Job Post Similarity

Tobia Ippolito, Syon Kadkade

Einführung

TF-IDF-Vektoren

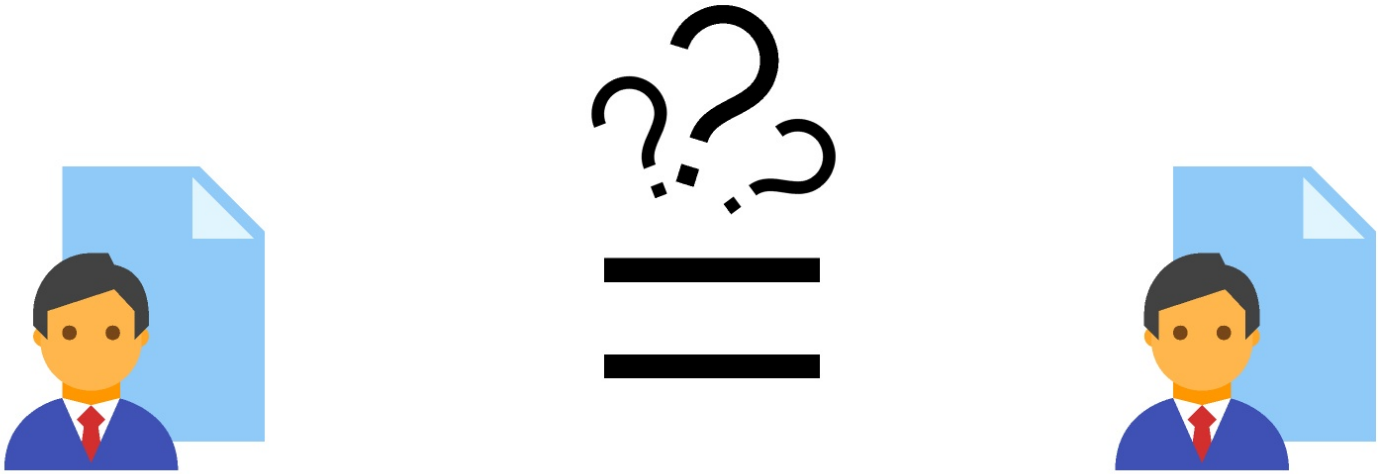
Topic-Modelling

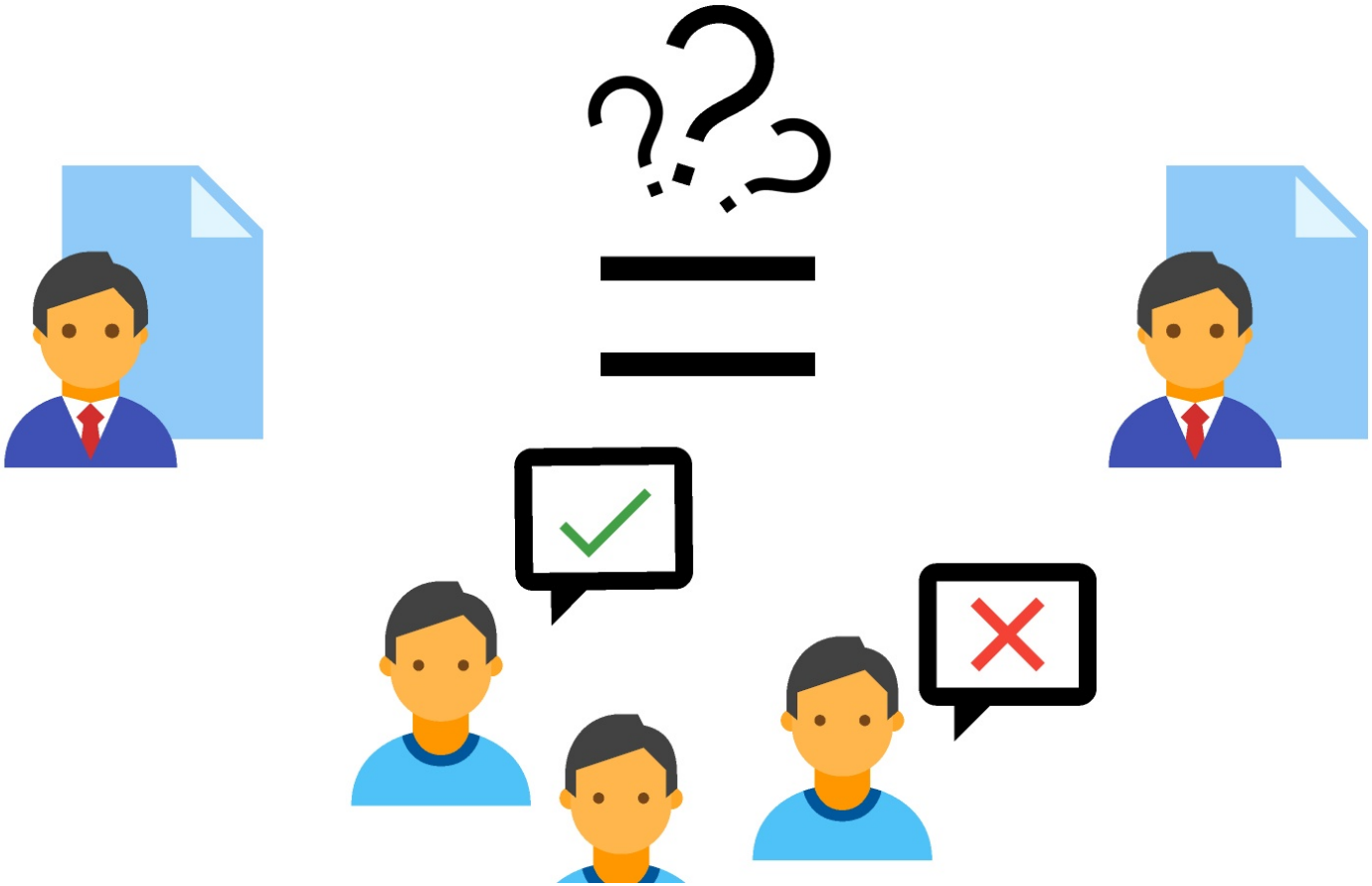
Word2Vec

Location + Type

Performance

Title, Category, Description





>>> Job Post <<<

Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required (masters or PhD preferred) Â Â

>>> Job Post <<<

Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required (masters or PhD preferred) Â Â

>>> Job Post <<<

Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required (masters or PhD preferred) Â Â

```
-----------------------------------------------------------------
                        >>> Job Post <<<
-----------------------------------------------------------------


Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:

 -----------------------------------------------------------------

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using
Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as
Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale
optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required
(masters or PhD preferred) Â  Â
-----------------------------------------------------------------
```

>>> Job Post <<<

Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required (masters or PhD preferred) Â Â

>>> Job Post <<<

Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required (masters or PhD preferred) Â Â

```
-----------------------------------------------------------------
                        >>> Job Post <<<
-----------------------------------------------------------------

Title: Data Scientist

Category: nan

Location: New York in Usa

Type: Full Time

Despription:
 -----------------------------------------------------------------

Candidates should have the following background, skills and characteristics: Â Experience developing analytic applications using Java/Scala and related technologies in the Spark/Hadoop ecosystem Seasoned experience with Big Data programming frameworks, such as Apache Spark and Kafka Expert-level proficiency in Scala and Java is must Experience in machine learning techniques, large scale optimization and building/supporting production ML systems is a big plus Bachelorâ€™s degree in a quantitative discipline required (masters or PhD preferred) Â  Â
-----------------------------------------------------------------
```

