Download Center Doku

Die Dokumentation zum Arbeiten mit den Datensätzen aus dem Download Center der Bundestags-Mine.

~~**Stand vom 28.03.2023**~~

Update: Stand vom 29.10.2024

info@bundestag-mine.de

Der Datensatz

Genereller Aufbau

Jedes Protokoll ist in seiner eigenen JSON-Datei gekapselt. Diese JSON-Datei beinhaltet sowohl Protokoll als auch die dazugehörigen Reden und Auswertungsdaten. Nachdem die ZIP-Datei entpackt wurde, sollten sich, je nach Filter, mehrere JSON-Dateien nach dem Format

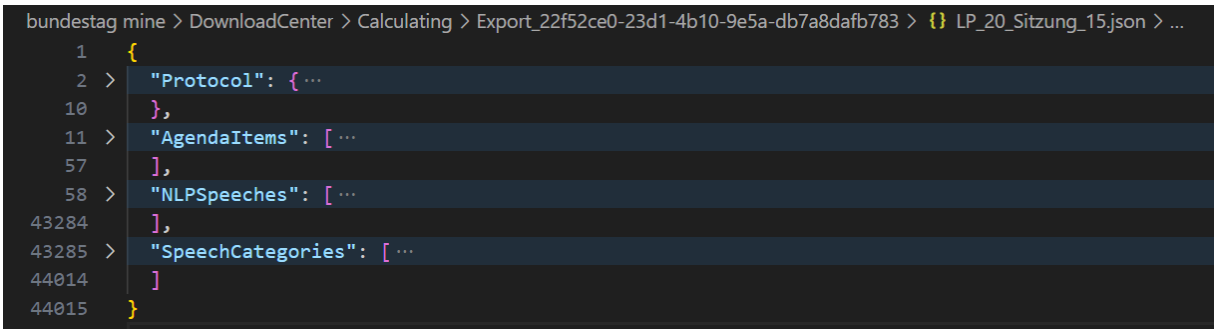
LP_{Legislaturperiode}_Sitzung_{Sitzungsnummer}.json

im Ordner befinden. Der Name der JSON-Datei bestimmt somit die genaue Sitzung des Protokolls.

Die JSON-Datei

Die JSON-Dateien haben alle den gleichen Aufbau, sodass diese programmatisch gleichbehandelt werden können. Es empfiehlt sich einen geeigneten Text-Editor zu nutzen, um die Dateien zu öffnen und das Format zu erkennen. Im Folgenden wird Visual Studio Code genutzt.

Eine Datei ist in 4 Objekten eingeteilt, die jeweils weitere Verschachtelungen besitzen:



Im Folgenden werden die Objekte erläutert.

Protocol

Eigenschaft	Erläuterung
Date	Das Datum, an dem die Sitzung stattfand (Ohne Uhrzeit)
LegislaturePeriod	Die Legislaturperiode, an dem die Sitzung stattfand.
Number	Die Nummer der Sitzung wie in „Die 5. Sitzung der 20. Legislaturperiode“
Title	Der Titel der Sitzung wie „5. Sitzung“
AgendaItemsCount	Die Anzahl der Tagesordnungspunkte der Sitzung. (Die Anzahl stammt aus den XML-Protokollen und weicht manchmal von der auf der Seite des Bundestags notierten Anzahl ab).
Mongold	-
Id	Die einzigartige Id des Protokolls (Alternativ kann auch die Legislaturperiode + Sitzungsnummer genutzt werden)

info@bundestag-mine.de

AgendaItems[]

Hierbei handelt es sich um eine **Liste** von mehreren Tagesordnungspunkten (Im Folgenden: „TOP“).

Eigenschaft	Erläuterung
Title	Der Titel des TOP.
Description	Eine Beschreibung des TOP in HTML, mit möglichen Hyperlinks zu dazugehörigen Drucksachen.
AgendaItemNumber	Die Bezeichnung des TOP wie in „1. Tagesordnungspunkt“ aber auch „ZP 1“ für „Zusatzpunkt 1“. Es ist nicht die Reihenfolge, sondern stammt nur aus der Seite des Bundestags.
Order	Die Reihenfolge der TOP wie in 1., 2., 3. Tagesordnungspunkt.
Date	Das Datum des TOP mit Uhrzeit
ProtocolId	Die Id des Protokolls, zu dem der TOP gehört.
Id	Die einzigartige Id des TOP.

