2_nlp_validation_schlaak_weise

August 23, 2020

Von Pascal Schlaak, Tim Weise - Natural Language Processing (SoSe 20)

1 Datenvalidierung

Bevor eine konkrete Verarbeitung der Daten durchgeführt werden kann, sollten diese auf Korrektheit validiert werden. Es soll ebenso überprüft werden, ob sich diese Daten für eine weitere Analyse eignen. Hierzu dienen einzelne Einträge der Datenbasis als Stichprobe.

1.1 Module importieren

Zur Verarbeitung der Datenbasis werden folgende Module benötigt und müssen zuerst importiert werden.

[1]:
```
import spacy
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from collections import Counter
```

1.2 Daten einlesen

Zum Einlesen der Daten nutzen wird das Modul `Pandas`, welches sich im Data Analytics-Bereich etabliert hat und eine geeignete Schnittstelle zur Darstellung und Verarbeitung von Daten bietet. Die `read_json()` Methode beinhaltet Funktionalität zum Einlesen der Datenbasis in ein *DataFrame*, ausgehend von einer `JSON`-Datei. Ein `DataFrame` stellt in Pandas eine Datenstruktur dar, welche Achsenbeschriftungen beinhaltet und arithmetische Operationen entlang der Achsen mit geringem Aufwand ermöglicht.

[2]:
```
# Pfad zu JSON-Datei
PATH_TO_DATA = '../data/movies.json'
# Einlesen der JSON-Datei in Dictionary
data = pd.read_json(PATH_TO_DATA)
```

Ein Aufrufen der Variable `data`, welche unsere Datenbasis beinhaltet, gibt uns die gecrawlten Daten als Tabelle aus.

[3]:
```
data
```

[3]:
```
                                       title  date  rank  \
0                      The Shawshank Redemption  1994     1
1                               The Dark Knight  2008     3
2                                  The Godfather  1972     2
3                        The Godfather: Part II  1974     4
4                                  Pulp Fiction  1994     6
..                                          …   …   …
245                               Green Book  2018   125
246                                  The Kid  1921   124
247                               Metropolis  1927   123
248  Three Billboards Outside Ebbing, Missouri  2017   126
249                                       M  1931   122

                                              synopsis
0    In 1947, Andy Dufresne (Tim Robbins), a banker…
1    The movie begins with a gang of men with clown…
2    In late summer 1945, guests are gathered for t…
3    The Godfather Part II presents two parallel st…
4    Late one morning in the Hawthorne Grill, a res…
..                                                 …
245  In 1962 in New York City, Tony "Tony Lip" Vall…
246  \n                It looks like we don't have …
247  (This is the synopsis of the full 150 minute v…
248  Mildred Hayes drives down Drinkwater Road near…
249  It's noon. Concerned parents are lined up outs…

[250 rows x 4 columns]
```

Die Struktur der Daten sieht auf den ersten Blick valide aus.

Pandas arithmetische Operationen erlauben beispielsweise das Sortieren der Dateneinträge. In diesem Fall kann die Methode `sort_values()` verwendet werden. Wir können unsere Dateneinträge somit nach ihrem Rang in der Top 250-Liste sortieren.

[4]:
```
data.sort_values(by=['rank'])
```

[4]:
```
                                         title  date  rank  \
0                        The Shawshank Redemption  1994     1
2                                   The Godfather  1972     2
1                                The Dark Knight  2008     3
3                          The Godfather: Part II  1974     4
5    The Lord of the Rings: The Return of the King  2003     5
..                                             …   …   …
183                                 The General  1926   246
179                                    The Help  2011   247
178                                 Before Sunset  2004   248
177                               Monsters, Inc.  2001   249
181                               The Terminator  1984   250
```

```
                                                synopsis
0    In 1947, Andy Dufresne (Tim Robbins), a banker…
2    In late summer 1945, guests are gathered for t…
1    The movie begins with a gang of men with clown…
3    The Godfather Part II presents two parallel st…
5    In the opening scene, a flashback, two hobbits…
..                                                 …
183  The Western & Atlantic Flyer "speeds into Mari…
179  In civil-rights era Jackson, Mississippi, 23-y…
178  Nine years have passed since the events of Bef…
177  The city of Monstropolis in a world entirely p…
181  Over an apocalyptic battlefield in the year 20…

[250 rows x 4 columns]
```

1.3 Tokenisierung

Um nun ebenfalls stichprobenartig unsere Daten auf Textebene validieren zu können, bietet es sich an, eine Tokenisierung der Zusammenfassung vorzunehmen. Hierbei nutzen wir das Modul `spaCy` für eine Segmentierung unseres Textes auf Wortebene. SpaCy bietet Funktionalitäten zur Verarbeitung natürlicher Sprache, in unserem Fall der englischen Sprache. Als Beispiel wird der erste Eintrag der Datenbasis untersucht, also der am besten bewertete Film auf IMDb: Es handelt sich hierbei um den Film *The Shawshank Redemption (deutsch: Die Verurteilten)*.

Wir benötigen dazu ein Sprachmodell, welches einzelne Token klassifizieren kann. Falls ein Importieren der Sprachmodelle scheitern sollte, können diese separat über den folgenden Befehl installiert werden:

```
python -m spacy download <sprachmodell z.B. en_core_web_sm>
```

Wir beginnen mit dem Laden eines schlanken, englischen Sprachmodells, um unsere Textdaten weiter analysieren zu können. Dieses schlanke Modell `en_core_web_sm` bringt den Vorteil mit sich, dass es effizienter ist, im späteren Verlauf soll ein umfangreicheres Sprachmodell zum Einsatz kommen. `en_core_web_sm` enthält ein Vokabular, Syntax und Eigennamen der englischen Sprache, welches wir für die Analyse eigener Texte nutzen können. Wir erstellen eine Variable `document`, die als Ergebnis von `nlp()`, unsere tokenisierte Zusammenfassung beinhaltet.

