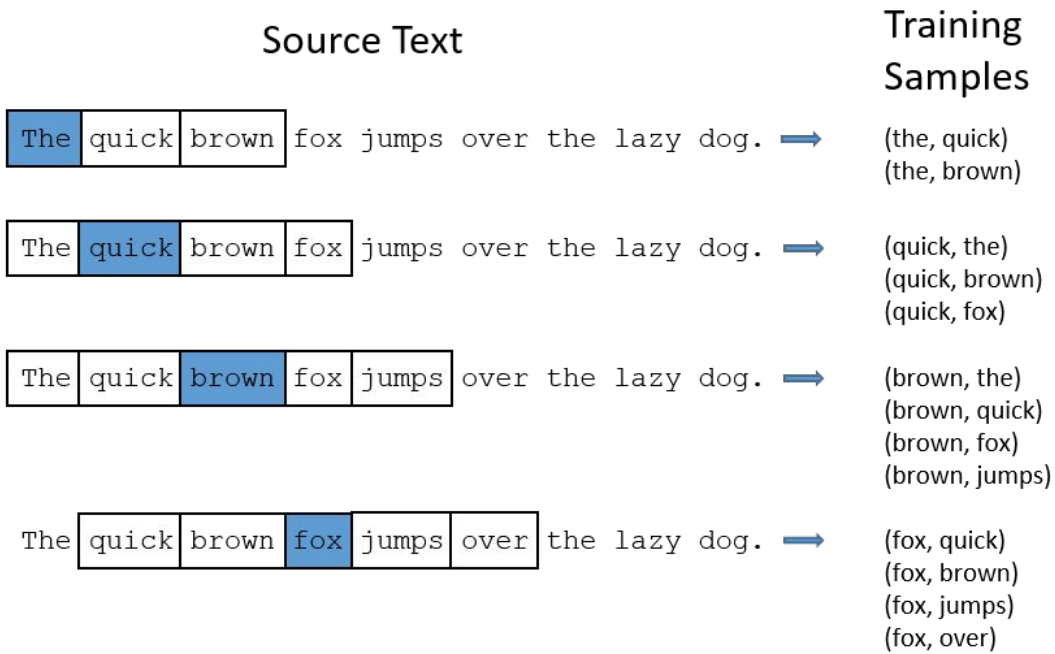
“king”

“Man”

“Woman”

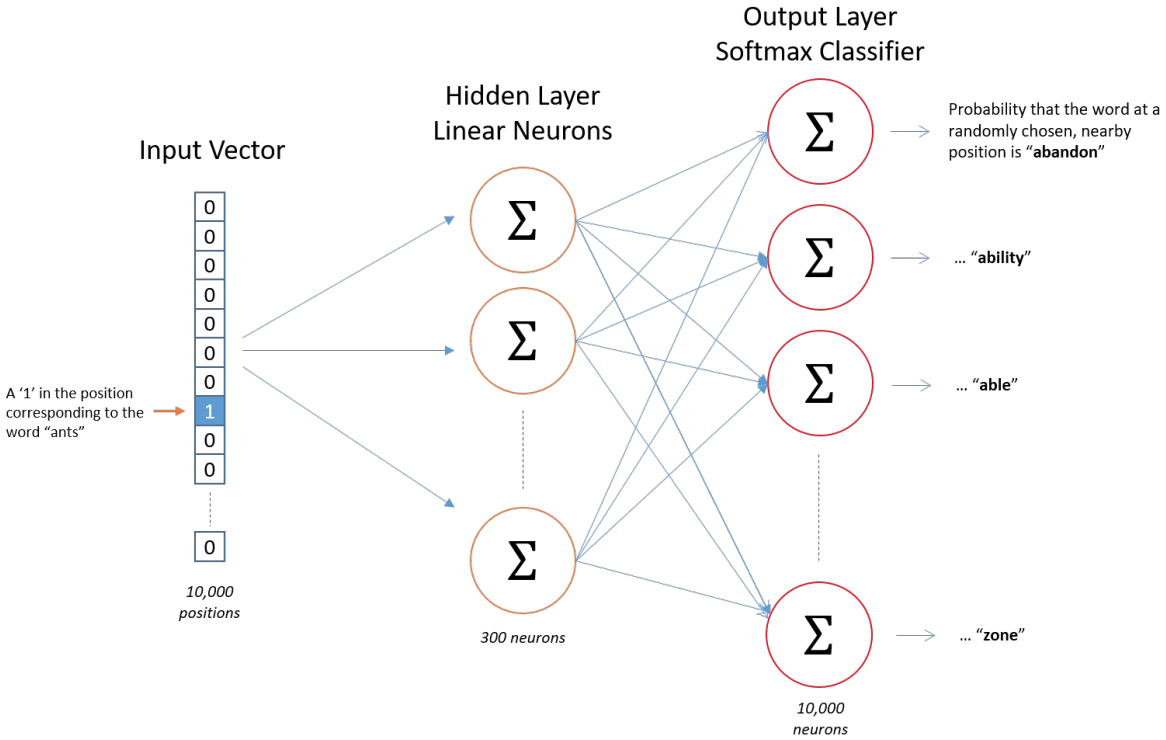
Quelle: [WED]