NLPSpeeches[]

Hierbei handelt es sich um eine **Liste** von mehreren Reden samt deren NLP-Auswertungen

Eigenschaft	Erläuterung
CategoryCoveredTags[]	Immer „null“, da grade nicht nutzbar.
Tokens[]	Immer „null“, da zu große Datenmengen.
Text	Der gesamte Text der Rede ohne Zwischenrufe.
SpeakerId	Die einzigartige Id des/r Redners/in
ProtocolNumber	Die Nummer der Sitzung zu der diese Rede gehört wie in „Rede der 3. Sitzung“
LegislaturePeriod	Die Legislaturperiode, in der diese Rede gehalten wurde. Die Periode samt der ProtocolNumber kann die Sitzung genau identifizieren.
AgendaItemNumber	Die Nummer des TOP zu dem diese Rede gehört. **Achtung:** Es handelt sich hierbei um die **Order** des TOP, nicht die gleichnamige „**AgendaItemNumber**“. Hier sind die Namen leider misslich gewählt worden.
MongoId	-
Id	Die einzigartige Id der Rede.
AbstractSummary	Eine abstrakte Zusammenfassung der Rede generiert durch BART.
AbstractSummaryPEGASUS	Eine abstrakte Zusammenfassung der Rede generiert durch PEGASUS.
ExtractiveSummary	Eine extrahierende Zusammenfassung der Rede generiert durch TextRank.
EnglishTranslationOfSpeech	Die Übersetzung der Rede ins Englische mithilfe von OPUS-MT.
EnglishTranslationScore	Ein Wert von 0-1, der bestimmt, wie gut die Übersetzung von Deutsch nach Englisch gelang mithilfe von LaBSE.

Für mehr Details zu den automatisch erstellen Zusammenfassungen verweise ich auf das Reasearch-Center der Bundestags-Mine. Es kann sein, dass nicht jede Rede durch die Summarization Pipeline gegangen ist, weshalb Eigenschaften wie „AbstractSummary“ null sein können. In diesem Fall sollten die Zusammenfassungen später vorhanden sein. Ansonsten bitte ich um Kontaktaufnahme. Die weiteren Eigenschaften NamedEntities[], Sentiments[] und Segments[] werden im Folgenden einzeln erklärt.

NLPSpeech.NamedEntities[]

Hierbei handelt es sich um eine **Liste** von mehreren NamedEntities. Eine Named-Entity (im Folgenden: „NE“) ist eine Klassifikation von Eigennamen. Ein Eigenname ist eine Folge von Wörtern, die eine real existierende Entität beschreibt, wie z. B. ein Firmenname. Jede Named-Entity wird in ORG (Organisation), PER (Person), LOC (Ort) oder MISC (Sonstiges) eingeordnet.

Eigenschaft	Erläuterung
LemmaValue	Der Wortlaut wie „Minister“ oder „VW-Chef“
NLPSpeechId	Die Id der Rede zu der diese NE gehört.
ShoutId	Wenn diese Id einer leeren GUID gleicht (00000000-0000-0000-0000-000000000000), wurde die NE im Text der Rede erwähnt. Ist die ShoutId eine nicht-leere GUID, dann wurde die NE in einem Zwischenruf der Rede erwähnt. In diesem Fall kann man die NE über die ShoutId dem genauen Zwischenruf zuordnen.
Begin	Der Index, an welchem der LemmaValue der NE im Text beginnt. Bsp.: „Der VW-Chef hat zu lange…“ hätte einen „Begin“ von **4**
End	Der Index, an welchem der LemmaValue der NE im Text endet. Bsp.: „Der VW-Chef hat zu lange…“ hätte einen „End“ von **11**
Value	Die Kategorie der NE. Entweder ORG, LOC, PER oder MISC (Siehe oben)
Id	Die einzigartige Id der NE.

