

# Location extraction and visualization

Presentation by Daniel T. Soukup (Faculty of Math, Uni Vienna)

Joint work with Regina Babo and Liad Magen.

# data4good hackathon



This work was mostly done at the data4good Hackathon (April 27-28, 2019).

- 2 days, 4 teams working paired with with 4 NGOs
- Our team lead: Liad Magen.



# Textual analysis of urban greening projects in the press

To analyse the urban greening literature and **determine where most of these projects happen.**

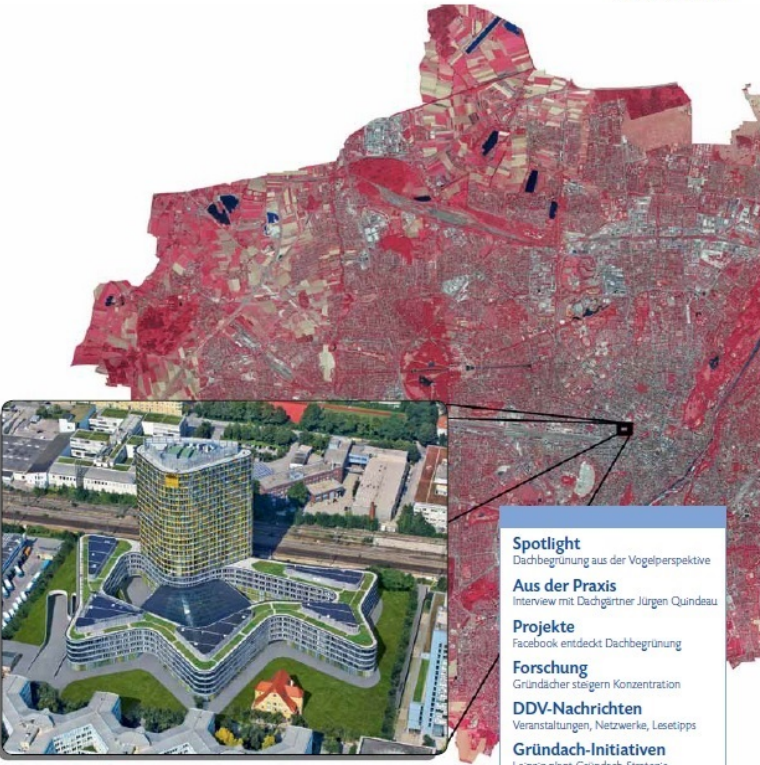


# Location extraction

The raw documents looked like this: are these locations relevant?

**GRÜNDACHAKTUELL**  
Der Deutsche Dachgärtner Verband informiert

Ausgabe 1 | 2016



**Spotlight**  
Dachbegrünung aus der Vogelperspektive

**Aus der Praxis**  
Interview mit Dachgärtner Jürgen Quindeau

**Projekte**  
Facebook entdeckt Dachbegrünung

**Forschung**  
Gründächer steigern Konzentration


**DDV-Nachrichten**  
Veranstaltungen, Netzwerke, Lesetipps

**Gründach-Initiativen**  
Leipzig plant Gründach-Strategie

Editorial

**GRÜNDACHAKTUELL**

**1'2016**



**03 Spotlight**  
Dachbegrünung aus der Vogelperspektive – „Rot ist das neue Grün“

**08 Aus der Praxis**  
Dachgärtner aus Leidenschaft – Interview mit Jürgen Quindeau, Fa. GRÜN + DACH

**10 Projekte**  
Leben und arbeiten auf dem Dach – Die neue Facebook-Zentrale in Menlo Park

**12 Forschung**  
Gründächer steigern Produktivität – Studie der Universität Melbourne

**14 DDV-Nachrichten**  
Veranstaltungen und Netzwerke

**22 Literaturtipps**

**23 Gründach-Initiativen**  
Leipzig plant Gründach-Strategie

**23 Vorschau/Impressum**

Titelbild: © Stadt München/DLR/DDV

Die Dachbegrünung kann in Deutschland auf eine jahrzehntelange Tradition zurückblicken. Trotzdem gibt es immer noch einige ungelöste Fragen. Wie viele Gründächer existieren bereits in Deutschlands Städten? Und welche Dächer bieten Potenzialflächen für den Ausbau der grünen Infrastruktur? Gemeinsam mit dem Deutschen Zentrum für Luft- und Raumfahrt (DLR) hat der Deutsche Dachgärtner Verband (DDV) ein Verfahren entwickelt, um diese Fragen in Zukunft schnell und effizient beantworten zu können. In der Rubrik „Spotlight“ präsentieren wir Ihnen eine Kurzbeschreibung der Methode.

Apropos Potenziale. Beispiele für die Nutzung von Firmendächern als „grüne Freiluftbüros“ und Erholungsbereiche sind hierzulande nur sehr selten zu finden. Wie sich ein Dachgarten für Mitarbeiter und Unternehmen doppelt bezahlt macht, zeigt die neue Facebook-Firmenzentrale in Kalifornien. Dass der Platzhirsch unter den Social Media Plattformen mit diesem Gebäude zusätzlich sein Umweltprofil in der Außendarstellung schärft, kommt als Bonus noch hinzu. Um Inspiration für den deutschen Gründach-Markt zu gewinnen, lohnt es sich, den Blick auch über die Ländergrenzen hinweg zu werfen.

Viel Vergnügen beim Lesen wünscht Ihnen

Wolfgang Ansel  
Geschäftsführer, Deutscher Dachgärtner Verband e. V.



# How about these locations?

Gründach-Initiativen

## Leipzig erarbeitet Gründach-Strategie als Anpassungsmaßnahme an den Klimawandel

Viele Kommunen folgen aktuell dem Beispiel der Stadt Hamburg und entdecken die Dachbegrünung als wichtiges Instrument einer nachhaltigen Städteplanung. In Leipzig beschäftigt sich das Amt für Umweltschutz mit der Entwicklung einer kommunalen Gründach-Strategie.