Context Window



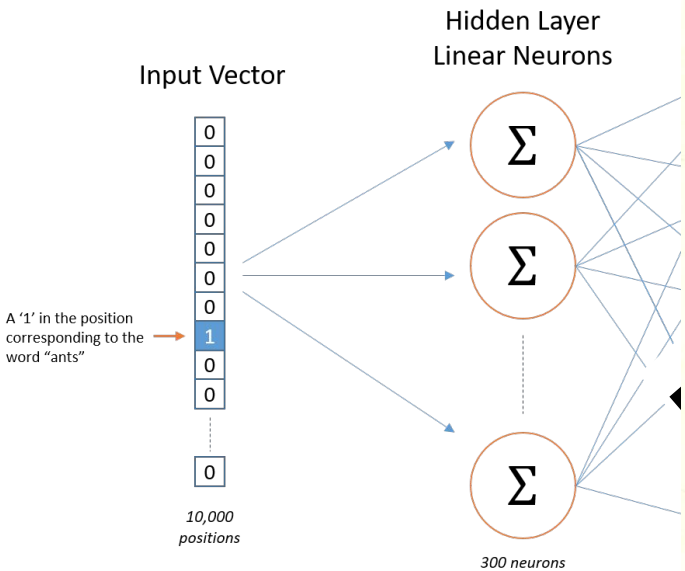
Quelle: [SWS]

Skip-Gram Architektur



Quelle: [SAM]

Skip-Gram Architektur



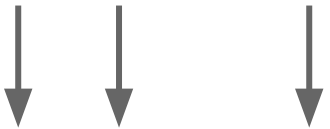
Document-Embeddings

Satz = [‘heute’, ‘schön’, ‘tag’]

Embeddings = [

[2, 4, 5], # ‘heute’
[6, 3, 8], # ‘schön’
[1, 7, 3] # ‘tag’