NLPSpeech.Sentiments[]

Hierbei handelt es sich um eine **Liste** von mehreren Sentiments. Ein Sentiment ist eine Dezimal-Zahl zwischen -1 und 1. Der Sentiment wird auf Satz-Basis berechnet und bestimmt demnach immer den Sentiment-Wert ganzer Sätze. Dabei gilt:

$$-1 \geq Sentiment < 0 \text{ Negativ}$$
$$Sentiment = 0 \text{ Neutral}$$
$$0 > Sentiment \leq 1 \text{ Positiv}$$

Eigenschaft	Erläuterung
NLPSpeechId	Die Id der Rede zu der dieser Sentiment gehört.
ShoutId	Wenn diese Id einer leeren GUID gleicht (00000000-0000-0000-0000-000000000000), wurde der Sentiment im Text der Rede berechnet. Ist die ShoutId eine nicht-leere GUID, dann wurde der

info@bundestag-mine.de

	Sentiment in einem Zwischenruf der Rede berechnet. In diesem Fall kann man den Sentiment über die ShoutId dem genauen Zwischenruf zuordnen.
Begin	Der Index, an welchem der Satz beginnt, der durch diesen Sentiment bestimmt wird. Bsp.: „Uns geht es nicht gut! Und wir machen zu wenig!“ hätte einen „Begin“ von **0**
End	Der Index, an welchem der Satz endet, der durch diesen Sentiment bestimmt wird. Bsp.: „Uns geht es nicht gut! Und wir machen zu wenig!“ hätte einen „End“ von **22**
SentimentSingleScore	Der Sentiment-Wert, der mit der obigen Skala zu interpretieren ist. Bsp: „Die Industrie scheint hier also deutlich weiter zu sein als die Bundesregierung, weil sich die Bundesregierung ihrerseits vom Klimaziel 2020 verabschiedet, was wir für ein fatales Signal halten“ Hätte einen **SSS** von **-0.3818** und wäre demnach **negativ** gestimmt.
Id	Die eindeutige Id des Sentiments

NLPSpeech.Segments[]

Hierbei handelt es sich um eine **Liste** von mehreren Segmenten. Ein Segment ist ein Stück der Rede, der mit einem Absatz oder Zwischenruf beendet wird. Der Text aller Segments ist vereint der Text der Rede. Die Struktur wird benötigt, um die Zwischenruf an die richtige Stellen der Rede zu platzieren.

Eigenschaft	Erläuterung
Text	Der Text dieses Segments/Absatzes.
SpeechId	Die Id der Rede zu welcher dieses Segment gehört.
Id	Die einzigartige Id dieses Segments.

NLPSpeech.Segment.Shouts[]

Hierbei handelt es sich um eine **Liste** von mehreren Zwischenrufen. Die Zwischenrufe beenden ein Segment automatisch und finden somit am Ende des Textes des dazugehörigen Segments statt.

Eigenschaft	Erläuterung
Text	Der Text/Inhalt des Zwischenrufs
SpeechSegmentId	Die Id des Segments zu welchem dieser Zwischenruf gehört.
FirstName	**Falls vorhanden:** Der Vorname des/r Rufenden
LastName	**Falls vorhanden:** Der Nachname des/r Rufenden
Fraction	**Falls vorhanden:** Die Fraktion des/r Rufenden
Party	**Falls vorhanden:** Die Partei des/r Rufenden
SpeakerId	**Falls vorhanden:** Die Id des/r Rufenden

Es ist wichtig zu wissen, dass die Stenographen nicht immer identifizieren können, wer grade rein ruft, weshalb nicht immer alle Informationen vorhanden sind.

SpeechCategories

Die Bundestags-Mine unterteilt jede Rede in sogenannte „Categories“ (Kategorien). Diese wurden durch **VecTop** erzeugt und den Reden zugeteilt. VecTop ist ein Topic Search Verfahren, dass unbekannte Texte in festgelegte Kategorien einordnet. Diese orientieren sich an den Kategorien, die man üblicherweise in Zeitungen finden (wie „Wirtschaft“, „Ausland“, „Deutschland“, etc.). VecTop gliedert sowohl in Kategorien als auch Unter-Kategorien ein.

Eigenschaft	Erläuterung
Name	Der Name der Kategorie, wie zB. „Wirtschaft“
SubCategory	Der Name der Sub-Kategorie dieser Category, wie zB. „Unternehmen“
NLPSpeechId	Die eindeutige Id, die diese Category zu einer NLPSpeech zuordnet.
Created	Das Datum, an dem diese Category erfasst wurde.
Id	Die eindeutige Id dieses Category Objektes.