WaPo

Readme:

ieses Archiv enthält Version 4 der TREC Washington Post Collection.

Diese Version enthält Artikel, die im Jahr 2020 veröffentlicht wurden. Das Format wurde dahingehend aktualisiert, dass der Inhaltsblock mit dem Typ "date" jetzt ein text/plain-Block mit einer ISO-formatierten Datumszeichenfolge ist. Das Feld "published_date" bleibt ein Millisekunden-Zeitstempel aus der Unix-Epoche (1/1/1970 00:00Z). Das für die Konvertierung verwendete Skript ist im Verzeichnis scripts/ enthalten.

Die Post änderte ihr JSON-Schema (erneut) im Jahr 2020; die Dokumente in dieser Sammlung wurden aus Gründen der Konsistenz auf das ursprüngliche Schema portiert (modulo das Datumsfeld, wie erwähnt). Eine Besonderheit ist, dass bei mehreren tausend neuen Dokumenten URLs als Bezeichner wiederholt wurden. Diesen Artikeln wurden neue UUID4-Bezeichner zugewiesen.

Ein weiterer Unterschied im Jahr 2020 besteht darin, dass es offenbar "parallele" Versionen von Artikeln gibt, von denen eine die Webversion und die andere ein Teaser für einen Podcast zu dem Artikel ist, mit identischen Titeln, aber unterschiedlichen Autorenlisten. Beide Versionen des Artikels wurden in der Sammlung beibehalten.

Ein dritter Unterschied ist, dass die Artikel aus dem Jahr 2020 keine Multimedia-Links in den Artikeln enthalten. Ich glaube, dass die Post die Art und Weise, wie sie Artikelinhalte intern speichert, geändert hat und diese Links deshalb nicht mehr enthalten sind.

Beachten Sie, dass in Version 3 eine Reihe von nahezu doppelten Dokumenten entfernt wurde. Einzelheiten zu diesem Vorgang finden Sie in README-v3.md. Dies führte zu Inkompatibilitäten mit TREC-Sammlungen, die auf v2 aufgebaut sind, da die docids in den topics- und qrels-Dateien in v3 möglicherweise nicht mehr existieren. WAS SOLLTEN SIE DAMIT TUN.

Dieses Paket enthält:

1. die WashingtonPost-Sammlung, eine einzelne Datei, in der jede Zeile ein JSON-Objekt ist, das ein Dokument darstellt,

2. eine Liste von nahezu doppelten Dokumenten, die in v3 entfernt wurden,

3. einige Beispiel-Python-Skripte:

convert-date.py: Konvertiert den Inhaltsblock "Datum" in der v3-Sammlung in das in v4 verwendete Format.

tag-wapo.py: Schmückt WaPo-Dokumente mit Spacy-NER-Daten aus.

index-wapo-spacy.py: Indexiert WaPo-Dokumente (optional mit tag-wapo.py getaggt) in ElasticSearch. Getestet mit Elastic 7.10.2.

wapo-print-docids.py: gibt die Dokumentenbezeichner aus den Dokumenten aus.

wapo-near-duplicates: Dies ist die Liste der nahezu doppelten Dokumente, die in v3 entfernt wurden. Siehe README-v3.md für Details.

update-topics-qrels-v2-v3: Skripte von John Foley zur Konvertierung von TREC news track topics und qrels, die auf v1 und v2 der Sammlung basieren, in v3/v4 docids (nach der Reduplikation)

Die Skripte benötigen Python 3 und hängen von den pip-Paketen für spacy, elasticsearch und tqdm ab. Es gibt eine Datei requirements.txt im Verzeichnis scripts.

ReadmeV3:

ieses Archiv enthält Version 4 der TREC Washington Post Collection.

Diese Version enthält Artikel, die im Jahr 2020 veröffentlicht wurden. Das Format wurde dahingehend aktualisiert, dass der Inhaltsblock mit dem Typ "date" jetzt ein text/plain-Block mit einer ISO-formatierten Datumszeichenfolge ist. Das Feld "published_date" bleibt ein Millisekunden-Zeitstempel aus der Unix-Epoche (1/1/1970 00:00Z). Das für die Konvertierung verwendete Skript ist im Verzeichnis scripts/ enthalten.

Die Post änderte ihr JSON-Schema (erneut) im Jahr 2020; die Dokumente in dieser Sammlung wurden aus Gründen der Konsistenz auf das ursprüngliche Schema portiert (modulo das Datumsfeld, wie erwähnt). Eine Besonderheit ist, dass bei mehreren tausend neuen Dokumenten URLs als Bezeichner wiederholt wurden. Diesen Artikeln wurden neue UUID4-Bezeichner zugewiesen.

Ein weiterer Unterschied im Jahr 2020 besteht darin, dass es offenbar "parallele" Versionen von Artikeln gibt, von denen eine die Webversion und die andere ein Teaser für einen Podcast zu dem Artikel ist, mit identischen Titeln, aber unterschiedlichen Autorenlisten. Beide Versionen des Artikels wurden in der Sammlung beibehalten.

Ein dritter Unterschied ist, dass die Artikel aus dem Jahr 2020 keine Multimedia-Links in den Artikeln enthalten. Ich glaube, dass die Post die Art und Weise, wie sie Artikelinhalte intern speichert, geändert hat und diese Links deshalb nicht mehr enthalten sind.

Beachten Sie, dass in Version 3 eine Reihe von nahezu doppelten Dokumenten entfernt wurde. Einzelheiten zu diesem Vorgang finden Sie in README-v3.md. Dies führte zu Inkompatibilitäten mit

TREC-Sammlungen, die auf v2 aufgebaut sind, da die docids in den topics- und qrels-Dateien in v3 möglicherweise nicht mehr existieren. WAS SOLLTEN SIE DAMIT TUN.

Dieses Paket enthält:

1. die WashingtonPost-Sammlung, eine einzelne Datei, in der jede Zeile ein JSON-Objekt ist, das ein Dokument darstellt,

2. eine Liste von nahezu doppelten Dokumenten, die in v3 entfernt wurden,

3. einige Beispiel-Python-Skripte:

convert-date.py: Konvertiert den Inhaltsblock "Datum" in der v3-Sammlung in das in v4 verwendete Format.

tag-wapo.py: Schmückt WaPo-Dokumente mit Spacy-NER-Daten aus.

index-wapo-spacy.py: Indexiert WaPo-Dokumente (optional mit tag-wapo.py getaggt) in ElasticSearch. Getestet mit Elastic 7.10.2.

wapo-print-docids.py: gibt die Dokumentenbezeichner aus den Dokumenten aus.

wapo-near-duplicates: Dies ist die Liste der nahezu doppelten Dokumente, die in v3 entfernt wurden. Siehe README-v3.md für Details.

update-topics-qrels-v2-v3: Skripte von John Foley zur Konvertierung von TREC news track topics und qrels, die auf v1 und v2 der Sammlung basieren, in v3/v4 docids (nach der Reduplikation)

Die Skripte benötigen Python 3 und hängen von den pip-Paketen für spacy, elasticsearch und tqdm ab. Es gibt eine Datei requirements.txt im Verzeichnis scripts.