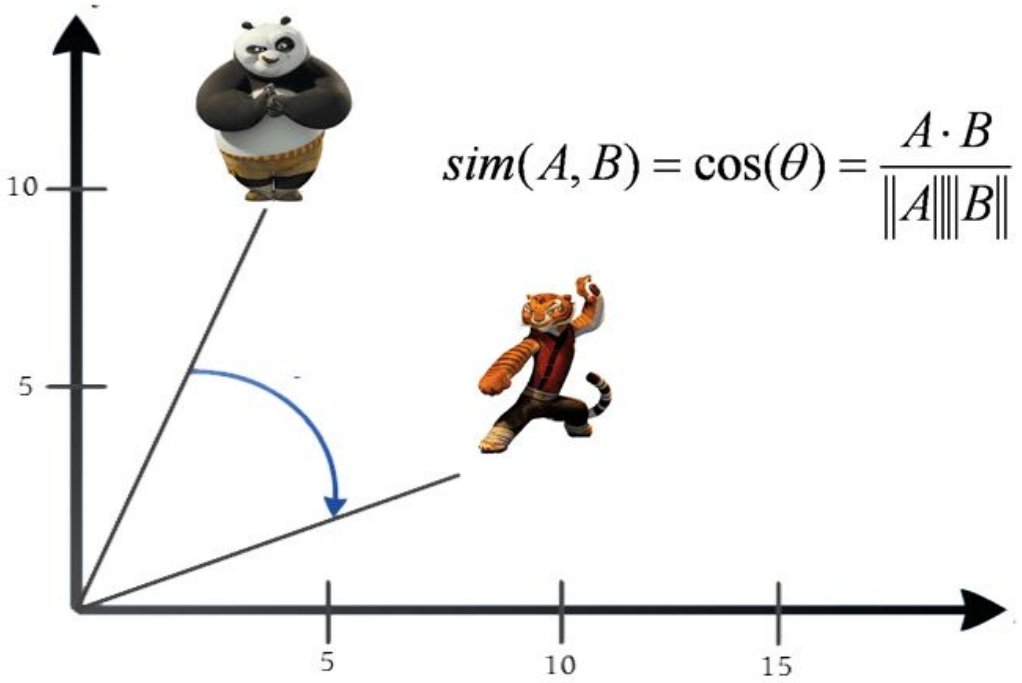
]



spaltenweise mitteln

Doc-Embedding = [3, 4.66, 5.33]

Cosine Similarity



$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Quelle: [CSM]

Implementierung

Preprocessing ➔ Lookup-Tables ➔ Context-Window ➔ Modelling ➔ Training ➔ ...

- Reduktion der Texte auf bedeutungstragende Begriffe
  - kleingeschrieben & lemmatisiert

Satz: **“Heute ist ein schöner sonniger Tag”**

Ergebnis: [**‘heute’, ‘schön’, ‘sonnig’, ‘tag’**]

Implementierung

Preprocessing → **Lookup-Tables** → Context-Window → Modelling → Training → ...

- Indizierung der Wörter und generieren von Lookup-Tables

Daten: **[[‘heute’, ‘schön’, ‘sonnig’, ‘tag’], [...], ...]**

Worthäufigkeiten Lookup-Table

```
Word2Index = {
        ‘heute’: 0,
        ‘schön’: 1,
        ...
}
```

```
Index2Word = {
        0:  ‘heute’,
        1:  ‘schön’,
        ...
}
```

SentencesAsIndex = [[0, 1, 2, 3], [...], ...]

SentencesAsIndexFlat = [0, 1, 2, 3, ...]

VocabSize = 4

Implementierung

Preprocessing → Lookup-Tables → **Context-Window** → Modelling → Training → ...

Daten: [[‘heute’, ‘schön’, ‘sonnig’, ‘tag’], [...], ...]

SentencesAsIndex = **[[0, 1, 2, 3], [...], ...]**

Window-Size = 2

Targets = [0, 0, 1, 1, 1, 2, 2, 2, 3, 3]
Labels = [1, 2, 0, 2, 3, 0, 1, 3, 1, 2]

0 ‘heute’ → 1 ‘schön’
0 ‘heute’ → 2 ‘sonnig’
1 ‘schön’ → 0 ‘heute’
1 ‘schön’ → 2 ‘sonnig’
1 ‘schön’ → 3 ‘tag’
2 ‘sonnig’ → 0 ‘heute’
...