[5]:
```
# Laden eines englischen Sprachmodells
nlp = spacy.load("en_core_web_sm")
# Tokenisieren der Zusammenfassung
document = nlp(data['synopsis'][0])
```

Wir erhalten folgende Sequenz von Token-Objekten.

[6]:
```
document
```

[6]:
```
In 1947, Andy Dufresne (Tim Robbins), a banker in Maine, is convicted of
murdering his wife and her lover, a golf pro. Since the state of Maine has no
```

death penalty, he is given two consecutive life sentences and sent to the notoriously harsh Shawshank Prison. Andy keeps claiming his innocence, but his cold and calculating demeanor leads everyone to believe he did it.Meanwhile, Ellis Boyd Redding (Morgan Freeman), known as Red is being interviewed for parole after having spent 20 years at Shawshank for murder. Despite his best efforts and behavior, Red's parole is rejected which doesn't phase him all that much. Red is then introduced as the local smuggler who can get inmates anything they want within reason. An alarm goes off alerting all prisoners of new arrivals. Red and his friends bet on whichever new fish will have a nervous break down during his first night in prison. Red places a huge bet on Andy.During the first night, an overweight newly arrived inmate, nicknamed ''fat ass', breaks down and cries hysterically allowing Heywood (William Sadler) to win the bet. However, the celebration is short lived when the chief guard, Byron Hadley (Clancy Brown), savagely beats up the fat man for not keeping quiet when he is asked to. Meanwhile, Andy remains steadfast and composed. The next morning, the inmates learn that ''fat ass'' died in the infirmary because the prison doctor had been out for the night. Andy inquires about the man's name only to get put down by Heywood.About a month later, Andy approaches Red having heard of his talents for finding things. He asks Red to find him a rock hammer, an instrument he claims is necessary for his hobby of rock collecting and sculpting. Red asks a few questions about his intentions which Andy laughs off. Red agrees to place the order and also warns Andy about ''the sisters'', a group of prisoners who sexually assaults other prisoners, most importantly their leader, Boggs (Mark Rolston) who has a crush on Andy. Though other prisoners consider Andy "a really cold fish," Red sees something in Andy, and likes him from the start. Red thinks Andy intends to use the hammer to engineer an escape in the future but when he finally sees the tool's actual size, he understands why Andy laughed and laughs too, putting aside the thought that Andy could ever use it to dig his way out of prison.During the first two years of his incarceration, Andy spends most of his time working in the prison laundry or fighting off Boggs and the Sisters. Though he persistently resists and fights them every time, Andy is beaten and raped on a regular basis but keeps quiet about it.When a work detail for tarring the roof of one of the prison's buildings is announced, Red pulls some strings to get Andy and a few of their mutual friends assigned to the job, giving everyone a break from the usual. During the job Andy overhears Hadley complaining about having to pay taxes for an upcoming inheritance. Drawing from his expertise as a banker, Andy lets Hadley know how he can shelter his money from the IRS by turning it into a one-time gift for his wife. He then offers to assist Hadley in filling out the paperwork in exchange for some cold beers for his fellow inmates while on the tarring job. Hadley first threatens to throw Andy off the roof, but eventually agrees and do provide the working inmates with cold beers before the job is finished. Red remarks that Andy may have engineered the privilege to build favor with the prison guards as much as with his fellow inmates, but he also thinks Andy did it simply to "feel normal again."While watching a movie, Andy approaches Red with another unusual demand and asks for the actress Rita Hayworth. Red is surprised by the demand but agrees to place the order. As he

exits the theater, Andy once more encounters the Sisters. Although he is able to talk his way out of being raped, he is brutally beaten within an inch of his life, putting him in the infirmary for a month. Boggs spends a week in solitary for the beating. When he comes out, he finds Hadley and his men waiting in his cell. They beat him so badly that he's left unable to walk or eat solid food for the rest of his life and is transferred to a prison hospital upstate. The Sisters move on and never bother Andy again. When Andy gets out of the infirmary, he finds a bunch of rocks for him to sculpt and a giant poster of Rita Hayworth in his cell; presents from Red and his friends.Warden Samuel Norton (Bob Gunton) hears about how Andy helped Hadley and uses a surprise cell inspection to size Andy up. He finds Andy reading his copy of the Holy Bible and they talk about their favorite verses while the guards are turning the cell upside down looking for illegal possessions. Satisfied with their encounter, the warden leaves and almost forget to give Andy his Bible back. He then encourages Andy to keep reading the Bible saying that ''Salvation lays within''.Andy is later advised that he will now work in the prison library with aging inmate Brooks Hatlen (James Whitmore). The reason for his transfer is made obvious when a prison guard shows up asking Andy for financial advising. Andy sets-up a makeshift desk and starts working, providing financial advising to most prison guards and helping them with their income tax returns. Andy also sees an opportunity to expand the prison library; he starts by asking the Maine state senate for funds. He writes letters every week. His financial support practice is so appreciated that even guards from other prisons, when they visit for inter-prison baseball matches, seek Andy's financial expertise. Even the warden himself has Andy preparing his tax returns.Not long afterwards, Brooks snaps and threatens to kill Heywood in order to avoid being paroled. Andy is able to talk him down. When his friends discuss Brooks 'behavior, Red sympathizes with Brooks having obviously become "institutionalized," after spending 50 years at Shawshank. He has become essentially conditioned to be a prisoner for the rest of his life and is unable to adapt to the outside world. Red remarks: "These walls are funny. First you hate 'em, then you get used to 'em. Enough time passes, you get so you depend on them." Brooks is paroled and goes to live in a halfway house. He is also given a job at a supermarket which he hates. Finding it impossible to adjust to life outside the prison, he eventually commits suicide, leaving the message "Brooks was here" carved on a wooden beam .After six years of writing letters, Andy receives $200 from the state for the library, along with a collection of old books and phonograph records. Though the state Senate thinks this will be enough to get Andy to halt his letter-writing campaign, he is undaunted and redoubles his efforts.When the donations of old books and records arrive at the warden's office, Andy finds a copy of Mozart's The Marriage of Figaro among the records. He locks the guard assigned to the warden's office in the bathroom and plays the record over the prison's PA system. The entire prison is soon captivated by the music. Red remarks that the voices of these women made everyone feel free, if only for a brief moment. Outside the office, Norton appears furious at the act of defiance, and orders Andy to turn off the record player. Andy responds by turning up the volume. The warden orders Hadley to break into the office and Andy is sent immediately to