Die Umsetzung begrünter Dächer kann auf städtischer Ebene durch verschiedene Maßnahmen gefördert werden. Neben der verstärkten Festsetzung in Bebauungsplänen werden in Leipzig auch die Einbindung in das wichtige Handlungsfeld Regenwassermanagement und die Begrünung von öffentlichen Gebäuden diskutiert. Um bei der Erstellung der Gründach-Strategie von den Erfahrungen anderer Kommunen zu profitieren, hat das Amt für Umweltschutz im März DDV-Geschäftsführer Wolfgang Ansel zu einem interdisziplinären Workshop mit Vertretern unterschiedlicher Fachbehörden, Forschungseinrichtungen, städtischer Unternehmen und Verbände eingeladen.



Auch die Leipziger Dachlandschaft bietet Potenziale für Dachbegrünungen.  
© Henry Pleifer [www.profilutbild.de](http://www.profilutbild.de)

Zu den Zielen der Gründach-Strategie gehört es auch, Architekten, Planer, Handwerksbetriebe und Bauherren durch Maßnahmen der Öffentlichkeitsarbeit über die Vorteile der Dachbegrünung zu informieren. Den Startschuss hierzu liefert das Gründach-Forum Leipzig, das die Stadt Leipzig mit dem Deutschen Dachgärtner Verband am 20. Oktober 2016 im Neuen Rathaus veranstalten wird.

### Themenvorschau GründachAktuell 2/2016

**Natur erleben:** Das Biodiversitäts-Gründach des Besucherzentrums der Internationalen Gartenschau 2017 in Berlin

**The Green Skyscraper:** Interview mit dem Pionier bioklimatischer Hochhäuser Dr. Ken Yeang (HAMZAH & YEANG, Malaysia)

**GründachAktuell kostenfrei abonnieren:**  
[www.dachgaertnerverband.de](http://www.dachgaertnerverband.de)

### Impressum GründachAktuell

**Herausgeber:** GründachAktuell ist die Verbandszeitschrift des Deutschen Dachgärtner Verbandes e.V. (DDV)

**Redaktion:** Wolfgang Ansel

**Fotos:** Sofern nicht anders angegeben, liegen die Bildrechte beim Herausgeber.

**Verlag:** DDV-Verlag Nürtingen

**Copyright:** Nachdruck – auch auszugsweise – nur mit Genehmigung des Herausgebers gestattet.

Deutscher Dachgärtner Verband e.V.  
Postfach 20 25  
72610 Nürtingen  
Tel. 07022 301378  
E-Mail: [contact@dachgaertnerverband.de](mailto:contact@dachgaertnerverband.de)  
[www.dachgaertnerverband.de](http://www.dachgaertnerverband.de)



Folgende Publikationen sind bei der DDV-Geschäftsstelle erhältlich:



„Das 1x1 der Dachbegrünung“



Leitfaden Dachbegrünung  
für Kommunen



Leitfaden Sicherer  
Gewerkeübergang



Fachbuch  
Moderne Dachgärten

### Unsere Aktivitäten im Überblick

**Fachberatung:** Für Behörden, Bauherren und Architekten liefern wir Informationsmaterial und Beratung zu allen wichtigen Fragen rund um das Thema Dachbegrünung. Auf unserer Internetseite finden Sie außerdem erfahrene Dachbegrünungsbetriebe aus dem gesamten Bundesgebiet.

**Netzwerk Kommune:** Der DDV fördert den Informationsaustausch zwischen kommunalen Fachbehörden. Durch „Best-Practice“-Beispiele wird die Entwicklung einer kommunalen Gründach-Strategie erleichtert.

**Seminare und Fachvorträge:** Mit der Informationsreihe „Gründach-Forum“ informieren wir Architekten, Baubeteiligte und Fachbehörden. Das DDV-Referenten-Team ist außerdem mit Fachvorträgen bei Umwelt-messen, Bürgerinfo-Abenden und kommunalen Indoor-Seminaren präsent.

**Richtliniendarbeit:** Wir beteiligen uns an der Erstellung neuer Richtlinien und Regelwerke, z. B. an den neuen FLL-Dachbegrünungsrichtlinien.

Hinweise zu unseren aktuellen Aktivitäten finden Sie auch im Internet unter:  
[www.dachgaertnerverband.de](http://www.dachgaertnerverband.de)



## **Plan:**

- Extract all locations from the text.
- Visualize and study the distribution (countries, cities).
- Try to classify the locations (relevant/irrelevant).

**Let's see some code!**

# Loading the modules

Let's start by loading all the necessary modules.



```
In [1]: # generally useful packages
import re, collections
import glob, os, requests, io, pickle

# data stuff
import pandas as pd
import numpy as np

# visualizing
import matplotlib.pyplot as plt
import seaborn as sns
import folium # visualizing on maps

# NLP packages, we work with German text mostly so we load that
import nltk
import spacy
import gensim
from sklearn.feature_extraction.text import CountVectorizer
from geotext import GeoText # for location detection
```

D:\anaconda\envs\data4good\lib\site-packages\smart\_open\ssh.py:34: UserWarning: paramiko missing, opening SSH/SCP/SFTP paths will be disabled. `pip install paramiko` to suppress

warnings.warn('paramiko missing, opening SSH/SCP/SFTP paths will be disabled. `pip install paramiko` to suppress')

D:\anaconda\envs\data4good\lib\site-packages\gensim\utils.py:1197: UserWarning: detected Windows; aliasing chunkize to chunkize\_serial

warnings.warn("detected Windows; aliasing chunkize to chunkize\_serial")