Implementierung

Context-Window → **Modelling → Training → Document-Embeddings** → Cosine-Similarity

- Modelling: Hyperparameter Tuning
- Training: Word-Embeddings
- Word-Embeddings -> Document Embeddings

Implementierung

Context-Window ➔ Modelling ➔ Training ➔ Document-Embeddings ➔ **Cosine-Similarity**

- **Positivbeispiel:**
  - Titel #1: Fußgänger angefahren
  - Titel #2: Radfahrerin bei Verkehrsunfall schwer verletzt

  ➔ Cosine-Sim.: 0.961

- **Negativbeispiel:**
  - Titel #1: Walpurgisnacht - Demonstrationen und Feiern nahezu störungsfrei
  - Titel #2: Fußgänger schwer verletzt

  ➔ Cosine-Sim.: 0.788

Problem 3

- Training eines Word2Vec Modells mit gensim
- t-SNE: Dimensionsreduktion auf 3D
- OPTICS Clustering

Word2Vec mit gensim

Word2Vec → Doc2Vec → t-SNE → OPTICS Clustering

```
model.wv.most_similar("polizist")
```

```
[('beamte', 0.9247633218765259),
 ('polizeibeamten', 0.8614054918289185),
 ('polizeikräfte', 0.8405277729034424),
 ('polizeibeamte', 0.7810276746749878),
 ('einsatzkräfte', 0.7457624673843384),
 ('fahnder', 0.7403584718704224),
 ('kraft', 0.7213431596755981),
 ('zivilpolizisten', 0.6968176364898682),
 ('polizeibeamter', 0.6833140850067139),
 ('unterstützungskräfte', 0.6811661720275879)]
```

```
model.wv.similar_by_word("moabit")
```

```
[('wedding', 0.9733237028121948),
 ('gesundbrunnen', 0.9623279571533203),
 ('charlottenburg', 0.9563843011856079),
 ('kreuzberg', 0.9533096551895142),
 ('wilmersdorf', 0.9530558586120605),
 ('reinickendorf', 0.9519842863082886),
 ('neukölln', 0.9516151547431946),
 ('lichtenberg', 0.9493242502212524),
 ('hellersdorf', 0.9483109712600708),
 ('lichterfelde', 0.9443531632423401)]
```

Word2Vec mit gensim

Word2Vec → Doc2Vec → t-SNE → OPTICS Clustering

```
model.wv.similar_by_word("pkw")
```

```
[('motorrad', 0.9521627426147461),
 ('lkw', 0.9259902834892273),
 ('renault', 0.9126736521720886),
 ('transporter', 0.9057509899139404),
 ('ford', 0.9015751481056213),
 ('roller', 0.8971459865570068),
 ('skoda', 0.8936470746994019),
 ('polizeifahrzeug', 0.8934042453765869),
 ('vw', 0.892335832118988),
 ('smart', 0.8907783031463623)]
```

```
model.wv.similar_by_word("fahrrad")
```

```
[('offensichtlich', 0.8779125213623047),
 ('entlang', 0.8543713688850403),
 ('rad', 0.8502610325813293),
 ('hermannplatz', 0.8460982441902161),
 ('kantstraße', 0.8174165487289429),
 ('dortig', 0.8112874031066895),
 ('oranienstraße', 0.8107872605323792),
 ('smart', 0.8093620538711548),
 ('königsheideweg', 0.7993432283401489),
 ('ufer', 0.7992792129516602)]
```

Word2Vec mit gensim

Word2Vec → Doc2Vec → t-SNE → OPTICS Clustering

```
model.wv.similar_by_word("messer")
```

```
[('machete', 0.8842913508415222),
 ('schuss', 0.863178014755249),
 ('waffe', 0.8625810742378235),
 ('schusswaffe', 0.8596803545951843),
 ('verkäufer', 0.8549923896789551),
 ('hals', 0.8465339541435242),
 ('baseballschläger', 0.8385649919509888),
 ('drohen', 0.8383954167366028),
 ('kassiererin', 0.8378837704658508),
 ('pistole', 0.8363977670669556)]
```

```
model.wv.similar_by_word("droge")
```

```
[('marihuana', 0.9189831018447876),
 ('menge', 0.9019356369972229),
 ('betäubungsmittel', 0.9001079797744751),
 ('verkaufen', 0.8964988589286804),
 ('cannabis', 0.8861299157142639),
 ('tütchen', 0.8838708400726318),
 ('kilogramm', 0.8829858303070068),
 ('verkauf', 0.8799164295196533),
 ('beweismittel', 0.8726941347122192),
 ('diebesgut', 0.8662963509559631)]
```