Job Post Similarity

Tobia Ippolito, Syon Kadkade

Einführung

TF-IDF-Vektoren

Topic-Modelling

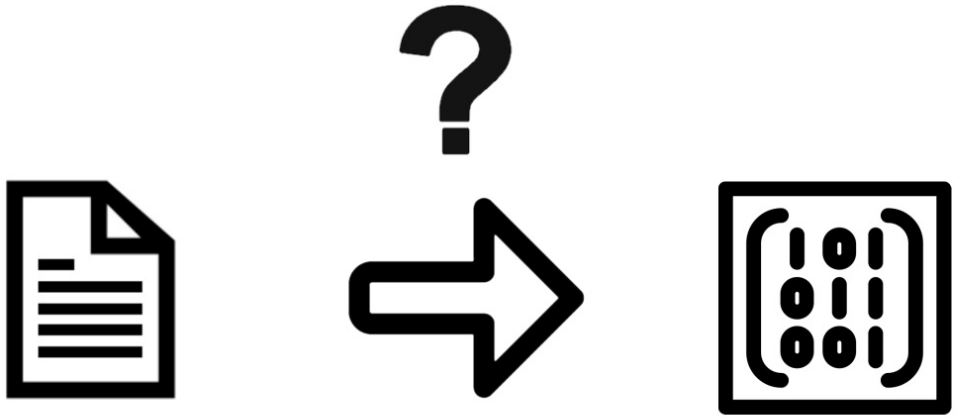
Word2Vec

Location + Type

Performance

Title, Category, Description

Word-Similarities

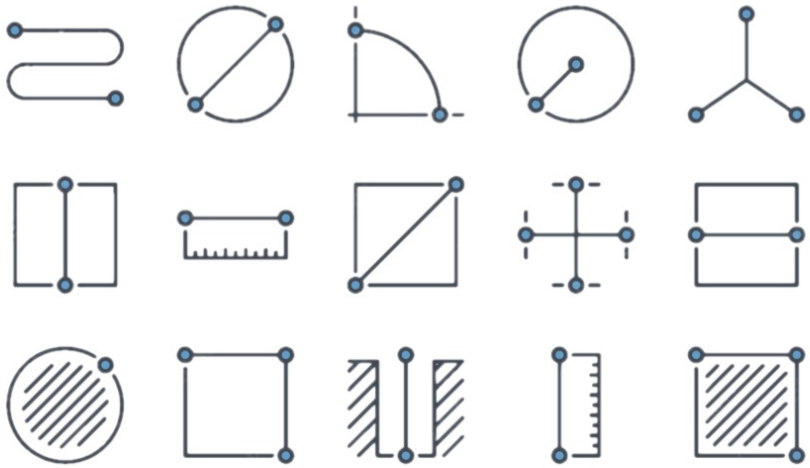


TF-IDF-Vektoren

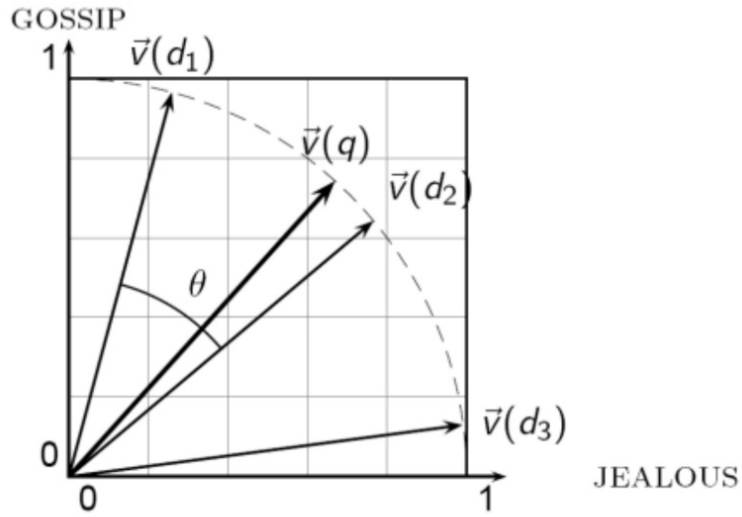
- Assign a tf-idf weight for each term t in each document d:

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

- N: total number of documents
- Increases with the number of occurrences within a document
- Increases with the rarity of the term in the collection



Cosinus Distanz



Implementierung

```python
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(use_idf=True))
X = vectorizer.fit_transform(new_corpus)
X
```

```
<100x39968 sparse matrix of type '<class 'numpy.float64'>'
        with 64457 stored elements in Compressed Sparse Row format>
```

Implementierung

Implementierung

```
[12] from sklearn.metrics.pairwise import cosine_similarity
     matrix = cosine_similarity(X.toarray())
     matrix
```

```
array([[1.        , 0.18206721, 0.15866824, ..., 0.18550827, 0.19767062,
        0.29209581],
       [0.18206721, 1.        , 0.20003613, ..., 0.16128113, 0.17246832,
        0.27128279],
       [0.15866824, 0.20003613, 1.        , ..., 0.14950255, 0.1491627 ,
        0.23637709],
       ...,
       [0.18550827, 0.16128113, 0.14950255, ..., 1.        , 0.186628  ,
        0.26513319],
       [0.19767062, 0.17246832, 0.1491627 , ..., 0.186628  , 1.        ,
        0.28023186],
       [0.29209581, 0.27128279, 0.23637709, ..., 0.26513319, 0.28023186,
        1.        ]])
```

Anwendung am Feature "job_description"

```
def job_description(sentence1, sentence2):

    sentences = [sentence1, sentence2]

    tf_vec = TfidfVectorizer(use_idf=True) # tfidvectorizer

    vectors = tf_vec.fit_transform(sentences).toarray() #transform documents into vectors
    cosine = cosine_similarity(vectors) #use cosine similarity on vectors

    similarity_df = pd.DataFrame(cosine) #save matrix in pandas dataframe to use melt method on it

    transform_df = pd.melt(similarity_df) #reshape matrix into column of dataframe

    similarity = transform_df["value"].values[1] #get value
```

