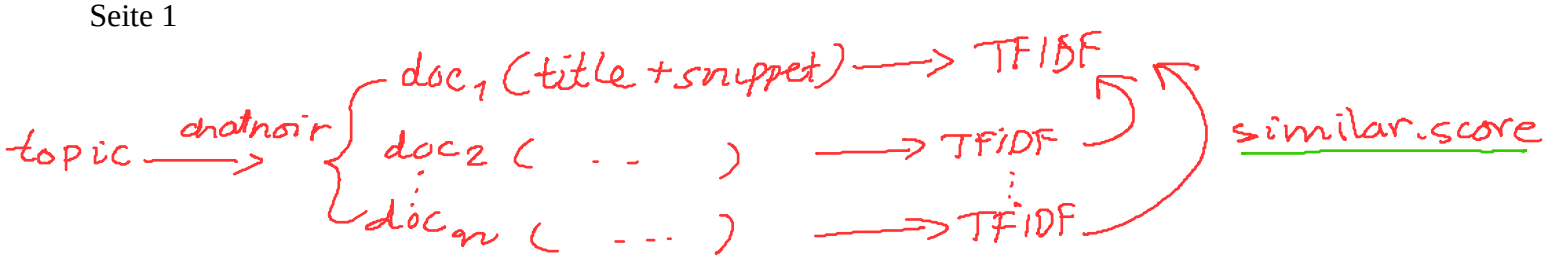
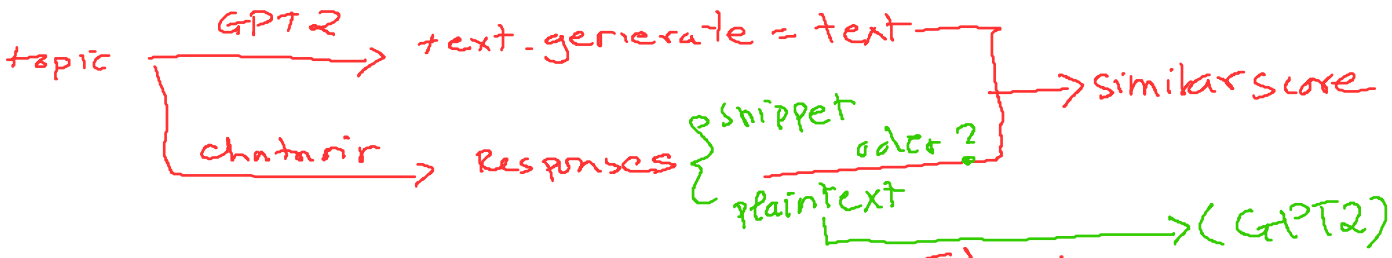
nächste Beobachtung:

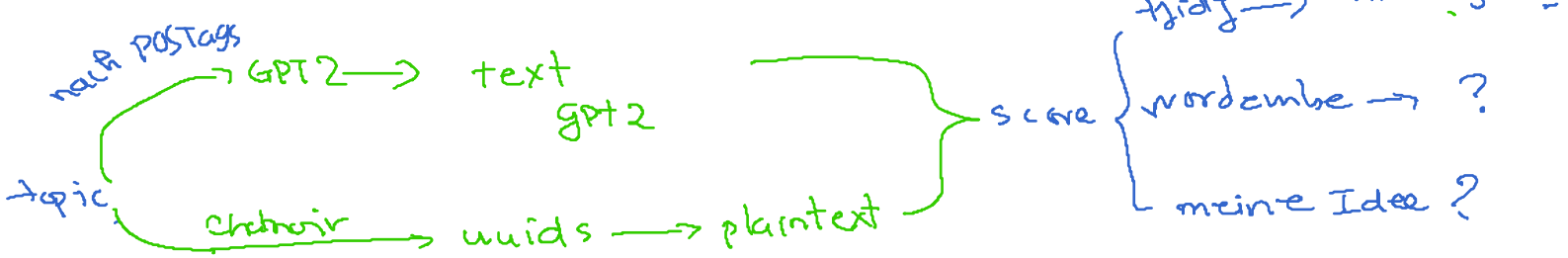


- similar (tfidf$_{doc1}$, tfidf$_{doc_i}$) $i \in [2, n]$
- similar 0,1



GPT2 text generate für wie viele Tokens

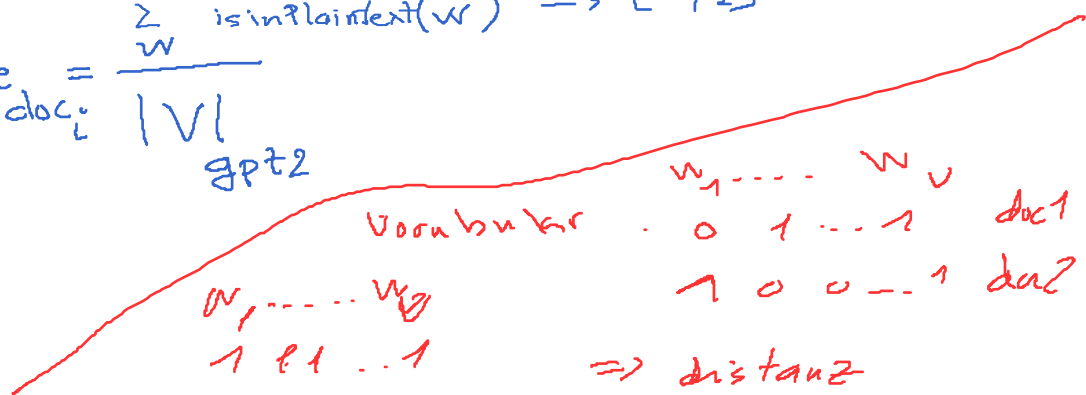
Modells: 124M, 355M, 1558 M



- text$_{gpt2}$ → preprocess → nur gute Wörter übrig → V_{gpt2} (Vocabular)
- Berechnen:

$$score_{doc_i} = \frac{\sum_{w}^{V_{gpt2}} isinPlaintext(w) \rightarrow [0,1]}{|V|_{gpt2}}$$

idf ?



Aktueller Stand (07.01.2021)

(1) Query Expansion:

- orginal — $w_1 \cdot score$
- POS Tags — $w_2 \cdot score$
- lemma, synonyms — $* w_3 \cdot score$

} best: { 1,75 ; 1,25 ; 1 = 0,585 } max

{ max / avg

(2) Target API

- orginal — responses / docs — target score → new = relscore * (1 + target score)

- target model { classify WD / classify WD_dep

- Best: { classify WD_dep / SW: 0,55 } => 0,582