Doc2Vec

Word2Vec → **Doc2Vec** → t-SNE → OPTICS Clustering

article	headline	processed_articles	processed_headlines	combinedText	doc_vector
Die Entschärfung der Weltkriegsbombe auf dem Gelände des Stadtguts Alt-Hellersdorf ist für Montagmittag geplant. Ab 9.30 Uhr beginnen die Absperr- und Evakuierungsmaßnahmen. Darüber hinaus wird eine Sicherheitszone eingerichtet, die folgende Straßenzüge umfassen wird: Zossener Straße zwischen Alte Hellersdorfer Straße ...	Geplante Bombenentschärfung am Montag - umfangreiche Absperr- und Evakuierungsmaßnahmen	...	...	[[geplante, bombenentschärfung, montag, umfangreiche, absperr-, evakuierungsmaßnahmen], [entschärfung, weltkriegsbombe, gelände, stadtguts, alt-hellersdorf, montagmittag, planen], [beginnen, absperr-, evakuierungsmaßnahmen], [hinaus, sicherheitszone, einrichten, folgend, straßenzüge, umfassen], [zossener, straße, alte, hellersdorfer], [straße, stendaler, straße, nördlich, stendaler, straße, zossener, straße, janusz-korczak-straße, östlich, cottbusser, straße, janusz-korczak-straße, ...]]	[0.002178671, -0.13364823, 0.06986375, -0.29327267, -0.086605564, -0.46204075, 0.060277093, -0.38269126, -0.051798865, 0.13829328, -0.14214486, 0.07997522, -0.1200359, 0.3690538, 0.24615626, -0.1824191, -0.037942264, -0.32919732, -0.08358068, -0.16176073, -0.017206812, 0.22714683, -0.033081092, -0.1987105, -0.03726253, -0.082241535, 0.31896037, 0.17566817, 0.2914497, -0.3590829, -0.06832161, -0.30331346, -0.31567678, -0.12815179, -0.02714962, 0.0565941, 0.18353476, 0.13733512, -0.02233113, -0.13247234, 0.03445474, -0.4400618, -0.02902167, 0.37574574, ...]

Doc2Vec

Word2Vec ➔ **Doc2Vec** ➔ t-SNE ➔ OPTICS Clustering

- **Positivbeispiel:**
  - Titel #1: Fußgänger angefahren
  - Titel #2: Radfahrerin bei Verkehrsunfall schwer verletzt

    ➔ **Cosine-Sim.: 0.978** (eigenes Modell: 0.961)

- **Negativbeispiel:**
  - Titel #1: Walpurgisnacht - Demonstrationen und Feiern nahezu störungsfrei
  - Titel #2: Fußgänger schwer verletzt

    ➔ **Cosine-Sim.: 0.48** (eigenes Modell: 0.788)