Job Post Similarity

Tobia Ippolito, Syon Kadkade

Einführung

TF-IDF-Vektoren

Topic-Modelling

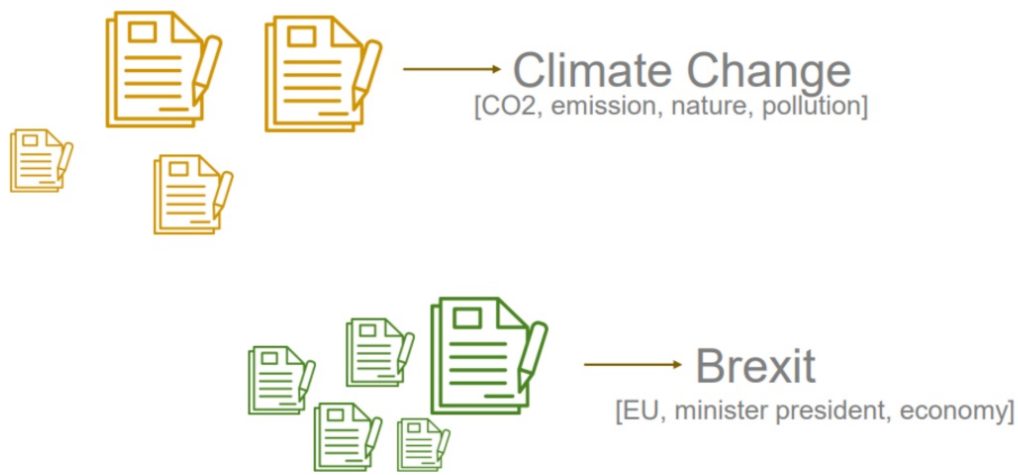
Word2Vec

Location + Type

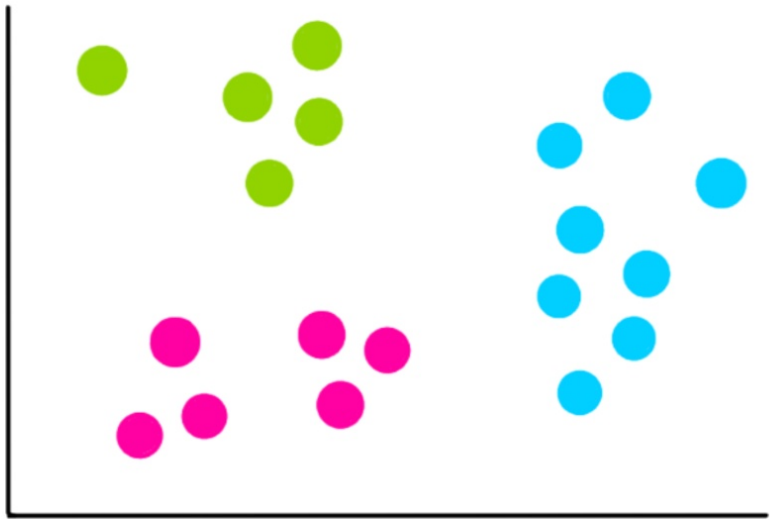
Performance

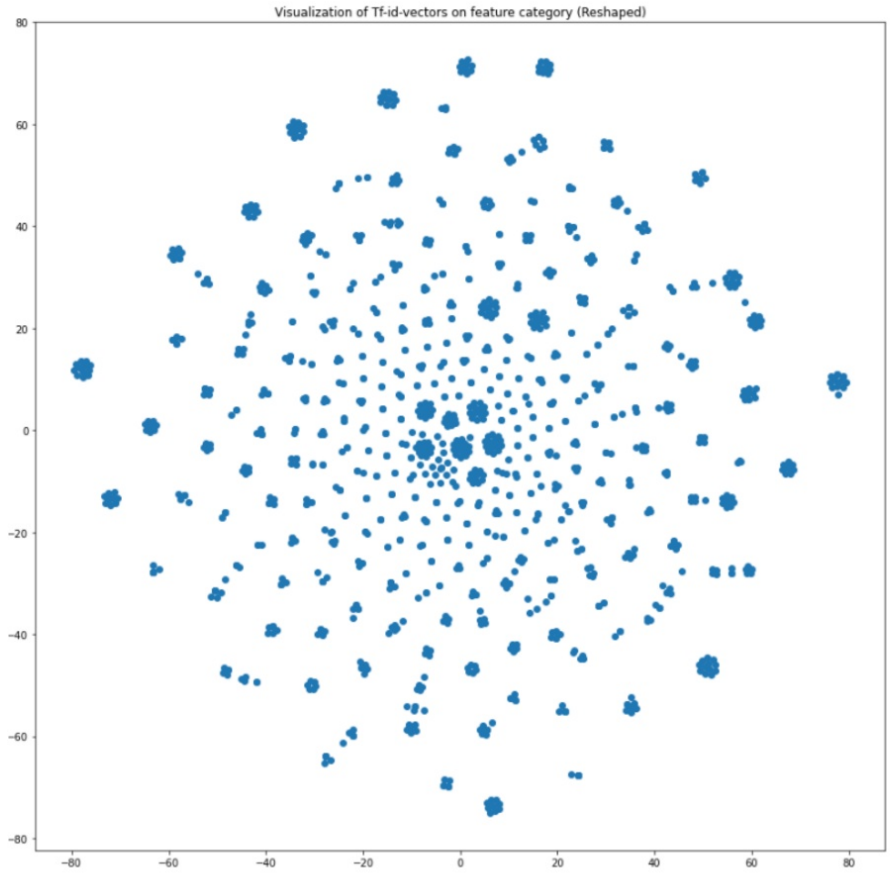
Title, Category, Description

Topic Modeling

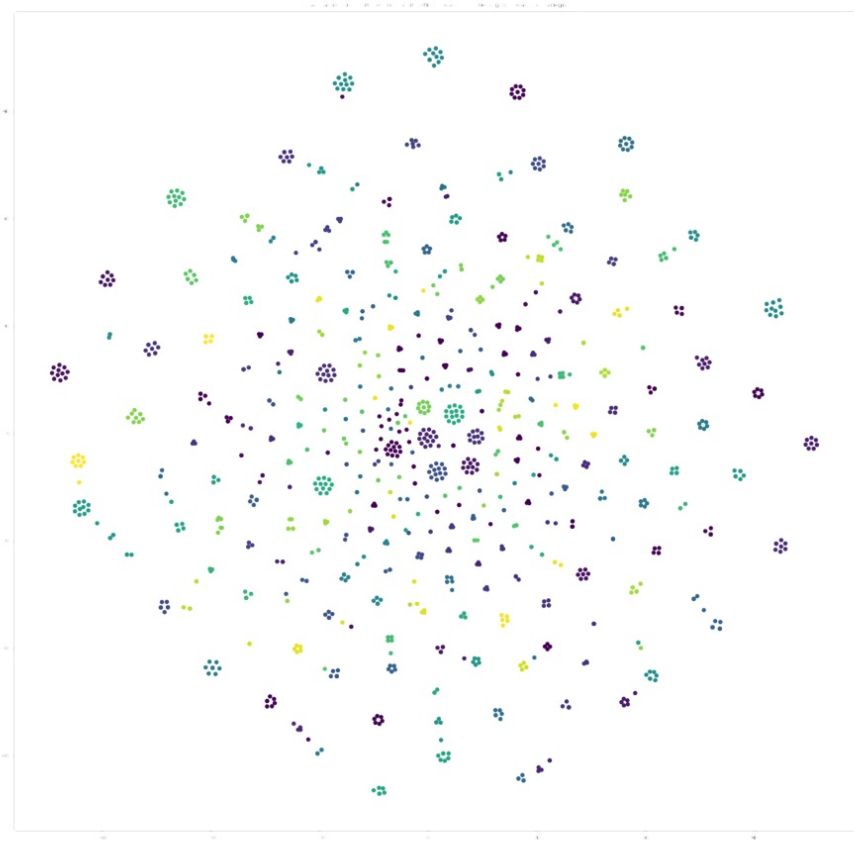


Text Clustering

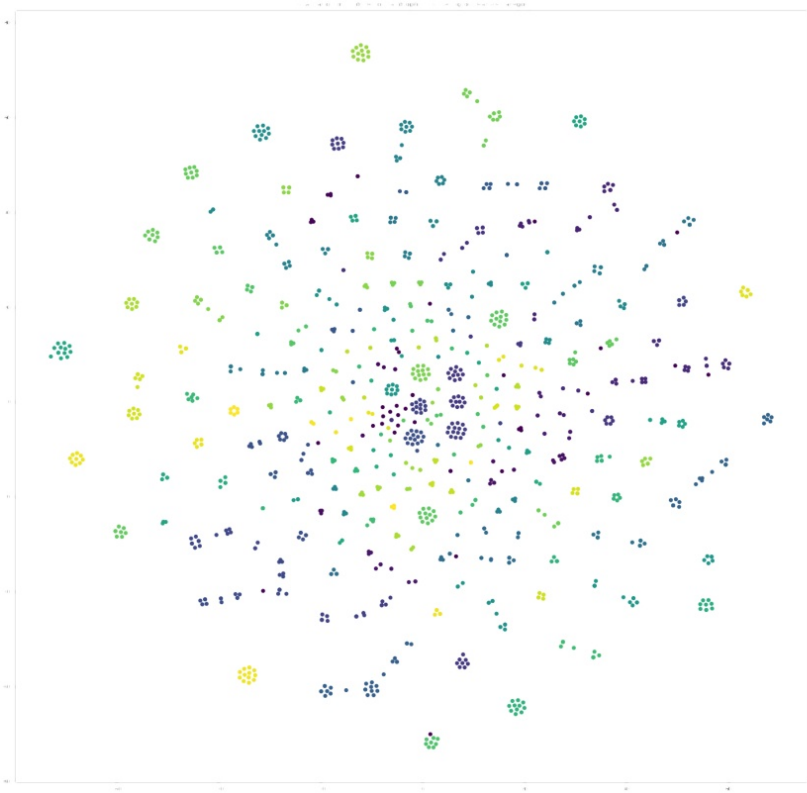



Visualization of Tf-id-vectors on feature category (Reshaped)
80
60
40
20
0
−20
−40
−60
−80
−80
−60
−40
−20
0
20
40
60
80