solitary confinement for two weeks. When he gets out, he tells his friends that the stretch was the "easiest time" he ever did in the hole because he spent it with Mozart's Figaro stuck in his head for comfort. When the other prisoners tell him how unlikely that is, he talks about the power that hope can have in prison and that hope can sustain them. Red strongly disagrees with Andy, claiming that hope is a dangerous thing in a place like Shawshank and tells Andy he should get used to living without it. Andy implies that this is exactly what Brooks did and Red leaves the table angry.Not long after, Red has a new parole hearing and realizes he's been in prison for 30 years now. He uses the exact same words he used ten years earlier only with no enthusiasm at all. His parole is rejected again. Andy gives him an harmonica to commemorate his 30 years which Red replies by offering Andy a giant poster of Marilyn Monroe to commemorate his 10 years.About four years after the Mozart incident, the state senate finally comes to the conclusion that they won't get rid of Andy with just another check. So they allow him a budget of $500 a year to build his library. Andy uses it wisely and makes deals with book clubs and charities to create the best prison library in the state and names it after Brooks. With the enlarged library and more materials, Andy begins to mentor inmates who want to receive their high school diplomas so they can get a decent job once they're out.Meanwhile, Warden Norton profits from Andy's knowledge and devises a scheme whereby he puts prison inmates to work on public projects which he wins by outbidding other contractors (prisoners are cheap labor). Occasionally, he allows other contractors to score projects as long as the bribe is good enough. Andy launders the money by setting up several accounts in several banks, along with several investments, using the fake identity of Randall Stephens, a man who only exist on papers, created by Andy himself through his knowledge of the system and mail ordered forms. Randall Stephens officially has a birth certificate, social security number and driving license. Should anyone ever investigate about the scheme; they will chase a man who only exists on paper. Andy shares the details with Red, noting that he had to "go to prison to learn how to be a crook."In 1965, a young prisoner named Tommy (Gil Bellows) comes to Shawshank to serve time for breaking and entering. Tommy is easy going, charismatic, and popular among the other inmates and is befriended by both Andy and Red. When Tommy explains that he's been going in and out of prison ever since he was 13 years old, Andy suggests that Tommy should consider another line of work besides theft because he seems to be not so good at it. The suggestion really gets to Tommy and he asks Andy to help him work on earning his high school equivalency diploma. Though Tommy is a good student, he is still frustrated when he takes the exam itself, crumpling it up and tossing it in the trash. Andy retrieves it and sends it in anyway. Tommy asks Red about Andy's case which Red explains. Upon hearing the story, Tommy is visibly upset. He then tells Andy and Red the story of a former cellmate of his from another prison who boasted about killing a man who was a pro golfer at the country club he worked at, along with his lover. The woman's husband, a banker, had gone to prison for those murders.With this new information, Andy, full of hope, meets with the warden, expecting Norton to help him get a new trial with Tommy as a witness. The reaction from Norton is completely contrary to what Andy hoped for. When Andy says emphatically that he would never reveal the money laundering