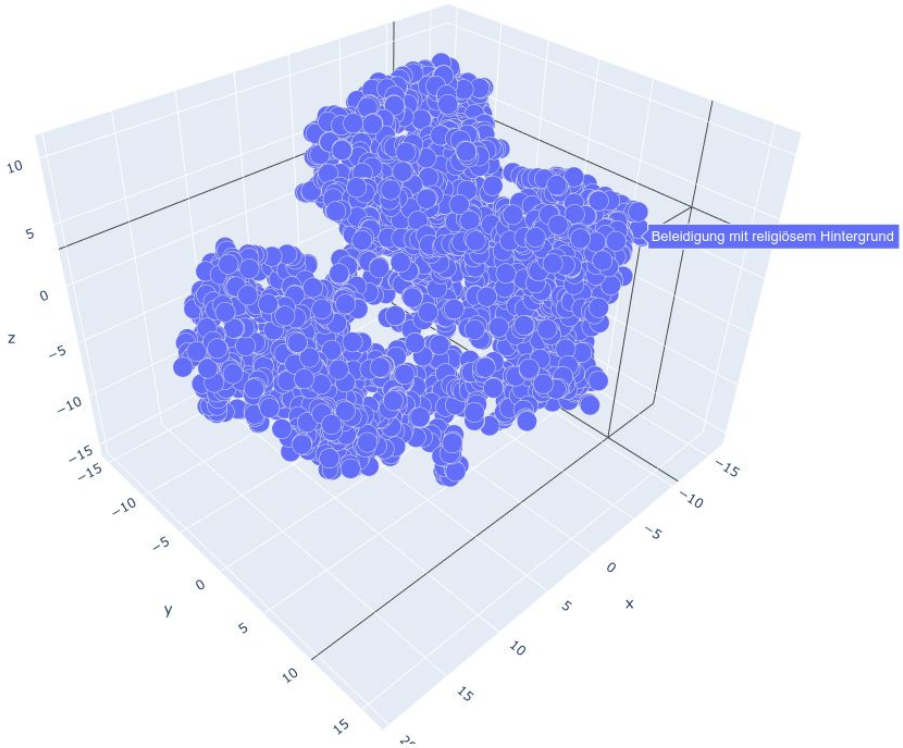
t-SNE Embeddings

Word2Vec → Doc2Vec → **t-SNE** → OPTICS Clustering

t-SNE: t-Distributed Stochastic Neighbor Embedding

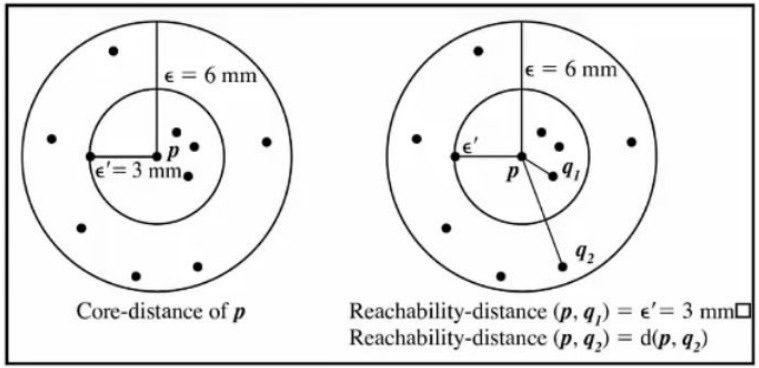
Haupteinsatzgebiet: Datenexploration und Visualisierung hochdimensionaler Datenräume

Reduzierung eines hochdimensionalen Datensets auf bspw. 2 oder 3 Dimensionen zur besseren Visualisierung
→ Anwendung auf unsere Dokumentenvektoren (hier: 100-dim → 3-dim)
→ Erkennen von Nachbarschaft/ semantischen Zusammenhängen möglich



OPTICS Clustering

Word2Vec ➔ Doc2Vec ➔ t-SNE ➔ **OPTICS Clustering**



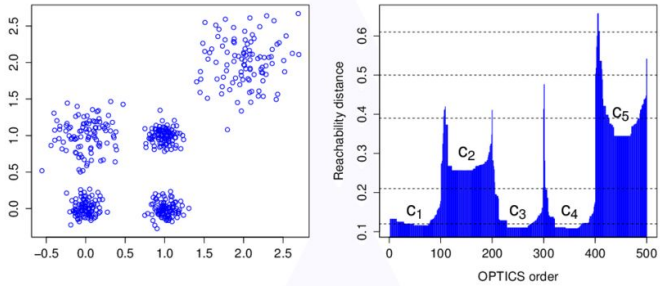
Quelle: [OPT1]