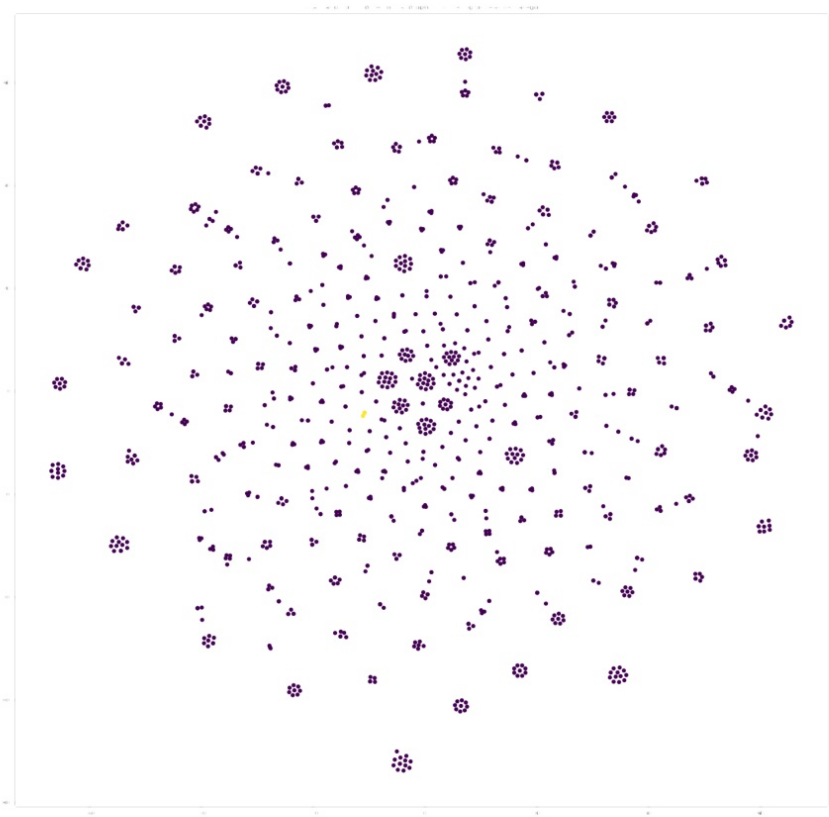
K-means (nltk)



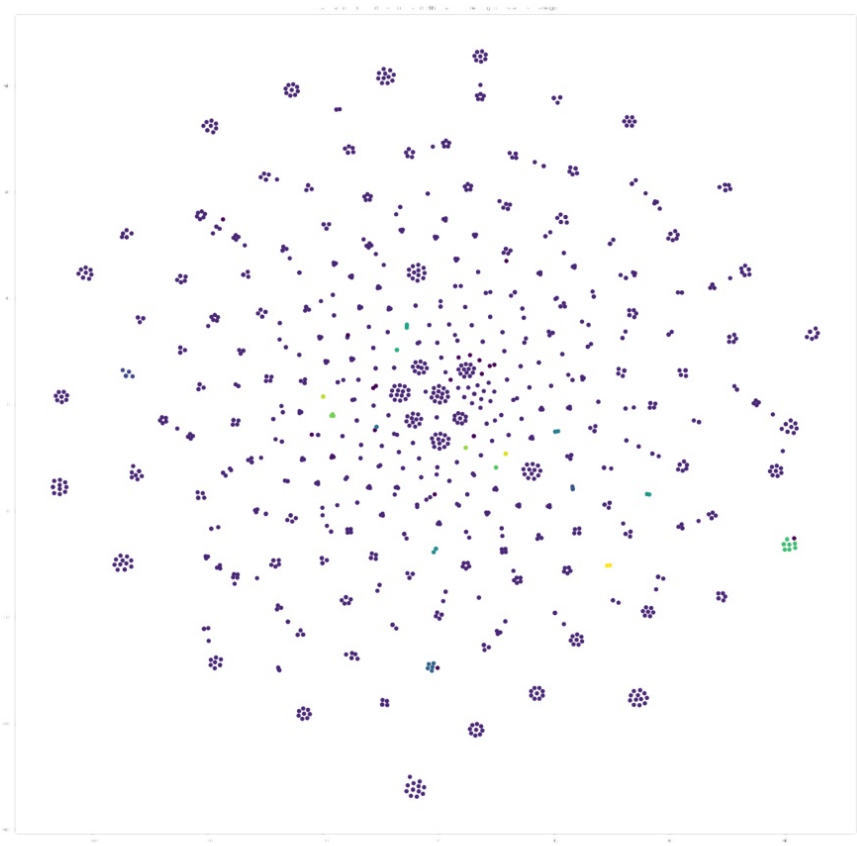
Optics (scikit-learn)