schemes he set up for Norton over the years, the warden becomes furious and orders him to solitary for a month. The inmates discuss the sentence mentioning it is the longest time in solitary that they've ever heard of. They also realize that Andy may truly be innocent after all and has spent almost 20 years in prison for a crime he didn't commit.Tommy receives a letter from the board of education announcing that he has passed the exam and now owns a high school diploma. A guard pass the news to Andy in his solitary cell which makes him smile a little.Later on, Tommy is escorted outside at night to have a private meeting with the warden. Warden Norton asks him if the story he told Andy is true and if he would be willing to testify on Andy's behalf. Tommy enthusiastically agrees. The warden smiles at him before nodding to Hadley to shoot him dead.When the warden visits Andy in solitary, he tells him that Tommy tried to escape and that Hadley had no choice but to shoot him. Andy doesn't buy that story and tells Norton that ''everything'' stops and that he's not going to work for him anymore. The warden threatens Andy to shut down the library, burn all the books, and move Andy to a much different cell in a much different part of the prison with the most hardened criminals should he stop working for him. He then leaves and orders Andy to another month in solitary to think about things.When Andy finally comes out of solitary, he and Red have a conversation where Andy talks about his wife and how much he loved her and feels responsible for her death even though he didn't pulled the trigger. He then talks about his projects should he ever get out of prison. He talks about Zihuatanejo, a beach town on the Pacific coast of Mexico where he'd like to live for the rest of his life and manage a hotel there. He then asks Red if he'd join him to which Red says no and that he believes he is too far gone like Brooks. He then criticizes Andy for allowing hope to mess with his mind like that and that it will only destroy him. Andy agrees and is about to leave when he asks Red if he knows the Buxton, Maine area. He then tells Red about a very specific hay field where there is a large oak tree at the end of a stone wall. He then asks Red to promise him that, should he ever get paroled, he will seek that oak tree and retrieve something that was hidden among the stones but refuses to say what it is. Red promises but is worried about his friend's state of mind. His worries are heightened further when he learns that Andy has asked Haywood for a six-foot rope. Red believes Andy may have finally reached his breaking point and is about to commit suicide. Meanwhile, Norton asks Andy to shine his shoes for him and put his suit in for dry-cleaning before retiring for the night. Andy returns to his cell and the guards turn the lights off for the night. Red remarks that it was the longest night of his life.The following morning, Andy has not answered the morning call and is not standing in front of his cell like every morning. The guard yells at Andy for putting him late and walks to his cell expecting to find a seriously sick or dead Andy. At the same time, Norton becomes alarmed when he finds Andy's shoes in his shoe box instead of his own. The alarm then goes off announcing a missing inmate. Norton rushes to Andy's empty cell and demands an explanation. Hadley brings in Red, but Red insists he knows nothing of Andy's plans. Becoming increasingly hostile and paranoid, Norton starts throwing Andy's sculpted rocks around the cell. When he throws one at Andy's poster of Raquel Welch (in the spot previously occupied by Marilyn Monroe, and

before that by Rita Hayworth), the rock punches through and into the wall. Norton tears the poster from the wall revealing a tunnel just wide enough for a man to crawl into.It is revealed in a series of flashback sequences narrated by Red that many years ago, not long after receiving his rock hammer, Andy innocently tried to carve his name on his cell wall when a chunk of it came off. Andy, being a fan of geology, realized that the material the wall was made off of could make it possible for him to dig a hole in case he ever needed to escape. Andy first ordered the giant poster of Rita Hayworth to hide the hole. He then spent years digging at night with his rock hammer and hiding the dirt from his job into his pockets which he would then empty in the courtyard during his morning walks. When Tommy was killed, Andy decided it was time to go.During the previous night's thunderstorm, Andy wore Norton's clothes underneath his own to his cell, catching a lucky break when no one notices Norton's shiny black shoes on his feet, including Red. He packed many of his belongings, some papers and Norton's clothes into a plastic bag which he tied to himself with the rope he'd asked for, and escaped through his hole. The tunnel he'd excavated led him to a space between two walls of the prison where he found a sewer main line. Using a rock, he hit the sewer line in time several times with the lightning strikes and eventually broke it open. After crawling through 500 yards of the raw sewage contained in the pipe, Andy emerged in a brook outside the walls. A search team later found his prison clothes, a bar of soap and a very worn out rock hammer.While the warden and Red are discovering Andy's genius escape, Andy walks into the Maine National Bank in Portland, where he had put Norton's money. Using his assumed identity as Randall Stephens, and with all the necessary documentation, he closes the account and walks out with a cashier's check. Before he leaves, he asks them to drop a package in the mail. He continues his visitations to nearly a dozen other local banks, ending up with some $370,000. The package contains Warden Norton's accounting books, which are delivered straight to the Portland Daily Bugle newspaper along with Andy's written confessions and testimony.Not long after, the Maine state police storm Shawshank Prison along with several reporters to cover the developing story. Hadley is arrested for the murder of Tommy and is taken away by the state police. According to Red, he heard unfounded rumors through the grapevine that Hadley allegedly started "crying like a little girl" in the back seat of the police squad car while he was being taken away. Seeing Hadley being taken away in a police squad car and the local district attorney entering the prison with several policemen holding a warrant for Norton's arrest, Warden Norton finally opens his safe in his office, which he hadn't touched since Andy escaped, and instead of his books, he finds the Bible he had given Andy with a note to the warden saying that he was right, "salvation did lay within". Norton then opens it to the book of Exodus and finds that the pages had all been cut out in the shape of Andy's rock hammer. Norton walks back to his desk as the police pound on his door, takes out a small revolver and commits suicide by shooting himself in the head. Red remarks that he wondered if the warden thought, right before pulling the trigger, how "Andy could ever have gotten the best of him."Shortly after, Red receives a postcard from Fort Hancock, Texas, with nothing written on it. Red takes it as a sign that Andy made it into Mexico to freedom. Red and his

```
buddies kill time talking about Andy's exploits (with a few embellishments), but
Red falls into a sort of depression from missing his friend.At Red's next parole
hearing in 1967, he talks to the parole board about how "rehabilitated" is just
a made-up word invented to justify their job. He then explains how much he
regrets his actions of the past, not because he's in jail but because he knows
how wrong it was. He then closes by saying that he has to live with that for the
rest of his life and ask the board to stop wasting his time and leave him alone.
His parole is finally granted. He goes to live and work at the same places that
Brooks did, even seeing Brooks 'message carved into the wooden beam. He
frequently walks by a pawn shop which has several guns in the window. At times
he contemplates trying to get back into prison feeling that he has no life
outside of prison where he has spent most of his adult life, but he remembers
the promise he made to Andy. He then reveals that he was not looking at the guns
but at the compasses behind the guns and he bought one.Red follows Andy's
instructions, hitchhiking to Buxton and finding the stone wall Andy described.
Just as Andy said, there is a large black stone. Underneath is a small box
containing a large sum of cash and instructions to come find him in Zihuatanejo
although he doesn't name the city just in case. He also says he needs somebody
"who can get things" for a "project" of his. Red suddenly understands all the
power of hope and feels exhilarated by the feelings inside of him.After carving
a new message in the wooden beam which reads: "Brooks was here, so was Red", Red
violates parole and leaves the halfway house, unconcerned since no one is likely
to do an extensive manhunt for "an old crook like [him]." Red takes a bus to
Fort Hancock, where he crosses into Mexico. The two friends are finally reunited
on a beach of the Pacific coast, just like Andy had been hoping for.
```