**Skipping the pdf2txt adventures...**

**...imagine we have 25 documents loaded.**

**Finding locations in the text (NER with Spacy)**

We use the Spacy's nlp pipeline to process the corpus. This includes tokenizing, tagging, parsing, **identifying and labeling named entities**.

```
In [3]: nlp = spacy.load("de_core_news_sm", disable=["tagger"])
nlp_corpus = []
for idx, raw in enumerate(corpus):
    joined = ' '.join(raw) # we join the docs (which are lists of strings)
    joined_nlp = nlp(joined)
    nlp_corpus.append(joined_nlp)
```

Now, let's find the sentences that include named entities labeled as LOC. As we go, we put the results into a dataframe.

```
In [4]: sents_with_loc = pd.DataFrame({'doc number': [], 'location': [], 'sentence': []})
for idx, doc in enumerate(nlp_corpus):
    for sent in doc.sents:
        locations = []
        for ent in sent.as_doc().ents:
            if ent.label_ == "LOC":
                locations.append(ent.text)
        if locations:
            sents_with_loc.loc[len(sents_with_loc)] = [idx, locations, sent]
print('We have', len(sents_with_loc), 'sentences with locations.')
```

We have 7728 sentences with locations.

```
In [8]: pd.options.display.max_colwidth = 100
sents_with_loc.sample(n=10, random_state=42)
```

Out[8]:

	doc number	location	sentence
3328	17.0	[GeWOFAG, Mehrkosten]	(Die, landeshauptstadt, München, hat, sich, 2015, aufgrund, einer, Stadtratsanfrage, einer, poli...
4059	19.0	[Stadt]	(Mit, einer, Grundstücksfläche, von, 25.000, Quadratmetern, und, einer, Bruttogeschossfläche, vo...
5767	22.0	[Steigenberger, Airport-Hotels, Frankfurt am Main, Fachverband]	(Für, Dach-, und, Fassadenbegrünung, ,, ,, Am, Montag, ,, dem, 19.02.1990, ,, erschienen, im, Sa...
263	1.0	[Oberflächengewässer]	(..., ..., \n \n , 7, Immission, Oberflächengewässer, ...)
2835	15.0	[Gründach, Fassadengrün, Photovoltaik]	(Bei, einem, EnergieGrünDach, oder, einer, EnergieGrünFassade, handelt, es, sich, um, eine, komb...
2519	14.0	[Wissen]	(Wissen, auf, sehr, anschauliche, und, greifbare, Weise, \n \n , Wiedner, Hauptstraße)
5395	21.0	[Fremdaufwuchs]	(Ebenso, Fremdaufwuchs, .)
2036	13.0	[MEISSE]	(MEISSE, \n \n )
3846	18.0	[SNFCC, Nationaloper, Nationalbibliothek]	(Das, neue, Kulturzentrum, SNFCC, mit, Nationaloper, und, Nationalbibliothek, ist, eingebettet, ...
5850	22.0	[Hauptkriterien]	(2, ,, Motivation, ,, ,, Umfeldverbesserung, ", Ökologie-/Umweltaspekte, ,, Aufenthaltsqualität,...

# Identifying countries and cities with GeoText

Now, let's catch the country and city mentions using GeoText. Here is a simple example how it works:

```
In [26]: gt_obj = GeoText('I love New York, Seattle, Beijing, and also Warsaw.')  
gt_obj.country_mentions, geo.cities
```

```
Out[26]: (OrderedDict([('US', 2), ('CN', 1), ('PL', 1)]),  
          ['New York', 'Seattle', 'Beijing', 'Warsaw'])
```

The GeoText function is not flawless as illustrated here:

```
In [25]: GeoText("Where is Munchen?").country_mentions
```

```
Out[25]: OrderedDict()
```



For our text:

```
In [4]: sents_with_loc = pd.read_pickle('sents_with_loc.pkl')
sents_with_loc['country mentions'] = pd.Series([GeoText(sent).country_mentions for sent
in sents_with_loc['sentence']],
                                                index = sents_with_loc.index)
sents_with_loc['cities'] = pd.Series([GeoText(sent).cities for sent in sents_with_loc['s
entence']],
                                    index = sents_with_loc.index)
sents_with_loc = sents_with_loc[sents_with_loc['country mentions'] != {}]

print('We have', len(sents_with_loc), 'sentences now.')
sents_with_loc.head(10)
```

We have 1362 sentences now.

Out[4]:

	doc number	location	sentence	country mentions	cities
0	0	['Biodiversitätsdach']	Biodiversitätsdach auf dem Besucherzentrum der...	{'DE': 1}	[Berlin]
1	0	['Berlin', 'Pflanzenauswahl']	Dachbegrünung und Biodiversität\n\n2\n\n Be...	{'DE': 1}	[Berlin]
11	0	['Schorndorf']	Ansprechpartner im Bereich der Nisthilfen war ...	{'DE': 1}	[Schorndorf]
12	0	['Falkensee']	Für das Totholz zeichnete sich die Firma Kusch...	{'DE': 1}	[Falkensee]
16	1	['Hamburg', 'Hannover', 'Berlin', 'Kornwestheim']	(Deutscher Dachgärtner Verband e.V. - DDV), Ju...	{'DE': 5}	[Hamburg, Stuttgart, Hannover, Berlin, Kornwes...
26	1	['Berlin', 'Hamburg', 'Hannover', 'Ludwigsburg...']	Neben aktuellen Gründach-Initiativen aus Berli...	{'DE': 5}	[Berlin, Hamburg, Hannover, Ludwigsburg, Stutt...
27	1	['Hannover']	25\n\n Hannover:	{'DE': 1}	[Hannover]
28	1	['Stuttgart']	Gebäudebegrünung als Bestandteil der Klimaanpa...	{'DE': 1}	[Stuttgart]
30	1	['Berlin']	32\n\n Berlin:	{'DE': 1}	[Berlin]
31	1	['Ludwigsburg']	Ökologische Gebäudekonzepte und Modellvorhaben...	{'DE': 1}	[Ludwigsburg]