OPTICS: Ordering Points To Identify Clustering Structure

Dichte basierte Form der Clusterbildung

Anzahl der Cluster ergibt sich durch ***reachability plot***

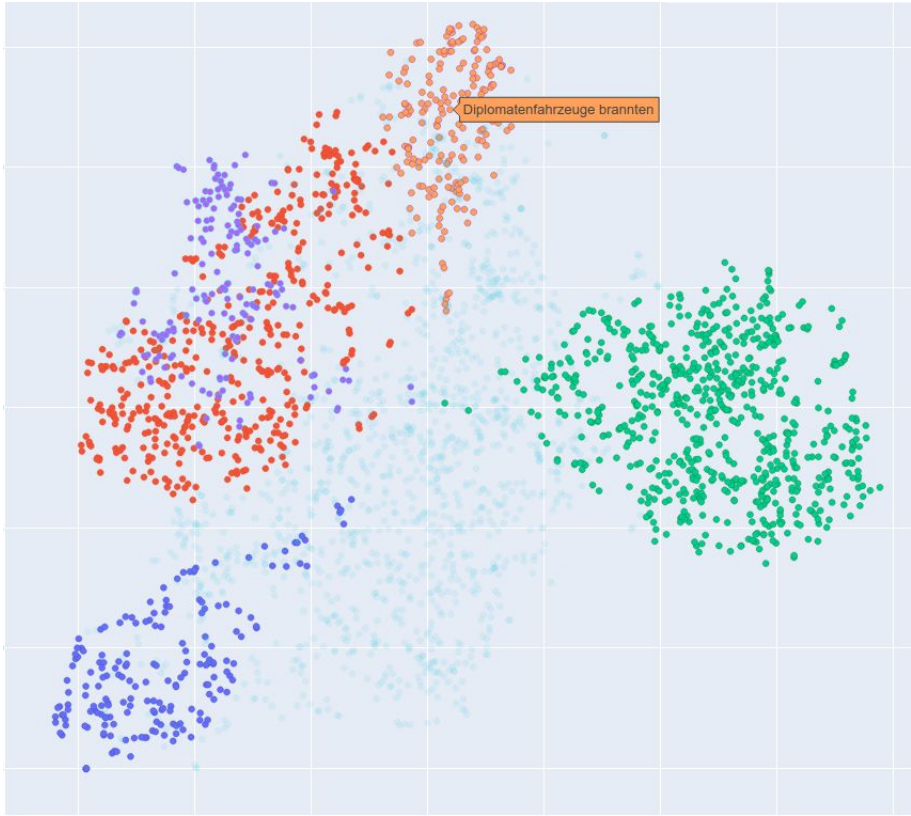
- Tal im Plot: ***reachability distance*** zwischen diesen Punkten ist geringer → Cluster-Bildung
- je geringer die reachability distance, desto dichter das Cluster



Quelle: [OPT2]

OPTICS Clustering

Word2Vec → Doc2Vec → t-SNE → **OPTICS Clustering**



Beispiel für ein OPTICS Clustering auf einem Sample von 3000 Polizeiberichten/ -dokumenten

VIELEN DANK!

Bildquellen

- WED:
  - http://jalammar.github.io/illustrated-word2vec/
- SWS:
  - http://jalammar.github.io/illustrated-word2vec/
- CSM:
  - http://i0.wp.com/techinpink.com/wp-content/uploads/2017/07/cosine.png
- OPT1:
  - https://scikit-learn.org/stable/modules/clustering.html#optics
- OPT2:
  - https://www.researchgate.net/figure/OPTICS-reachability-plot-and-randomly-generated-density-levels_fig2_304480189