Linkage



DBSCAN

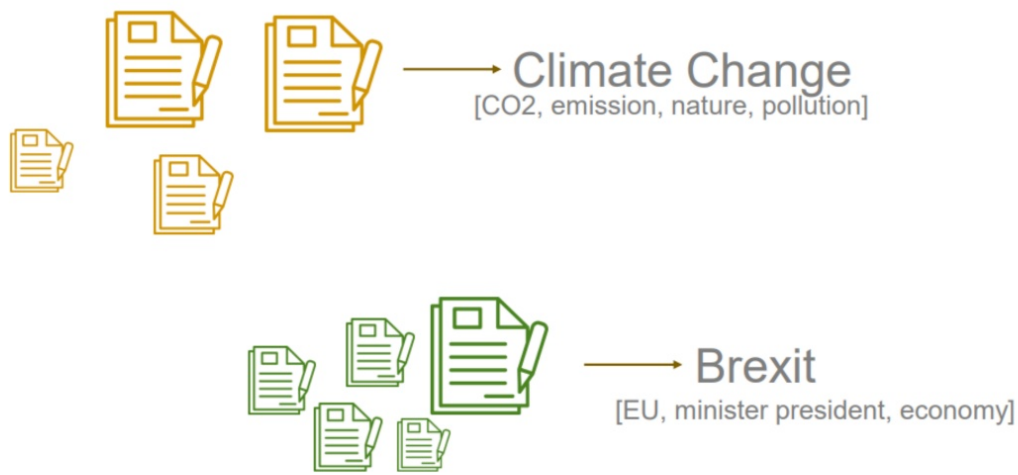


```
ALGORITHM: NLTK KMEANS (COSINE DISTANTS)


Cluster: 281
Words: ['accounting', 'accounting finance', 'finance', 'finance jobs']
Cluster: 0
Words: ['actuarial', 'actuarial data', 'actuarial data scientist', 'ad', 'ad tech', 'administrative support', 'agriculture', 'agriculture fishing', 'alpha', 'alpha insights', 'astrophysics', 'automotive', 'biotech', 'building', 'building
ist', 'government', 'government military', 'healthcare data', 'healthcare data scientist', 'healthcare jobs', 'human', 'human resources', 'insights', 'insights practice', 'legal', 'liquidity', 'maintenance', 'manufacturing data', 'manufa
t marketplace liquidity', 'scientist max', 'scientist oci', 'scientist principal', 'scientist provider', 'scientist provider performance', 'scientist sas', 'scientist sas programmer', 'scientist stellar', 'scientist stellar astrophysics'
Cluster: 236
Words: ['admin', 'admin laboratory', 'admin laboratory research', 'laboratory', 'laboratory research', 'research']
Cluster: 252
Words: ['admin administrative', 'admin administrative positions', 'administrative', 'administrative jobs', 'administrative positions', 'positions']
Cluster: 49
Words: ['admin admissions', 'admin admissions enrollment', 'admissions', 'admissions enrollment', 'enrollment']
Cluster: 142
Words: ['admin health', 'admin health medical', 'health medical', 'health medical services', 'medical', 'medical services']
Cluster: 206
Words: ['admin information', 'admin information technology', 'information', 'information technology', 'technology']
Cluster: 76
Words: ['admin institutional', 'admin institutional research', 'institutional', 'institutional research', 'institutional research planning', 'research planning']
Cluster: 227
Words: ['admin instructional', 'admin instructional technology', 'instructional', 'instructional technology', 'instructional technology design', 'planning admin', 'planning admin instructional', 'research planning admin', 'technology des
Cluster: 101
Words: ['administrator', 'database', 'database administrator', 'information technology database', 'technology database', 'technology database administrator']
Cluster: 160
Words: ['advanced', 'data scientist advanced', 'scientist advanced']
Cluster: 226
Words: ['advanced analytics', 'scientist advanced analytics']
Cluster: 284
Words: ['advanced analytics data', 'analytics data', 'analytics data scientist', 'brand', 'brand analytics', 'brand analytics data']
Cluster: 158
Words: ['advanced analytics hia', 'analytics hia', 'hia', 'manager advanced', 'manager advanced analytics', 'scientist manager advanced']
Cluster: 230
Words: ['advanced marketing', 'advanced marketing analytics', 'marketing analytics', 'scientist advanced marketing']
Cluster: 89
Words: ['advanced video', 'advanced video advertising', 'scientist advanced video', 'video', 'video advertising']
Cluster: 55
Words: ['advertising', 'advertising pr', 'data scientist advertising', 'marketing', 'marketing advertising', 'marketing advertising pr', 'pr', 'scientist advertising']
Cluster: 42
Words: ['aerial', 'aerial imagery', 'banking', 'banking loans', 'data', 'data scientist', 'data scientist aerial', 'data scientist graph', 'data scientist growth', 'data scientist healthcare', 'data scientist home', 'data scientist netwo
Cluster: 7
Words: ['agile']
Cluster: 123
Words: ['agile hub', 'data scientist agile', 'hub', 'scientist agile', 'scientist agile hub']
Cluster: 31
Words: ['agile software', 'agile software developer', 'software', 'software developer']
Cluster: 97
Words: ['ai', 'ai ml', 'data scientist ai', 'ml', 'scientist ai', 'scientist ai ml']
Cluster: 166
Words: ['ai data', 'ai data scientist', 'learning ts', 'learning ts sci', 'machine learning ts', 'sci', 'ts', 'ts sci']
Cluster: 13
Words: ['ai implementation', 'ai implementation leader', 'apaas', 'apaas lead', 'apaas lead data', 'implementation', 'implementation leader', 'implementation leader apaas', 'leader', 'leader apaas', 'leader apaas lead']
```