1.4 Part-Of-Speech (POS)-Tagging

Um nun spezifischere Informationen zum Inhalt der Zusammenfassung zu erhalten, wenden wir ein POS-Tagging als Bestandteil der Vorverarbeitung der Daten auf unsere zuvor zerlegten Token-Objekte an. Während des Zerlegens der Zusammenfassung in Token-Objekte, führt spaCy bereits ein POS-Tagging durch. Dafür trifft ein *Convolutional Neural Network* (CNN) Vorhersagen über den Inhalt der Token. Um die Korrektheit des POS-Taggings von spaCy zu validieren, geben wir für jedes Token in unserem Dokument verschiedene Merkmale aus. Wir nutzen hier Pythons List Comprehension und übergeben dessen Ergebnis einem neuen DataFrame.

Merkmale:

- Text: Unverändertes Token
- Lemma: Grundform
- POS: Universal POS tag
- Shape: Form des Wortes (Groß-, Kleinschreibung, Satzzeichen, Zahl)
- Alpha: Ist Alpha Charakter?
- Stop: Stopwort?
- Entity: Beschreibung, falls Eigenname

[7]:

```
pos = [{'Text': token.text, 'Lemma': token.lemma_, 'POS': token.pos_, 'Shape':␣
 ↪token.shape_, 'Alpha': token.is_alpha, 'Stop': token.is_stop, 'Entity':␣
 ↪token.ent_type_} for token in document]
tokens = pd.DataFrame(pos)
tokens
```

[7]:

	Text	Lemma	POS	Shape	Alpha	Stop	Entity
0	In	in	ADP	Xx	True	True	
1	1947	1947	NUM	dddd	False	False	DATE
2	,	,	PUNCT	,	False	False	
3	Andy	Andy	PROPN	Xxxx	True	False	PERSON
4	Dufresne	Dufresne	PROPN	Xxxxx	True	False	PERSON
…	…	…	…	…	…	…	…
4139	had	have	AUX	xxx	True	True	
4140	been	be	AUX	xxxx	True	True	
4141	hoping	hope	VERB	xxxx	True	False	
4142	for	for	ADP	xxx	True	True	
4143	.	.	PUNCT	.	False	False	

[4144 rows x 7 columns]

Beispielhafte Validierung Wir nutzen hierfür das 18. Token in unserem Text. Es handelt sich um das Wort: *murdering*. spaCy detektiert korrekterweise die Grundform, das Lemma: *murder*. Basierend auf dessen Kontext erkennt spaCy, dass es sich um ein Verb handelt, kein Stoppewort und kein Eigenname.

[8]:
```
tokens.loc[18]
```

[8]:
```
Text      murdering
Lemma        murder
POS             VERB
Shape           xxxx
Alpha           True
Stop           False
Entity
Name: 18, dtype: object
```

1.5 Bag of Words und Worthäufigkeit

Zur Identifizierung inhaltlich wichtiger Informationen wird der Ansatz der *Bag of Words (BoW)* verwendet: Dabei wird mithilfe der Worthäufigkeit versucht, die Relevanz einzelner Tokens zu messen. Des weiteren wird die Tokenrelevanz in einer neuen, kompakten Struktur, bestehend aus einer Sammlung aller Wörter der Zusammenfassung, für die spätere Analyse gesichert. Relevantere Token liefern die charakteristischen Informationen über den Inhalt des jew. Films. Man muss jedoch beachten, dass es bestimmte Token gibt, die keine relevanten Informationen liefern. Dazu später mehr.

Wir erstellen zunächst ein eine neue Liste, welche alle Token der Zusammenfassung beinhaltet. Das Modul `Counter` bietet eine Schnittstelle zum Zählen der Häufigkeit aller Token.

[9]:
```
# Alle Token in Liste speichern
words = [token.text for token in document]
# Häufigkeit für jedes Token berechnen
counters = Counter(words)
```

Wir erstellen ein neues DataFrame basierend auf diesem Bag of Words und sortieren die Häufigkeiten in absteigender Reihenfolge.

[10]:
```
# Neues DataFrame erstellen
frequency = pd.DataFrame(columns=['Frequency'])
# Häufigkeit der Token zuweisen
frequency['Frequency'] = pd.Series(counters)
# Absteigend sortieren
frequency = frequency.sort_values(by=['Frequency'], ascending=False)
```

[11]:
```
frequency
```

[11]:

	Frequency
the	193
.	174
,	147
to	119
Andy	110
…	…
weeks	1
stretch	1
easiest	1
stuck	1
hoping	1

