بخش ۳- سؤال ۴-

stopword ها کلماتی هستند که شامل اطلاعات مهمی درباره متن نمی‌شوند و تعداد آن‌ها در متن زیاد است.

به عبارت بهتر: کلماتی هستند که یا ناچیز هستند یا تعداد آنقدر زیاد است که سیستم نمی‌تواند آن‌ها را هندل کند

مانند: a an and are as at from has - - - - will with

به طور کلی stopword ها تعداد زیاد کلمات حاوی اطلاعات غیرمفید در یک مدل زبانی هستند که تأثیر زیادی در به دست آوردن احتمال هر کلمه ایجاد می‌کنند زیرا تعداد آن‌ها زیاد است و در فرمول احتمالی مدرج را با کلماتی که حاوی اطلاعات خاصی نیستند زیادی می‌کنند.

stopwords کلمات بدون

اطلاعات ضروری و پرکاربرد را از مجموعه زبانی حذف

کنیم تا هم به سرعت و دقت بهتری برسیم.

(برگرفته از Search در گوگل!)

(در حد و اندازه اشاره شده)

بخش ۳- سؤال ۵.

فرمول Perplexity:

$$\sqrt[N]{\frac{1}{P(w_1 \dots w_N)}}$$

برای 3-grams ⇒ $PP(w) = \sqrt[N]{\prod_{i=2}^{N} \frac{1}{P(w_i | w_{i-1}, w_{i-2})}}$

برای 4-grams ⇒ $PP(w) = \sqrt[N]{\prod_{i=3}^{N} \frac{1}{P(w_i | w_{i-1}, w_{i-2}, w_{i-3})}}$

در ۲ فرمول بالا N ← تعداد توکن‌های تست

$P(w_i | w_{i-1,i-2}) \geq P(w_i | w_{i-1,i-2,i-3})$ می‌دانیم ←

← با 4-grams خروجی کمتر بالا کوچکتری می‌شود ،

← مقدار زیر رادیکال اضافه می‌شود ،

← مقدار سرگشتگی بزرگ‌تری می‌شود .

(معایب سیستم)

بخش ۳ ، سؤال ۶-

<s> B C A B A </s>

$P(B|<s>) \times P(C|B) \times P(A|C) \times P(B|A) \times$

$P(A|B) \times P(</s>|A)$

$$= \frac{P(<s>, B)}{P(<s>)} \times \frac{P(B, C)}{P(B)} \times \frac{P(C, A)}{P(C)} \times \frac{P(A, B)}{P(A)}$$

$$\times \frac{P(B,A)}{P(B)} \times \frac{P(A,</s>)}{P(A)}$$

$$= \frac{1}{5} \times \frac{6}{8} \times \frac{6}{16} \times \frac{5}{11} \times \frac{0}{8} \times \frac{1}{11} = 0$$

به نظرم نباید عبارت بالا ۰ میشه، ولی هرچه جزوه رو نگاه می‌کنم هیچ راه‌کاری برای این شرایط پیدا نمی‌کنم. ⇐ در این شرایط $PP(w) = \infty$

بخش ٤ ـ سؤال ١

فاصله = ۱