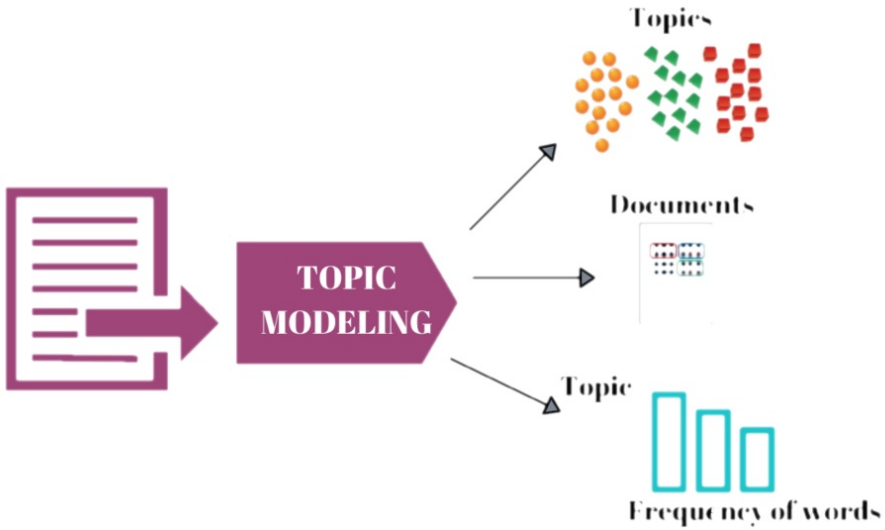
```
ALGORITHM: NLTK KMEANS (COSINE DISTANTS)


Cluster: 281
Words: ['accounting', 'accounting finance', 'finance', 'finance jobs']
Cluster: 0
Words: ['actuarial', 'actuarial data', 'actuarial data scientist', 'ad', 'ad tech', 'administrative support', 'agriculture', 'agriculture fishing', 'alpha', 'alpha insights', 'astrophysics', 'automotive', 'biotech', 'building', 'building
ist', 'government', 'government military', 'healthcare data', 'healthcare data scientist', 'healthcare jobs', 'human', 'human resources', 'insights', 'insights practice', 'legal', 'liquidity', 'maintenance', 'manufacturing data', 'manufa
t marketplace liquidity', 'scientist max', 'scientist oci', 'scientist principal', 'scientist provider', 'scientist provider performance', 'scientist sas', 'scientist sas programmer', 'scientist stellar', 'scientist stellar astrophysics'
Cluster: 236
Words: ['admin', 'admin laboratory', 'admin laboratory research', 'laboratory', 'laboratory research', 'research']
Cluster: 252
Words: ['admin administrative', 'admin administrative positions', 'administrative', 'administrative jobs', 'administrative positions', 'positions']
Cluster: 49
Words: ['admin admissions', 'admin admissions enrollment', 'admissions', 'admissions enrollment', 'enrollment']
Cluster: 142
Words: ['admin health', 'admin health medical', 'health medical', 'health medical services', 'medical', 'medical services']
Cluster: 206
Words: ['admin information', 'admin information technology', 'information', 'information technology', 'technology']
Cluster: 76
Words: ['admin institutional', 'admin institutional research', 'institutional', 'institutional research', 'institutional research planning', 'research planning']
Cluster: 227
Words: ['admin instructional', 'admin instructional technology', 'instructional', 'instructional technology', 'instructional technology design', 'planning admin', 'planning admin instructional', 'research planning admin', 'technology des
Cluster: 101
Words: ['administrator', 'database', 'database administrator', 'information technology database', 'technology database', 'technology database administrator']
Cluster: 160
Words: ['advanced', 'data scientist advanced', 'scientist advanced']
Cluster: 226
Words: ['advanced analytics', 'scientist advanced analytics']
Cluster: 284
Words: ['advanced analytics data', 'analytics data', 'analytics data scientist', 'brand', 'brand analytics', 'brand analytics data']
Cluster: 158
Words: ['advanced analytics hia', 'analytics hia', 'hia', 'manager advanced', 'manager advanced analytics', 'scientist manager advanced']
Cluster: 230
Words: ['advanced marketing', 'advanced marketing analytics', 'marketing analytics', 'scientist advanced marketing']
Cluster: 89
Words: ['advanced video', 'advanced video advertising', 'scientist advanced video', 'video', 'video advertising']
Cluster: 55
Words: ['advertising', 'advertising pr', 'data scientist advertising', 'marketing', 'marketing advertising', 'marketing advertising pr', 'pr', 'scientist advertising']
Cluster: 42
Words: ['aerial', 'aerial imagery', 'banking', 'banking loans', 'data', 'data scientist', 'data scientist aerial', 'data scientist graph', 'data scientist growth', 'data scientist healthcare', 'data scientist home', 'data scientist netwo
Cluster: 7
Words: ['agile']
Cluster: 123
Words: ['agile hub', 'data scientist agile', 'hub', 'scientist agile', 'scientist agile hub']
Cluster: 31
Words: ['agile software', 'agile software developer', 'software', 'software developer']
Cluster: 97
Words: ['ai', 'ai ml', 'data scientist ai', 'ml', 'scientist ai', 'scientist ai ml']
Cluster: 166
Words: ['ai data', 'ai data scientist', 'learning ts', 'learning ts sci', 'machine learning ts', 'sci', 'ts', 'ts sci']
Cluster: 13
Words: ['ai implementation', 'ai implementation leader', 'apaas', 'apaas lead', 'apaas lead data', 'implementation', 'implementation leader', 'implementation leader apaas', 'leader', 'leader apaas', 'leader apaas lead']
```

```
EXPERIMENT WITH ALL SAMPLES (actual 9118) DIFFERENCE BETWEEN MENTIONED ALGORYTHM ==========================================================================================
ALGORITHM: OPTICS

FOUND CLUSTER = 289

Cluster: 0
Words: ['accounting', 'accounting finance', 'finance']
Cluster: 255
Words: ['actuarial', 'actuarial data', 'actuarial data scientist']
Cluster: 108
Words: ['ad', 'ad tech', 'data scientist ad', 'scientist ad', 'scientist ad tech']
Cluster: 151
Words: ['admin', 'admin laboratory', 'admin laboratory research', 'laboratory', 'laboratory research', 'research']
Cluster: 178
Words: ['admin administrative', 'admin administrative positions', 'administrative positions', 'positions']
Cluster: 184
Words: ['admin admissions', 'admin admissions enrollment', 'admissions', 'admissions enrollment', 'enrollment']
Cluster: 195
Words: ['admin health', 'admin health medical', 'health medical', 'health medical services', 'medical', 'medical services']
Cluster: 33
Words: ['admin information', 'admin information technology', 'information', 'information technology']
Cluster: 152
Words: ['admin institutional', 'admin institutional research', 'institutional', 'institutional research', 'institutional research planning', 'planning', 'research planning']
Cluster: 154
Words: ['admin instructional', 'admin instructional technology', 'instructional', 'instructional technology', 'instructional technology design', 'planning admin', 'planning admin instructional', 'research planning admin', 'technology des
Cluster: 208
Words: ['administrative', 'administrative jobs']
Cluster: 269
Words: ['administrative support', 'office', 'office administrative', 'office administrative support', 'support']
Cluster: 34
Words: ['administrator', 'database', 'database administrator', 'information technology database', 'technology database', 'technology database administrator']
Cluster: 73
Words: ['advanced', 'advanced analytics', 'data scientist advanced', 'scientist advanced']
Cluster: -1
Words: ['advanced analytics data', 'analyst', 'automotive', 'biotech', 'education jobs', 'engineering jobs', 'experience', 'finance jobs', 'financieel', 'insurance', 'iv', 'learning', 'legal', 'level', 'management', 'management jobs', 'm
Cluster: 72
Words: ['advanced analytics hia', 'analytics hia', 'hia', 'manager advanced', 'manager advanced analytics', 'scientist manager advanced']
Cluster: 74
Words: ['advanced marketing', 'advanced marketing analytics', 'marketing analytics', 'scientist advanced marketing']
Cluster: 77
Words: ['advanced video', 'advanced video advertising', 'scientist advanced video', 'video', 'video advertising']
Cluster: 180
Words: ['advertising', 'advertising pr', 'marketing', 'marketing advertising', 'marketing advertising pr', 'pr']
Cluster: 238
Words: ['aerial', 'aerial imagery', 'data scientist aerial', 'imagery', 'scientist aerial', 'scientist aerial imagery']
Cluster: 90
Words: ['agile', 'agile software', 'agile software developer', 'software', 'software developer']
Cluster: 91
Words: ['agile hub', 'data scientist agile', 'hub', 'scientist agile', 'scientist agile hub']
Cluster: 283
Words: ['agriculture', 'agriculture fishing', 'fishing']
Cluster: 81
Words: ['ai', 'ai data', 'ai data scientist']
```