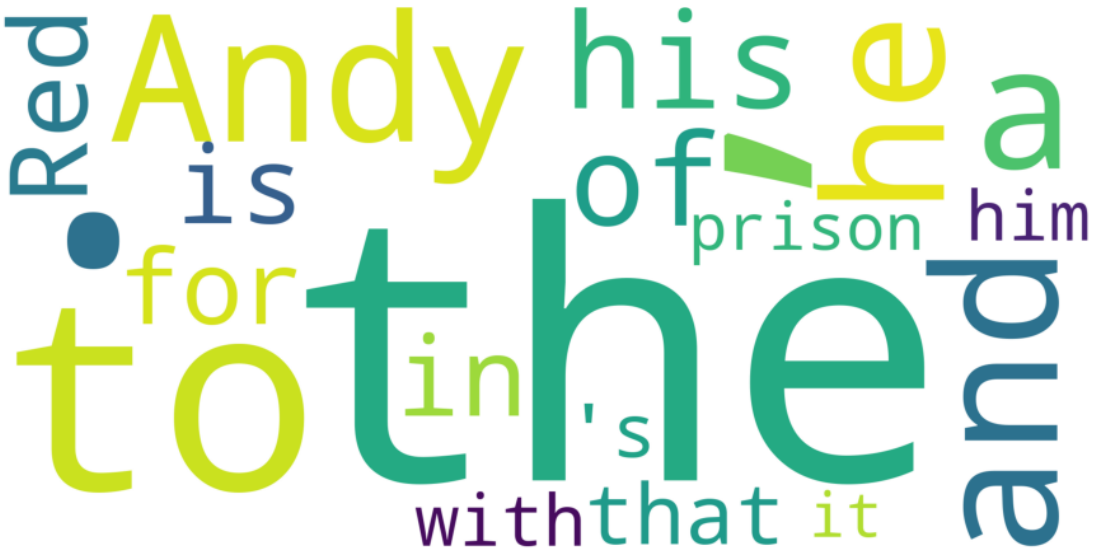
[1161 rows x 1 columns]

Im nächsten Schritt werden die relevantesten Tokens der Zusammenfassung untersucht, indem ein Dictionary mit den 20 am häufigsten verwendeten Token erstellt wird. Das Modul Counter bietet hierfür die Methode `most_common()` an. Mithilfe von `Matplotlib` und `WordCloud` können diese in einer geeigneten Form visualisiert werden.

[12]:
```
common_tokens = dict(Counter(words).most_common(20))
```

[13]:
```
word_cloud = WordCloud(background_color="white", width=2000, height=1000,).
 ↪generate_from_frequencies(common_tokens)
plt.rcParams["figure.figsize"] = (16, 8)
plt.axis("off")
plt.imshow(word_cloud)
```

[13]: <matplotlib.image.AxesImage at 0x7f54beb54d60>



Man kann erkennen, dass es sich bei den 20 am häufigsten verwendeten Token hauptsächlich um Satzzeichen, Stoppwörter und Eigennamen handelt. Da es sich bei Stopwörtern und Satzzeichen jedoch um inhaltlich irrelevante Informationen handelt, erhalten wir nur wenige Informationen über den Inhalt des Films. Eine Reduzierung der Token durch eine Vorverarbeitung der Daten, wird benötigt.

Hierfür wird analog wie zuvor verfahren und eine Liste aller Tokens unseres Textes erstellt, wobei dieses Mal jedoch alle Stoppwörter und Satzzeichen ignoriert werden.

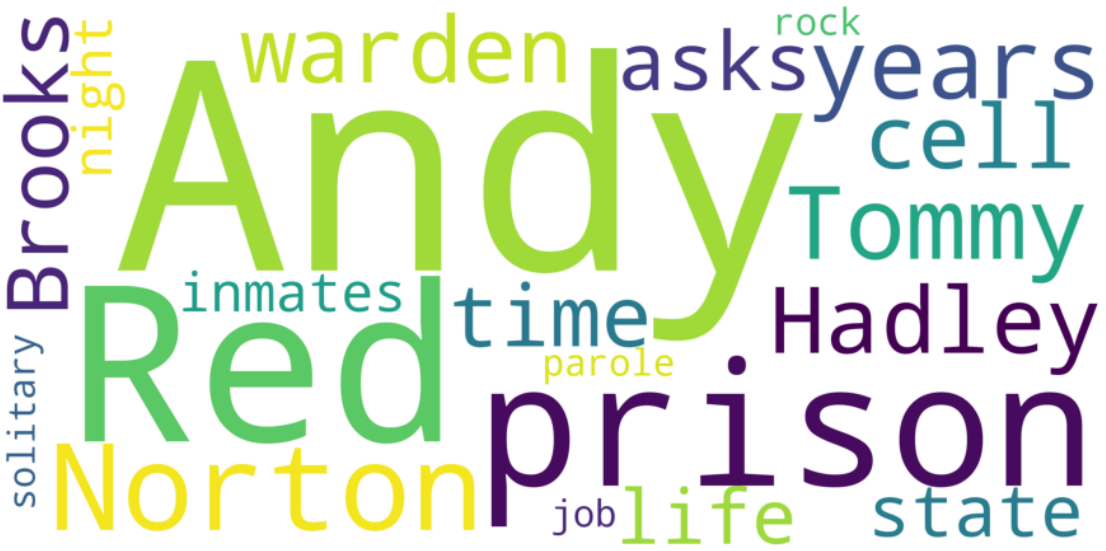
[14]:
```
# Alle Wörter (keine Satzzeichen) in Liste speichern
words = [token.text for token in document if not token.is_stop and not token.
 ↪is_punct]
# Worthäufigkeit berechnen
counters = Counter(words)
```

[15]:
```
# Neues DataFrame erstellen
frequency = pd.DataFrame(columns=['Frequency'])
# Häufigkeit der Token zuweisen
frequency['Frequency'] = pd.Series(counters)
```

[16]:
```
common_words = dict(Counter(words).most_common(20))
```

[17]:
```
word_cloud = WordCloud(background_color="white", width=2000, height=1000,).
 ↪generate_from_frequencies(common_words)
plt.rcParams["figure.figsize"] = (16, 8)
plt.axis("off")
plt.imshow(word_cloud)
```