## Country mentions

Next, we'll count: which countries were mentioned the most?

```

In [3]: country_count = collections.defaultdict(int)
        for index, row in sents_with_loc.iterrows():
            for code, value in row['country mentions'].items():
                country_count[code] += value

        city_count = collections.defaultdict(int)
        for index, row in sents_with_loc.iterrows():
            for city in row['cities']:
                city_count[city] += 1

        country_data = pd.DataFrame({
            'country code': [code for code in country_count.keys()],
            'freq': [country_count[code] for code in country_count.keys()]})

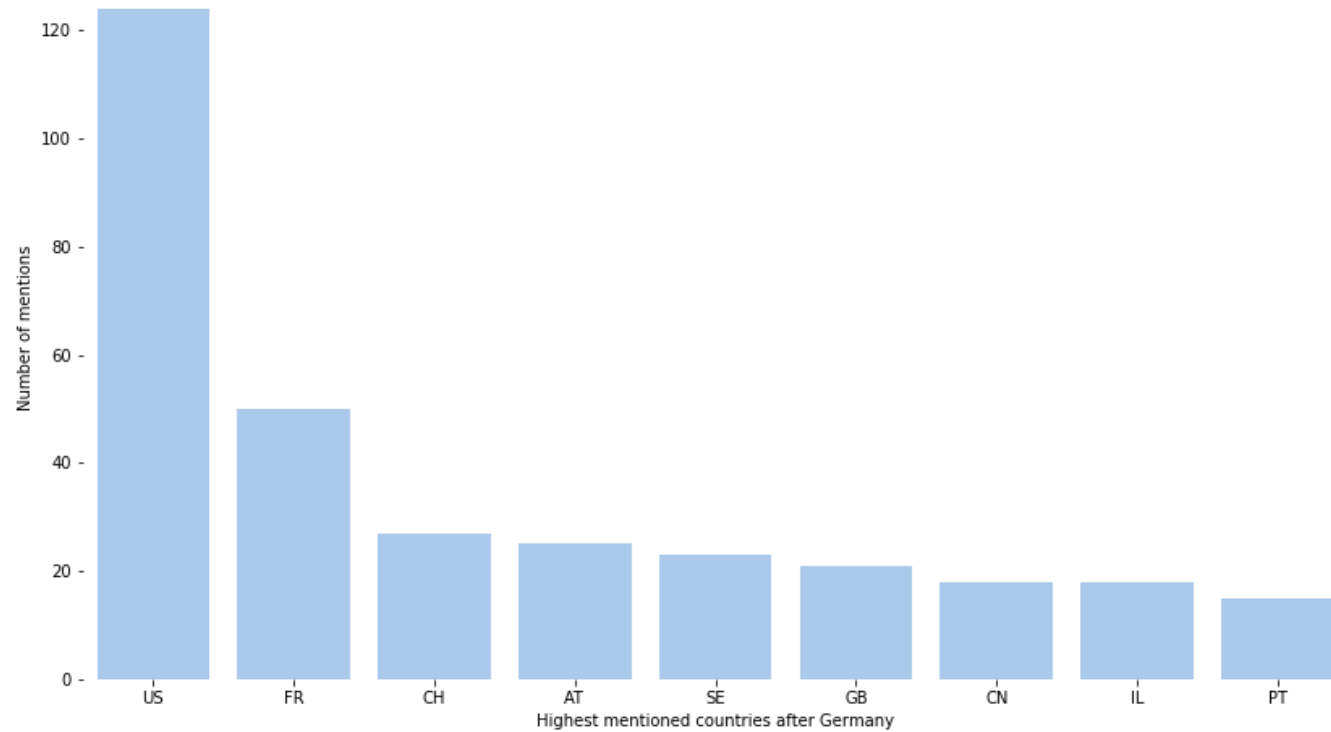
        country_data = country_data.sort_values(by = 'freq', ascending = False)
        country_data.head(10)

```

Out[3]:

	country code	freq
0	DE	1290
4	US	124
3	FR	50
20	CH	27
23	AT	25
2	SE	23
8	GB	21
7	CN	18
17	IL	18
18	PT	15

Omitting Germany, we have the following barchart:



# Visualization on maps

Let's visualize this! We get coordinates for the countries first:

```
In [4]: url = 'https://developers.google.com/public-data/docs/canonical/countries_csv'
html = requests.get(url).content
df_coords = pd.read_html(html)
df_coords = df_coords[0]
df_coords.head()
```

Out[4]:

	country	latitude	longitude	name
0	AD	42.546245	1.601554	Andorra
1	AE	23.424076	53.847818	United Arab Emirates
2	AF	33.939110	67.709953	Afghanistan
3	AG	17.060816	-61.796428	Antigua and Barbuda
4	AI	18.220554	-63.068615	Anguilla

We use folium to put markers proportional to the frequencies:

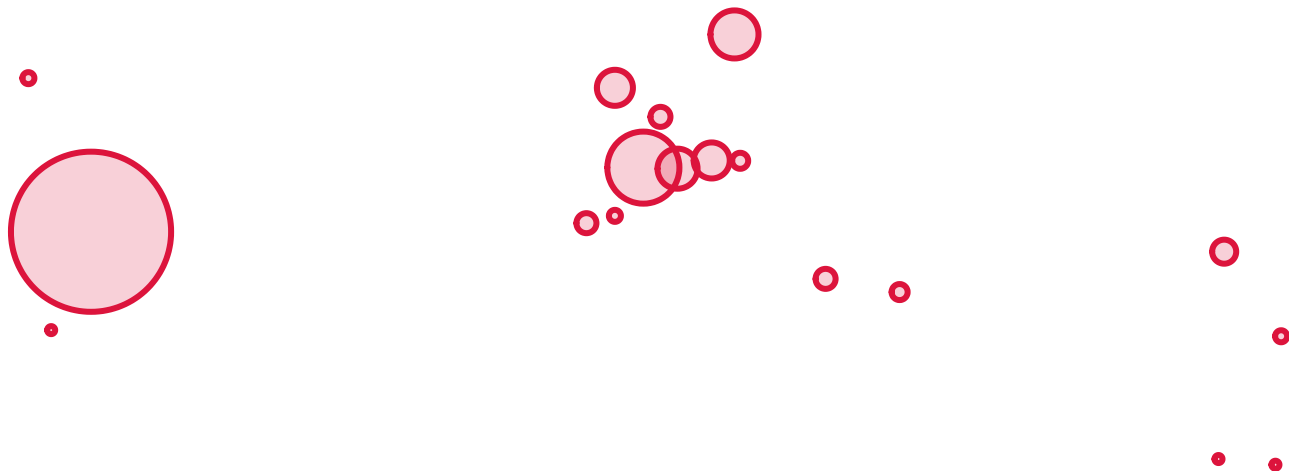
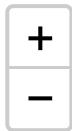
```
In [5]: # Make an empty map
m = folium.Map(location=[20,0], tiles="Mapbox Bright", zoom_start=2)
data = country_data

# we add markers one by one to the map
for i in range(1,20):
    code = data.iloc[i]['country code']
    lon = int(df_coords[df_coords['country'] == code]['longitude'])
    lat = int(df_coords[df_coords['country'] == code]['latitude'])
    folium.Circle(
        location=[lat, lon],
        popup = code + ': ' + str(data.iloc[i]['freq']) + ' mentions', # pop-up label above marker
        radius = int(data.iloc[i]['freq']) * 10000,
        color = 'crimson',
        fill = True,
        fill_color = 'crimson'
    ).add_to(m)

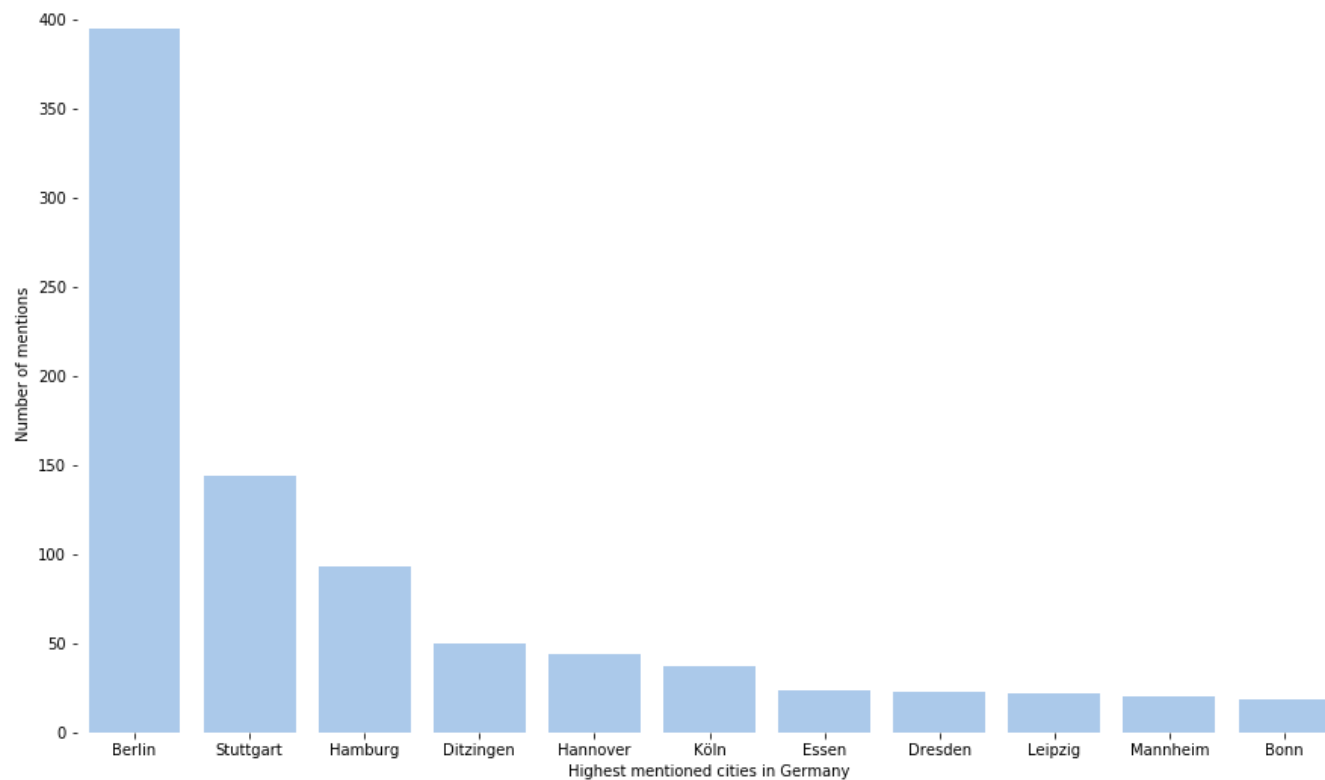
m.save('world_mentions.html')
```



```
In [1]: %%html
<iframe src="world_mentions.html" width="1200" height="800"></iframe>
```



Leaflet (<http://leafletjs.com>)





# **Classifying location mentions**

How can we tell if a locations is relevant? (What is relevant?)

Let's see a sample of our sentences with locations:

```
In [28]: sample = sents_with_loc.sample(n = 5, random_state = 42)

for index, row in sample.iterrows():
    print(row['sentence'].replace('\n', ' '))
    print('--->The countries were:', [c for c in row['country mentions'].keys()],
          'The cities were:', row['cities'], '\n')
```

Gründach-Siedlung in Berlin

--->The countries were: ['DE'] The cities were: ['Berlin']

48329 Havixbeck D 50226 Frechen D 50739 Köln      www.marcel-nadorf.com www.benning-dachbegruenung.de

--->The countries were: ['DE'] The cities were: ['Frechen', 'Köln']

Im Pommerfeld 2 56630 Kretz / Andernach fon +49 (0) 26 32 - 95 48-0 fax +49 (0) 26 32 - 95 48-20      www.vulkatec.de info@vulkatec.de

--->The countries were: ['DE'] The cities were: ['Andernach']

FBB-Symposium Gebäudegrün      Am 20. Februar 2018 findet in Berlin im Rahmen der Grünbau während der Messe Bautech das FBB-Symposium Gebäudegrün statt.

--->The countries were: ['DE'] The cities were: ['Berlin']

Auch hier zeigt sich, dass durch Bewässerung eine höhere CO<sub>2</sub>-Aufnahme im Jahresgang erwartet werden kann.

--->The countries were: ['FR'] The cities were: ['Auch']

## First approach: topic modelling

```
In [29]: nltk.download('stopwords')
# we have some English text in the German too so add those as well
from nltk.corpus import stopwords
stop_words = set(stopwords.words('german'))
stop_words = stop_words.union(set(stopwords.words('english')))

corpus = [sent for sent in sents_with_loc['sentence']]

vect = CountVectorizer(min_df=20, max_df=0.2, stop_words=stop_words,
                      token_pattern='(?u)\\b\\w\\w+\\b')

# create the sparse matrix
X = vect.fit_transform(corpus)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\admin1\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

```
In [30]: # IDs to words dict
id_map = dict((v, k) for k, v in vect.vocabulary_.items())

# Convert sparse matrix to gensim corpus.
corpus = gensim.matutils.Sparse2Corpus(X, documents_columns=False)

# Train the LDA model
from gensim.models.ldamodel import LdaModel

ldamodel = LdaModel(corpus=corpus,
                    id2word=id_map,
                    num_topics=10,
                    passes=10,
                    random_state = 42)
```



If we check the topics, nothing stands out as super irrelevant but **topic 3 and 5** might worth our attention.

```
In [31]: ldamodel.print_topics(10)
```

```
Out[31]: [(0,
  '0.257*"hannover" + 0.160*"2018" + 0.145*"dach" + 0.137*"2013" + 0.113*"rahmen" +
  0.049*"stadt" + 0.028*"universität" + 0.025*"leipzig" + 0.016*"gebäudegrün" + 0.014
  *"dächer"'),
  (1,
    '0.224*"2017" + 0.121*"green" + 0.097*"roofs" + 0.086*"frankfurt" + 0.069*"dachbegr
    ünung" + 0.068*"london" + 0.067*"internationalen" + 0.061*"leipzig" + 0.051*"essen" +
    0.043*"gründach"'),
  (2,
    '0.177*"fbb" + 0.148*"ditzingen" + 0.120*"bugg" + 0.103*"gründachsymposium" + 0.071
    *"dachbegrünungen" + 0.067*"2018" + 0.063*"stuttgart" + 0.060*"gebäudegrün" + 0.042
    *"statt" + 0.028*"beim"'),
  (3,
    '0.425*"gmbh" + 0.354*"stuttgart" + 0.067*"usa" + 0.045*"mannheim" + 0.020*"dresde
    n" + 0.020*"dachbegrünung" + 0.019*"dach" + 0.008*"www" + 0.005*"bad" + 0.003*"fb
    b"'),
  (4,
    '0.178*"city" + 0.144*"begrünung" + 0.133*"boden" + 0.123*"köln" + 0.113*"dass" +
    0.059*"grün" + 0.048*"dach" + 0.047*"ddv" + 0.042*"gebäude" + 0.027*"essen"'),
  (5,
    '0.558*"www" + 0.111*"com" + 0.064*"bad" + 0.059*"tel" + 0.052*"institut" + 0.035
    *"dächer" + 0.022*"köln" + 0.017*"gmbh" + 0.017*"dresden" + 0.015*"universität"'),
  (6,
    '0.168*"2016" + 0.165*"fassadenbegrünung" + 0.132*"gebäudebegrünung" + 0.129*"fbb"
    + 0.094*"prof" + 0.040*"dach" + 0.038*"münchen" + 0.033*"grün" + 0.032*"institut" +
    0.029*"2014"'),
  (7,
    '0.156*"gründach" + 0.123*"hamburg" + 0.090*"stuttgart" + 0.078*"umwelt" + 0.070*"m
    ünchen" + 0.066*"sowie" + 0.066*"teil" + 0.059*"jahr" + 0.058*"universität" + 0.055
    *"gebäude"'),
  (8,
    '0.202*"hamburg" + 0.184*"wurde" + 0.114*"stadt" + 0.083*"2017" + 0.065*"gebäudegrü
    n" + 0.060*"jahren" + 0.043*"beim" + 0.041*"dresden" + 0.035*"2014" + 0.029*"grün"'),
  (9,
    '0.261*"green" + 0.194*"deutschland" + 0.127*"roof" + 0.119*"urban" + 0.089*"statt"
```

Let's look at the topic distribution of each sentence and filter for sentences with Topic 3 or 5 being dominant.

```
In [32]: def topic_distribution(new_doc_list):
new_transformed = vect.transform(new_doc_list)
new_corpus = gensim.matutils.Sparse2Corpus(new_transformed, documents_columns=False)
return ldamodel[new_corpus[0]]

def dominant_topic(new_doc_list):
topics = topic_distribution(new_doc_list)
topics.sort(key = lambda x: x[1], reverse=True)
return topics[0]

sents_with_loc['dominant topic'] = pd.Series([dominant_topic([sent])
for sent in sents_with_loc['sentence']])
sents_with_loc
```

Out[32]:

	index	doc number	location	sentence	country mentions	cities	dominant topic
0	0	0	['Biodiversitätsdach']	Biodiversitätsdach auf dem Besucherzentrum der...	{'DE': 1}	[Berlin]	(1, 0.5499903)
1	1	0	['Berlin', 'Pflanzenauswahl']	Dachbegrünung und Biodiversität\n\n2\n\nBe...	{'DE': 1}	[Berlin]	(1, 0.69998765)
2	11	0	['Schorndorf']	Ansprechpartner im Bereich der Nisthilfen war ...	{'DE': 1}	[Schorndorf]	(3, 0.5499983)
3	12	0	['Falkensee']	Für das Totholz zeichnete sich die Firma Kusch...	{'DE': 1}	[Falkensee]	(3, 0.5499983)
4	16	1	['Hamburg', 'Hannover', 'Berlin', 'Kornwestheim']	(Deutscher Dachgärtner Verband e.V. - DDV), Ju...	{'DE': 5}	[Hamburg, Stuttgart, Hannover, Berlin, Kornwes...	(7, 0.77570164)
5	26	1	['Berlin', 'Hamburg', 'Hannover', 'Ludwigsburg...]	Neben aktuellen Gründach- Initiativen aus Berli...	{'DE': 5}	[Berlin, Hamburg, Hannover, Ludwigsburg, Stutt...	(7, 0.84997654)
6	27	1	['Hannover']	25\n\nHannover:	{'DE': 1}	[Hannover]	(0, 0.54999775)
7	28	1	['Stuttgart']	Gebäudebegrünung als Bestandteil der Klimaanpa...	{'DE': 1}	[Stuttgart]	(6, 0.378283)
8	30	1	['Berlin']	32\n\nBerlin:	{'DE': 1}	[Berlin]	(0, 0.1)
9	31	1	['Ludwigsburg']	Ökologische Gebäudekonzepte und Modellvorhaben...	{'DE': 1}	[Ludwigsburg]	(0, 0.1)

	index	doc number	location	sentence	country mentions	cities	dominant topic
10	37	1	['Berlin', 'Hamburg', 'Hannover', 'Ludwigsburg...]	Die in dieser Broschüre vorgestellten praktisc...	{'DE': 5}	[Berlin, Hamburg, Hannover, Ludwigsburg, Stutt...	(8, 0.27521625)
11	44	1	['Pilotprojektes', 'Hamburg', 'BUE', 'Stuttgar...]	Zu den weiteren Kooperationspartnern des Pilot...	{'DE': 2}	[Karlsruhe, Stuttgart]	(7, 0.8150366)
12	65	1	['Straße', '\n ']	Man sieht an dieser Stelle bereits, dass die v...	{'CI': 1, 'SE': 1}	[Man, Boden]	(4, 0.77499545)
13	119	1	['Ökosystemleistungen']	Die damit einhergehende Quantifizierung der ge...	{'SE': 1}	[Boden]	(4, 0.36666808)
14	126	1	['Kommunales', 'Berlin']	/ Kommunales Grünprogramm Berlin:	{'DE': 1}	[Berlin]	(0, 0.1)
15	128	1	['\n\n\x0cHamburg\n ']	TH Treibhaus Landschaftsarchitekten, Luftbild:...	{'DE': 1}	[Hamburg]	(7, 0.7749931)
16	129	1	['Hamburg']	Hamburg:	{'DE': 1}	[Hamburg]	(8, 0.5499831)
17	154	1	['Hannover', 'Hannover', 'Zahl']	Nach der von der Landeshauptstadt in 2010 beau...	{'DE': 2}	[Hannover, Hannover]	(0, 0.69999844)
18	158	1	['Hannover']	Mögliche Folgen des Klimawandels für Hannover	{'DE': 1}	[Hannover]	(0, 0.54999757)
19	172	1	['\n\n\x0cHannover']	Teil II Kommunale Gründach-Strategien 29\n\n ...	{'DE': 1}	[Hannover]	(7, 0.61583203)
20	180	1	['Hannover']	Die Nord/LB in Hannover verfügt über eine knap...	{'DE': 1}	[Hannover]	(0, 0.3666658)
21	183	1	['Hannover']	Von diesen Gründachflächen fließen (bezogen au...	{'DE': 1}	[Hannover]	(7, 0.35741398)
22	185	1	['Stadt Hannover', 'Stadt', 'Hannover']	Ein Beispiel für Fördermöglichkeiten bietet da...	{'DE': 1}	[Hannover]	(0, 0.72287095)
23	192	1	['\n\n\x0cStuttgart']	Teil II Kommunale Gründach-Strategien 31\n\n ...	{'DE': 1}	[Stuttgart]	(7, 0.7749838)
24	193	1	['Stuttgart']	Stuttgart:	{'DE': 1}	[Stuttgart]	(3, 0.54998803)
25	196	1	['Stuttgart']	Urbanes Gärtnern in Stuttgart	{'DE': 1}	[Stuttgart]	(3, 0.5499881)
26	198	1	['Stuttgart']	Doch Nutzgärten haben in vielen Städten eine l...	{'DE': 1}	[Stuttgart]	(3, 0.5499881)
27	202	1	['Stuttgart']	Weitere Gärten in Stuttgart sind geplant.	{'DE': 1}	[Stuttgart]	(3, 0.5499881)

	index	doc number	location	sentence	country mentions	cities	dominant topic
28	210	1	['Stadt Stuttgart', 'Stuttgart', 'Stuttgart']	© Stadt Stuttgart\n\n In den letzten Jahren s...	{'DE': 2}	[Stuttgart, Stuttgart]	(8, 0.41395968)
29	217	1	['Landes-\n\n hauptstadt', 'Stuttgart']	Die Gemeinderäte haben der neuen Richtlinie fü...	{'DE': 1}	[Stuttgart]	(6, 0.62410635)
...	...	...	...	...	...	...	...
1332	7616	24	['Darmstadt']	Darmstadt\n\n	{'DE': 1}	[Darmstadt]	(0, 0.1)
1333	7617	24	['Bulgaria']	Bulgaria\n\n	{'BG': 1}	[]	(0, 0.1)
1334	7618	24	['South Korea']	South Korea\n\n	{'KR': 1}	[]	(0, 0.1)
1335	7619	24	['Seoul', 'Korea\n\n Germany Germany']	curve and keyword analysis for biodiversity co...	{'KR': 1}	[Seoul]	(0, 0.1)
1336	7620	24	['Design-for-Safety']	Why we need Design-for-Safety (Dfs) for Skyris...	{'FR': 1}	[]	(9, 0.69999999)
1337	7621	24	['Deutschland']	Switzerland\n\n Artenschutz über den Köpfen -...	{'CH': 1, 'AT': 1, 'KR': 1}	[Seoul]	(9, 0.87141585)
1338	7623	24	['Denver Colorado', 'Denver', 'Colorado', 'USA']	Denver Colorado\n\n BIO-TECTURE Living Wall S...	{'DE': 1, 'US': 1, 'BR': 1}	[Denver, Colorado]	(3, 0.54999947)
1339	7624	24	['Poland\n ']	Living walls in public spaces in Poland\n	{'PL': 1}	[]	(0, 0.1)
1340	7630	24	['Vienna', 'Tel.', '\n Prof. Dr.-Ing']	Architects Bernardgasse 21 A - 1070 Vienna, Te...	{'US': 1}	[Vienna]	(5, 0.5912214)
1341	7634	24	['Madrid']	2 E - 28040 Madrid\n Dr. rer. hort.	{'ES': 1}	[Madrid]	(0, 0.1)
1342	7637	24	['Pokfulam']	Y. Chen Department of Geography, The Universit...	{'HK': 4}	[]	(0, 0.1)
1343	7638	24	['Berlin']	D - 10965 Berlin	{'DE': 1}	[Berlin]	(0, 0.1)
1344	7639	24	['\n ']	DD-Berlin.de\n	{'DE': 1}	[Berlin]	(0, 0.1)
1345	7642	24	['Bolzano', 'Bozen']	22 Mühlbachpromenade 22 I - 39100 Bolzano - Bo...	{'IT': 1}	[Bolzano]	(5, 0.55)
1346	7643	24	['Budapest']	Elnök utca 24 HUN - 1089 Budapest pdezsenyi@gr...	{'HU': 1}	[Budapest]	(0, 0.1)
1347	7645	24	['USA', 'Boston']	20 Custom House Street Suite 800 USA - Boston,...	{'US': 1}	[Boston]	(5, 0.52477914)
1348	7653	24	['\n\n Prof:']	Babul Asociacion de Infraestructura Verde de Ch...	{'CL': 1}	[]	(5, 0.3983512)
1349	7656	24	['Bonn\n Marco Fritz European Commission DG Re...	II 6 Bauen und Umwelt Deichmanns Aue 31-37 D...	{'DE': 2}	[Aue, Bonn]	(7, 0.54999994)

```
In [35]: def filter_for_topic(df, top_id_list):
    filtered_df = df[[topic[0] in top_id_list for topic in df['dominant topic']]]
    return filtered_df[['sentence', 'dominant topic']]

topic5 = filter_for_topic(sents_with_loc, [3, 5])
topic5['weight'] = [p[1] for p in topic5['dominant topic']]
topic5.sort_values(by='weight', ascending = False, inplace = True)
topic5.drop(columns = ['weight'], inplace = True)

print('We have', len(topic5), 'sentences here with Topic 3 or 5 dominant which is',
      100 * len(topic5)//len(sents_with_loc), "percent." )
topic5.head(10)
```

We have 234 sentences here with Topic 3 or 5 dominant which is 17 percent.

Out[35]:

	sentence	dominant topic
991	122\n\n D 63768 Hösbach D 97486 Königsberg\n\n www.ild-group.com www.benkert-dachbegruenung.de...	(5, 0.9181739)
843	www.iasp.asp-berlin.de www.gruendach-mv.de www.fbg.fh-wiesbaden.de www.baubotanik.org\n\n 73730...	(5, 0.90999997)
1120	Viernheim\n\n www.novihum.de www.aco-hochbau.de www.sachsenband.de www.xeroflor.com www.xeroflo...	(5, 0.8874999)
844	Bad Urach-Hengen 77933 Lahr 88633 Heiligenberg-Steigen\n\n www.xeroflor.de www.xeroflor.com www.	(5, 0.8499999)
965	www.gruendach-siebert.de www.vedag.de whawiw.vedag.com.cn\n\n D 01307 Dresden\n\n www.novihum....	(5, 0.84999585)
1089	Die Wiese - Naturnahe Gärten GmbH Immo Herbst Dach- + Innenraumbegrünungs GmbH Helix Pflanzensys...	(3, 0.8499852)
955	D 81927 München\n\n www.kramer-gartenbau.de\n\n D 86477 Adelsried\n\n www.garten-koenig.com\n...	(5, 0.8272736)
836	China www.vedag.com\n\n DE\n\n 25494 Heist bei Hamburg\n\n www.sachsenband.de\n\n DE\n\n 28...	(5, 0.8272732)
1358	+43-699-19744304 bernhard.scharf@green4cities.com www.green4cities.com\n\n Rudi Scheuermann Aru...	(5, 0.82631105)
988	www.schadenberg.nl\n\n SI 4260 Bled\n\n www.greenwalls.si\n\n US MD 21211 Baltimore\n\n www....	(5, 0.8199999)



**Thank you for your attention!**

**([https://docs.google.com/presentation/d/1PzeVghmDcjDCG20aRibq6cDBoHCELQHdyn0zAVlag\\_w/edit](https://docs.google.com/presentation/d/1PzeVghmDcjDCG20aRibq6cDBoHCELQHdyn0zAVlag_w/edit))**