Topic Modeling



Latent Dirichlet Allocation



Spezielle Bibliotheken

GENSIM
topic modelling for humans

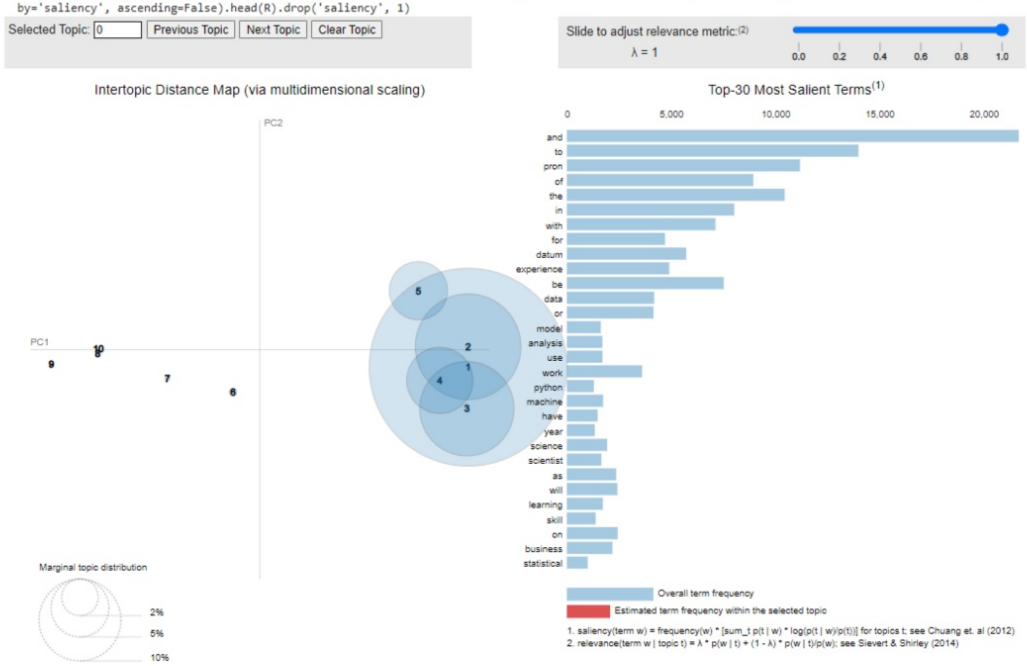
pyLDAvis

LDA Model

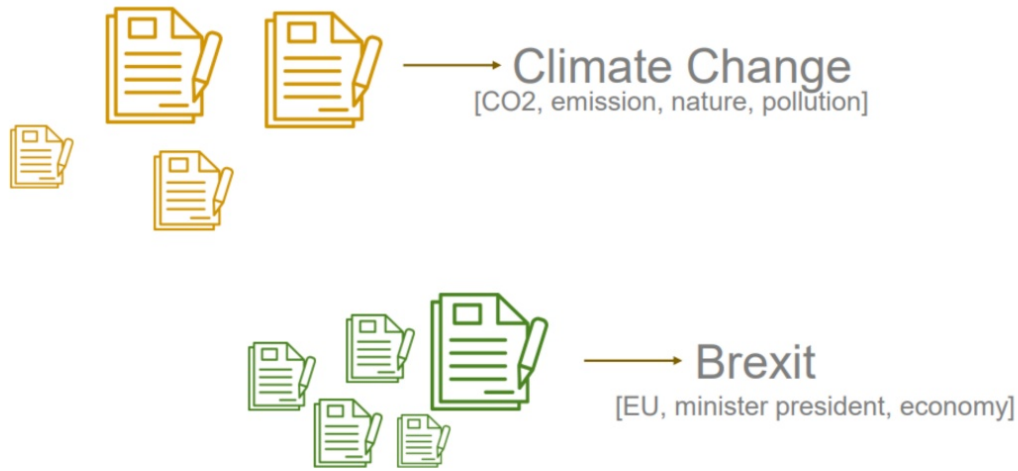
```
model = LdaModel(
    corpus=new_corpus3,
    id2word=dictionary,
    alpha='auto',
    eta='auto',
    iterations=400,
    num_topics=10,
    eval_every=None)
```

PyLDAvis

```
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  by='saliency', ascending=False).head(R).drop('saliency', 1)
```



Topic Modeling



Job Post Similarity

Tobia Ippolito, Syon Kadkade

Einführung

TF-IDF-Vektoren

Topic-Modelling

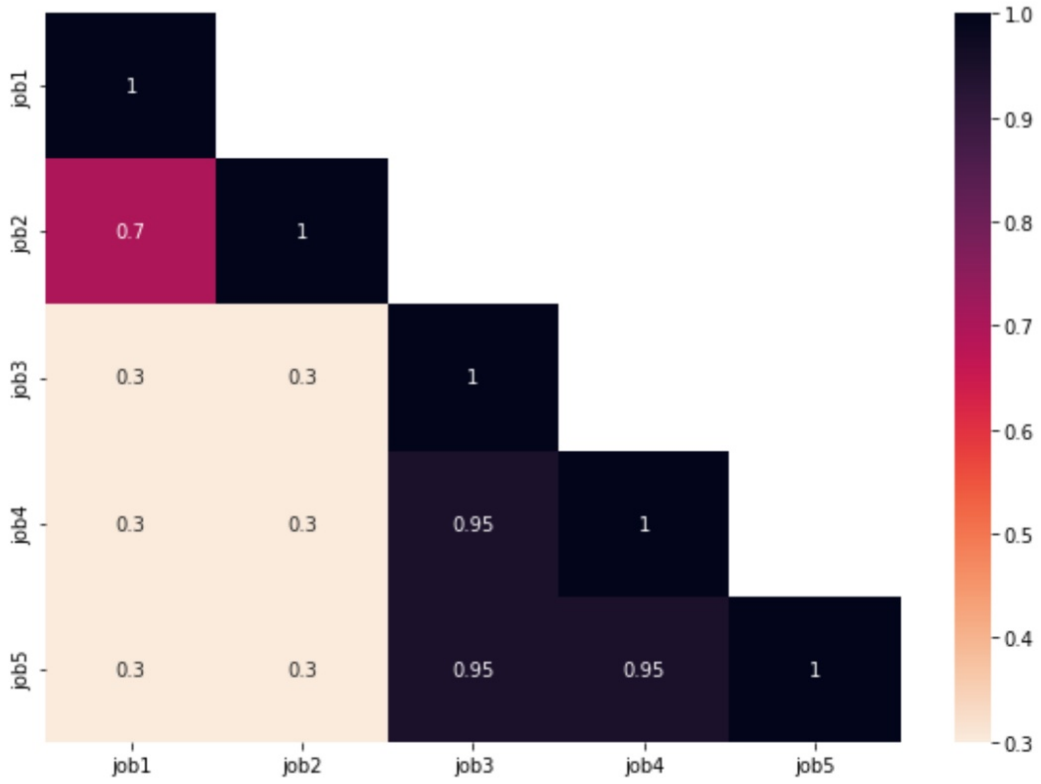
Word2Vec

Location + Type

Performance

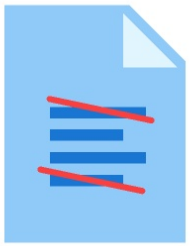
Title, Category, Description

Evaluierungsmaß



Experiment 1:
Job Description mit nur wichtigen Wörtern

Experiment 1:
Job Description mit nur wichtigen Wörtern

1. 

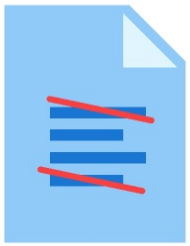
Nomen

Adjektive

Verben

Experiment 1:
Job Description mit nur wichtigen Wörtern

1.



Nomen

Adjektive

Verben

2.

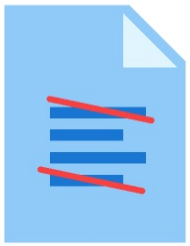
```
doc1.similarity(doc2)
```

✓ 0.1s

```
0.9615480470385989
```

Experiment 1:
Job Description mit nur wichtigen Wörtern

1.



Nomen

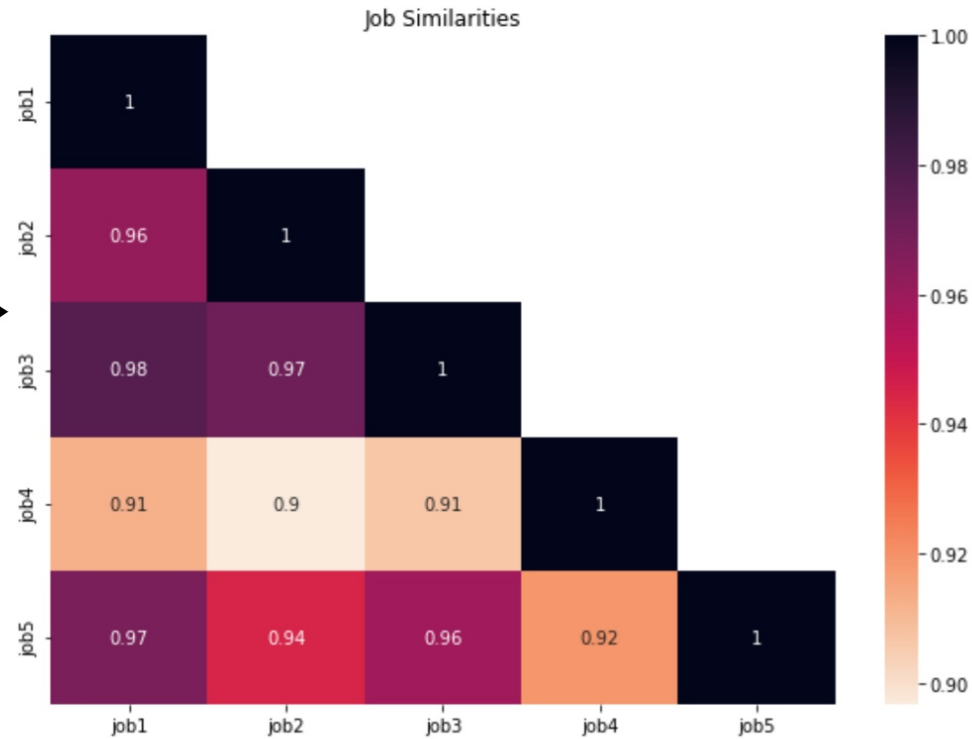
Adjektive

Verben

2.

```
doc1.similarity(doc2)
✓ 0.1s
```

```
0.9615480470385989
```



Experiment 2:
Job-Titel mit Word2Vec

Experiment 2:
Job-Titel mit Word2Vec

```
job1_title = data['job_title'][0]
job2_title = data['job_title'][3]
```