[17]: <matplotlib.image.AxesImage at 0x7f54b7fc37f0>



Man kann erkennen, dass es sich bei den nun relevantesten Wörtern um deutlich inhaltsbezogenere Informationen handelt. Generell sind nun Eigennnamen sehr präsent. Diese erlauben zwar Rückschlüsse auf Figurenmerkmale, Schauspieler, sowie Beziehungen der Figuren, liefern jedoch wenig Rückschlüsse über die generelle Handlung des Films. Ebenso sind Verben weniger relevant. Beispielsweise wird die Grundform von *asks* in wahrscheinlich fast allen Filmen vorkommen. Neben Eigennamen werden nun ebenfalls alle Verben entfernt, da diese, gerade in lemmatisierter Form, zu häufig in allen Filmzusammenfassungen vorkommen werden und somit nur ein generelles Clustern der Daten erlauben.

[18]:
```
# Eigennamen identifizieren
entities = sorted(set([entity.text for entity in document.ents]))
# Token auf Wörter ohne Eigennamen, Stoppwörter, Satzzeichen, Verben reduzieren
words = [str(token.lemma_).lower() for token in document if not token.ent_type_␣
 ↪and not token.is_stop and not token.is_punct and token.pos_ != 'VERB']
# Worthäufigkeit berechnen
counters = Counter(words)
```

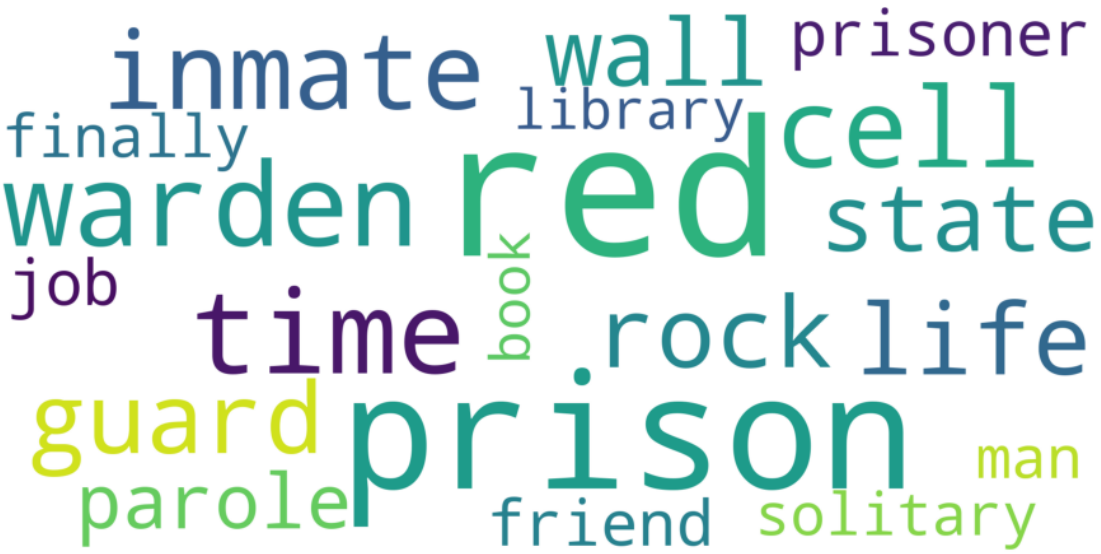
[19]:
```
# Neues DataFrame erstellen
frequency = pd.DataFrame(columns=['Frequency'])
# Häufigkeit der Token zuweisen
frequency['Frequency'] = pd.Series(counters)
```

[20]:
```
common_words = dict(Counter(words).most_common(20))
```

[21]:
```
word_cloud = WordCloud(background_color="white", width=2000, height=1000,).
 ↪generate_from_frequencies(common_words)
```

```
plt.rcParams["figure.figsize"] = (16, 8)
plt.axis("off")
plt.imshow(word_cloud)
```

[21]: <matplotlib.image.AxesImage at 0x7f54b62d7c40>



Basierend auf diesen reduzierten Daten, welche zum Großteil aus Substantiven und Adjektiven bestehen, sollte eine inhaltliche Analyse möglich sein. Wir können beispielsweise erkennen, dass es sich bei dem Film *Die Verurteilten*, um einen Film handelt, der sich mit einem Männergefängnis, Insassen, Wächtern und Einzelhaft beschäftigt. Fälschlicherweise erkennt spaCy *Red* des Öfteren als Adjektiv und nicht als Spitznamen einer Rolle. Betrachtet man die Merkmal aller Token, dessen Text *Red* ist, so kann man erkennen, dass spaCy erstaunlicherweise *Red* dennoch als Person erkennt.

[22]:
```
tokens.loc[tokens["Text"] == "Red"]
```

[22]:

	Text	Lemma	POS	Shape	Alpha	Stop	Entity
87	Red	Red	PROPN	Xxx	True	False	
110	Red	Red	PROPN	Xxx	True	False	ORG
124	Red	Red	PROPN	Xxx	True	False	
153	Red	Red	PROPN	Xxx	True	False	
175	Red	red	ADJ	Xxx	True	False	
311	Red	Red	PROPN	Xxx	True	False	
323	Red	Red	PROPN	Xxx	True	False	
346	Red	Red	PROPN	Xxx	True	False	
359	Red	Red	PROPN	Xxx	True	False	
416	Red	Red	PROPN	Xxx	True	False	
429	Red	Red	PROPN	Xxx	True	False	
553	Red	Red	PROPN	Xxx	True	False	

```
686   Red   red    ADJ   Xxx   True  False
731   Red   Red  PROPN   Xxx   True  False
744   Red   Red  PROPN   Xxx   True  False
901   Red   Red  PROPN   Xxx   True  False
1169  Red   red    ADJ   Xxx   True  False
1212  Red   red    ADJ   Xxx   True  False
1421  Red   red    ADJ   Xxx   True  False
1561  Red   red    ADJ   Xxx   True  False
1601  Red   Red  PROPN   Xxx   True  False
1611  Red   Red  PROPN   Xxx   True  False     ORG
1664  Red   red    ADJ   Xxx   True  False
1929  Red   Red  PROPN   Xxx   True  False
1991  Red   Red  PROPN   Xxx   True  False  PERSON
2098  Red   Red  PROPN   Xxx   True  False  PERSON
2104  Red   Red  PROPN   Xxx   True  False
2122  Red   Red  PROPN   Xxx   True  False
2527  Red   Red  PROPN   Xxx   True  False
2607  Red   Red  PROPN   Xxx   True  False
2615  Red   Red  PROPN   Xxx   True  False     ORG
2662  Red   Red  PROPN   Xxx   True  False
2675  Red   Red  PROPN   Xxx   True  False     ORG
2700  Red   Red  PROPN   Xxx   True  False
2735  Red   red    ADJ   Xxx   True  False
2768  Red   Red  PROPN   Xxx   True  False
2827  Red   red    ADJ   Xxx   True  False
2935  Red   Red  PROPN   Xxx   True  False     GPE
2938  Red   Red  PROPN   Xxx   True  False
3034  Red   Red  PROPN   Xxx   True  False
3201  Red   Red  PROPN   Xxx   True  False     GPE
3334  Red   Red  PROPN   Xxx   True  False
3490  Red   Red  PROPN   Xxx   True  False
3664  Red   red    ADJ   Xxx   True  False
3695  Red   Red  PROPN   Xxx   True  False
3711  Red   red    ADJ   Xxx   True  False
3726  Red   Red  PROPN   Xxx   True  False
3745  Red   Red  PROPN   Xxx   True  False
3758  Red   Red  PROPN   Xxx   True  False     ORG
3958  Red   Red  PROPN   Xxx   True  False
4037  Red   Red  PROPN   Xxx   True  False
4074  Red   Red  PROPN   Xxx   True  False
4077  Red   Red  PROPN   Xxx   True  False
4108  Red   Red  PROPN   Xxx   True  False
```

1.6 Eigennamen

spaCy verfügt des Weiteren über Schnittstellen zur Identifizierung der Typen von Eigennamen. Diese sollen zunächst nicht im Fokus der kommenden Ausführungen liegen und könnten lediglich in

einer späteren Optimierung hinzugezogen werden. Mit der integrierten `render()` Methode lassen sich Eigennamen im Text hervorheben. Dessen Typen werden hierbei ebenfalls dargestellt.

[23]:
```
spacy.displacy.render(document, style="ent", jupyter=True)
```

```
<IPython.core.display.HTML object>
```

Man könnte somit auch beispielsweise Personen, Orte und Zeiten identifizieren.

1.6.1 Personen

Beinhalten Schauspieler, Rollen, Spitznamen.

[24]:
```
set([token.text for token in document.ents if token.label_ == 'PERSON'])
```

[24]:
```
{'Andy',
 'Andy Dufresne',
 'Bible',
 'Bob Gunton',
 'Boggs',
 'Brooks',
 'Brooks Hatlen',
 'Byron Hadley',
 'Clancy Brown',
 'Ellis Boyd Redding',
 'Gil Bellows',
 'Hadley',
 'Heywood',
 'James Whitmore',
 'Marilyn Monroe',
 'Mark Rolston',
 'Mozart',
 'Randall Stephens',
 'Raquel Welch',
 'Red',
 'Rita Hayworth',
 'Shawshank Prison',
 'Tim Robbins',
 'Tommy',
 'Warden Norton',
 "Warden Norton's",
 'Warden Samuel Norton',
 'William Sadler',
 'Zihuatanejo',
 'hammer'}
```

1.6.2 Orte

Ermöglichen Rückschlüsse auf Schauplätze und Drehorte.

[25]:
```
set([token.text for token in document.ents if token.label_ == 'GPE'])
```

[25]:
```
{'Buxton',
 'Fort Hancock',
 'Haywood',
 'Maine',
 'Mexico',
 'Portland',
 'Red',
 'Texas',
 'Zihuatanejo'}
```

1.6.3 Zeiten

Ermöglichen beispielsweise Rückschlüsse in welchem Jahrzehnt die Geschichte spielt oder welche Zeitintervalle vorkommen.

[26]:
```
set([token.text for token in document.ents if token.label_ == 'DATE'])
```

[26]:
```
{'10 years',
 '13 years old',
 '1947',
 '1965',
 '1967',
 '20 years',
 '30 years',
 '50 years',
 'About a month later',
 'About four years',
 'a month',
 'a week',
 'almost 20 years',
 'another month',
 'every week',
 "lays within''",
 'ten years earlier',
 'the first two years',
 'the years',
 'two weeks',
 'years'}
```

1.7 Fazit

Eine Validierung der Zusammenfassung des Films “Die Verurteilten” mit Python, Pandas und spaCy ermöglicht eine Reduzierung der Textdaten auf inhaltlich relevante Informationen, welche

in einer weiteren Analyse der Daten untersucht werden können. Im folgenden Kapitel soll nun eine Verarbeitung aller Dateneinträge angewendet werden, sowie eine Analyse genereller Information erfolgen.

- Weiter zu: Verarbeitung aller Dateneinträge
- Zurück zur Übersicht

[ ]: