

Anreicherung von News-Artikeln durch automatisch generierte Inhalte

Studiengang: BSc in Informatik | Vertiefung: Data Engineering
Betreuer: Prof. Dr. Jürgen Vogel
Experte: Dr. Igor Metz (GLUE Software Engineering AG)

Können Artikel von News-Portalen mit automatisch generierten Inhalten angereichert werden, sodass der Leser einen Mehrwert hat? In dieser Arbeit wurden zwei Aspekte genauer untersucht: Die Generierung von Zusammenfassungen und von Fragen und Antworten zum Artikel.

Daten

Die Artikel für diese Arbeit stammen von News-Portal des Tages-Anzeigers. Täglich wurden die neuesten 50 Artikel abgerufen und in einer MongoDB gespeichert. Gegen Ende der Arbeit lagen über 4000 Artikel vor.

Zusammenfassungen

Drei Arten von Zusammenfassungen wurden untersucht:

- Extrahierende Zusammenfassungen, dabei werden die relevantesten Sätze 1:1 aus dem Originaltext übernommen
- Abstrahierende Zusammenfassungen, dabei werden die Sätze ganz neu geschrieben
- Keynotes, drei Sätze mit den wichtigsten Inhalten des Artikels

Für die extrahierenden Zusammenfassungen wurde ein eigener Algorithmus entwickelt und mit bestehenden Tools verglichen.

Abstrahierende Zusammenfassungen wurden mit einem Modell der OpenAI API generiert. Es wurden

folgende Zusammenfassungen erstellt und mit einer Metrik für die Lesbarkeit bewerten: Zusammenfassungen in normaler Sprache, in einfachen Worten und für Kinder. Die generierten Zusammenfassungen in normaler Sprache wurden mit von Menschen erstellten Zusammenfassungen verglichen. Dabei lieferte eine menschliche Bewertung sehr gute Ergebnisse für die automatisch erstellten Zusammenfassungen.

Keynotes wurden mit drei verschiedenen Modellen von OpenAI erstellt, wobei die Modelle unterschiedliches Verhalten zeigten.

Quiz (Frage/Antworten)

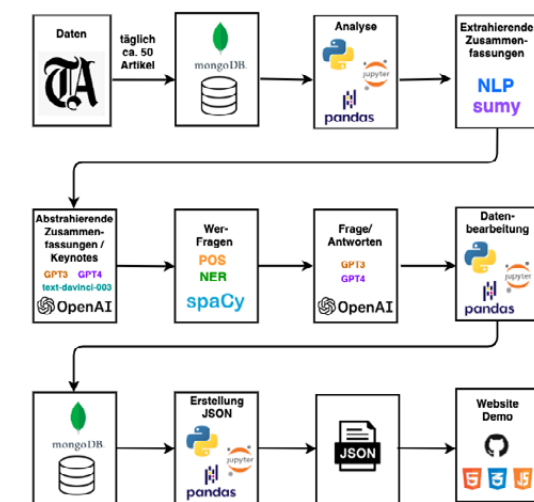
Aus dem Artikel sollten eine Frage und drei Antworten (eine richtig, zwei falsch) zum Inhalt generiert werden. Dies wurde mit SpaCy und zwei Modellen von OpenAI realisiert. Im Vergleich lieferte OpenAI deutlich bessere Ergebnisse, auch wenn die Fragen und Antworten im Detail überprüft werden müssen.

Webseite

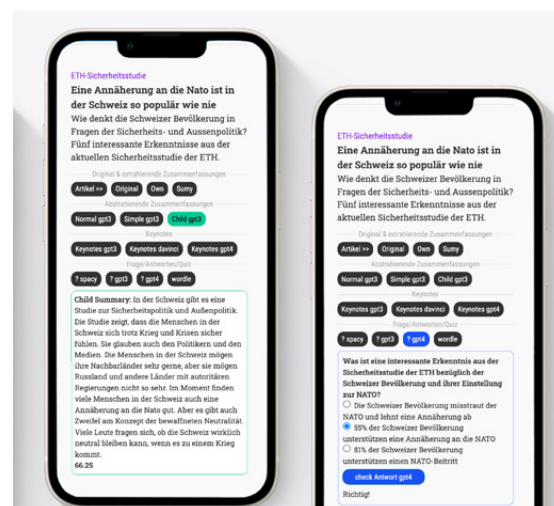
Zu Demonstrationszwecken wurde eine Website mit Artikeln und den dazu generierten Texten erstellt.



Nicole Herold
herold.nicole@gmx.ch



Daten-Pipeline



Zusammenfassung für Kinder und Frage mit Antworten