Experiment 2:
Job-Titel mit Word2Vec

```
job1_title = data['job_title'][0]
job2_title = data['job_title'][3]
```

```
title1 = nlp(job1_title)
title2 = nlp(job2_title)
title1.similarity(title2)
```

✓ 0.1s

```
0.7552776180724826
```

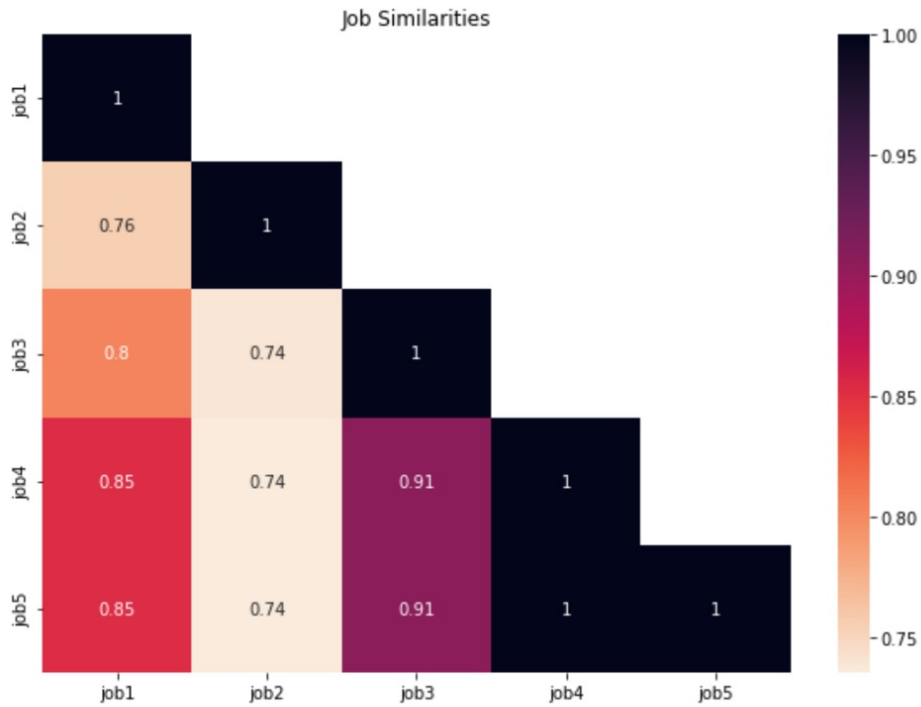
Experiment 2:
Job-Titel mit Word2Vec

```
job1_title = data['job_title'][0]
job2_title = data['job_title'][3]
```

```
title1 = nlp(job1_title)
title2 = nlp(job2_title)
title1.similarity(title2)
```

✓ 0.1s

```
0.7552776180724826
```



Experiment 3:
Job-Kategorien mit Word2Vec

Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```

Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

```
categories_
```

✓ 0.1s

```
['Accounting/Finance', 'Accounting/Finance', 'biotech', nan, nan]
```

```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```

Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

```
categories_
```

✓ 0.1s

```
['Accounting/Finance', 'Accounting/Finance', 'biotech', nan, nan]
```

```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```

```
data.shape[0] / data['category'].isna().sum()
```

✓ 0.1s

```
11.337868480725623
```

Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```

Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

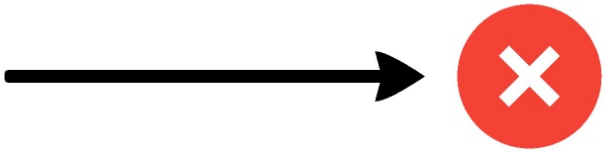
```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```

Experiment 3:
Job-Kategorien mit Word2Vec

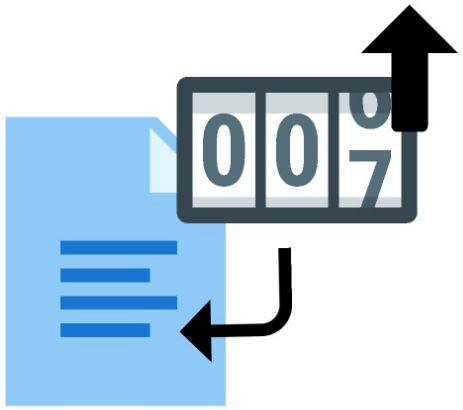
```
job1_category = data['category'][0]
job2_category = data['category'][3]
```



```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```



Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```

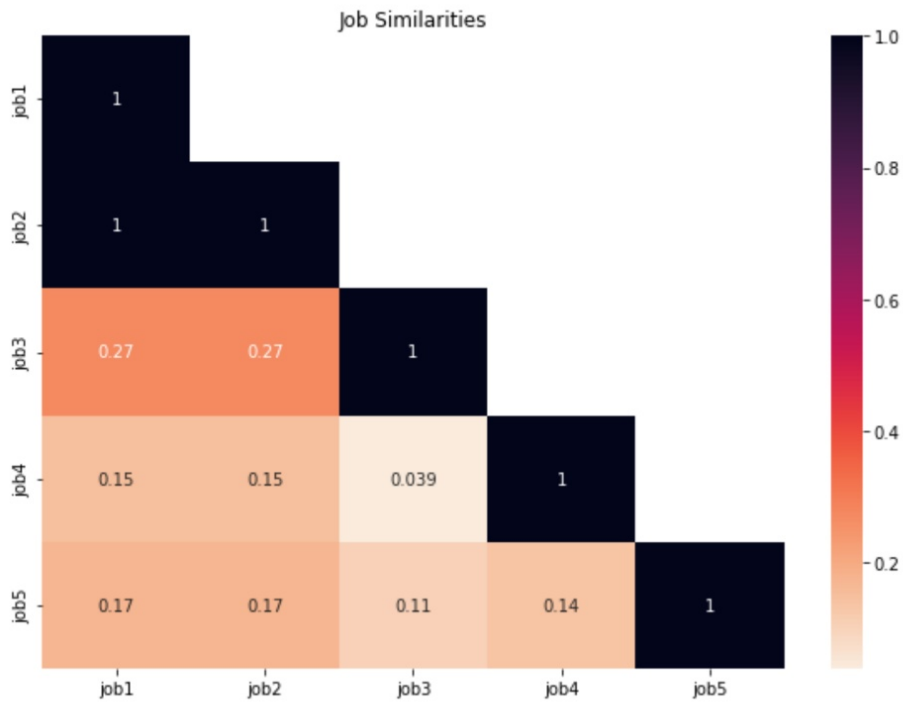
Experiment 3:
Job-Kategorien mit Word2Vec

```
job1_category = data['category'][0]
job2_category = data['category'][3]
```

```
cat1 = nlp(job1_category)
cat2 = nlp(job2_category)
cat1.similarity(cat2)
```

✓ 0.1s

```
1.0
```



Job Post Similarity

Tobia Ippolito, Syon Kadkade

Einführung

TF-IDF-Vektoren

Topic-Modelling

Word2Vec

Location + Type

Performance

Title, Category, Description