Top K, Top P en Temperature in AI-tekstgeneratie

INLEIDING **1**

OORSPRONG VAN DE NAMEN "TOP K" EN "TOP P" **2**

Top K 2
Top P 2

VERGELIJKING **2**

WAT DOEN TOP K, TOP P EN TEMPERATURE PRECIES? **3**

Top K: De Beperkte Boekenkast 3
Top P: De Selectieve Boekenkast 3
Temperature: De Thermostaat van Creativiteit 3

HOE WERKEN ZE? **4**

Top K in Actie 4
Top P in Actie 4
Temperature in Actie 5

OVEREENKOMSTEN EN VERSCHILLEN **5**

VOOR- EN NADELEN **6**

Top K 6
Top P 6
Temperature 6

PRAKTISCHE TOEPASSING **7**

Kiezen van Instellingen 7
Combineren van Methoden 7
Experimenteren en Finetunen 7

SAMENVATTING **8**

Inleiding

Stel je voor dat je een AI-assistent bent in een grote bibliotheek. Je taak is om boeken te kiezen voor bezoekers op basis van hun vragen. Hoe je deze boeken kiest, bepaalt hoe gevarieerd en verrassend je aanbevelingen zijn. In de wereld van AI-tekstgeneratie spelen Top K, Top P en Temperature een vergelijkbare rol. Ze bepalen hoe de AI woorden kiest bij het maken van zinnen.

In deze gids zullen we deze drie belangrijke concepten verkennen aan de hand van een doorlopend voorbeeld. We zullen zien hoe een AI-model het volgende woord kiest in de zin ***"Het weer is vandaag..."***.

Top K, Top P en Temperature in AI-tekstgeneratie

Oorsprong van de namen "Top K" en "Top P"

Top K

"Top K" staat voor "Top K-selectie".

- De "K" in Top K is simpelweg een variabele die het aantal top keuzes aangeeft dat we willen behouden.
- In de wiskunde en informatica wordt de letter "K" vaak gebruikt als een algemene constante of parameter, vooral wanneer het gaat om het tellen of rangschikken van items.
- "Top" verwijst naar het feit dat we de K meest waarschijnlijke (of "top") opties selecteren.

Dus "Top K" betekent letterlijk "de top K items", waarbij K een getal is dat we zelf kunnen kiezen.

Top P

"Top P" staat voor "Top Probability" of "Top Percentage".

- De "P" verwijst naar de cumulatieve waarschijnlijkheid (Probability) of het percentage (Percentage) dat we als drempelwaarde gebruiken.
- "Top" duidt hier op het feit dat we de meest waarschijnlijke opties selecteren totdat we de P-drempel bereiken.

Top P wordt ook wel "nucleus sampling" genoemd, omdat het de "kern" (nucleus) van de meest waarschijnlijke opties selecteert.

Vergelijking

- Top K selecteert een vast aantal (K) opties, ongeacht hun waarschijnlijkheden.
- Top P selecteert een variabel aantal opties, maar met een vaste totale waarschijnlijkheid (P).

Beide methoden zijn ontworpen om de "top" of meest relevante keuzes te behouden, maar doen dit op verschillende manieren: Top K kijkt naar aantallen, terwijl Top P kijkt naar cumulatieve waarschijnlijkheden.

Top K, Top P en Temperature in AI-tekstgeneratie

Wat doen Top K, Top P en Temperature precies?

Top K: De Beperkte Boekenkast

Top K is als een bibliothecaris die werkt met een vaste, beperkte boekenkast.

- Als K is ingesteld op 5:
  - De bibliothecaris plaatst alleen de 5 meest populaire boeken in de kast.
  - Alle keuzes worden gemaakt uit deze 5 boeken, ongeacht hoe populair ze zijn ten opzichte van elkaar.

Top P: De Selectieve Boekenkast

Top P, 'nucleus sampling', werkt als een selectieve boekenkast.

- Stel, Top P is ingesteld op 0.8 (80%):
  - De AI kijkt alleen naar de meest waarschijnlijke woorden die samen 80% van de totale waarschijnlijkheid vormen.
  - Het is alsof de bibliothecaris alleen kijkt naar de populairste 80% van de boeken en de rest negeert.

Temperature: De Thermostaat van Creativiteit

Stel je Temperature voor als een thermostaat die de "temperatuur" van de creativiteit van de AI regelt.

- **Lage Temperature** (dicht bij 0): De AI is als een voorzichtige bibliothecaris die alleen de meest voor de hand liggende boeken kiest. De antwoorden zijn voorspelbaar en consistent, maar mogelijk minder creatief.
- **Hoge Temperature** (richting 1 of hoger): De AI wordt een avontuurlijke bibliothecaris die ook obscure of onverwachte boeken durft aan te bevelen. De antwoorden zijn gevarieerder en creatiever, maar mogelijk minder samenhangend.

Top K, Top P en Temperature in AI-tekstgeneratie

Hoe werken ze?

Laten we ons doorlopende voorbeeld gebruiken om te laten zien hoe Top K, Top P en Temperature samenwerken. De AI moet het volgende woord kiezen in de zin:
"Het weer is vandaag...".

Top K in Actie

1. De AI rangschikt alle mogelijke volgende woorden op waarschijnlijkheid.
2. Alleen de top K meest waarschijnlijke woorden worden behouden.
3. De AI kiest uit deze K woorden, gewogen naar hun waarschijnlijkheid.

Voorbeeld:
- K = 5
- Alle mogelijke woorden en hun kansen:
  1. "zonnig" (30%)
  2. "regenachtig" (25%)
  3. "bewolkt" (20%)
  4. "warm" (10%)
  5. "koud" (8%)
  6. "stormachtig" (5%)
  7. "mistig" (2%)

Na Top K (K=5) blijven over:
- "zonnig" (30%)
- "regenachtig" (25%)
- "bewolkt" (20%)
- "warm" (10%)
- "koud" (8%)

Top P in Actie

1. Van de overgebleven woorden na Top K, telt de AI de waarschijnlijkheden op.
2. Zodra de som de Top P-waarde bereikt, worden alle overige woorden uitgesloten.

Voorbeeld:
- Top P = 0.8 (80%)
- Overgebleven woorden na Top K:
  "zonnig" (30%)
  "regenachtig" (25%)
  "bewolkt" (20%)
  "warm" (10%)
  "koud" (8%)

Na Top P (P=0.8) blijven over:
- "zonnig" (30%)
- "regenachtig" (25%)
- "bewolkt" (20%)
(Deze tellen op tot 75%, wat het dichtst bij 80% komt zonder er overheen te gaan)

Temperature in Actie

1. Temperature past de waarschijnlijkheden van de overgebleven woorden aan.
2. Lagere Temperature vergroot de verschillen, hogere Temperature verkleint ze.

Voorbeeld met overgebleven woorden na Top P:
- Bij lage Temperature (0.3):
  "zonnig" (sterk verhoogde kans, bijvoorbeeld 60%)
  "regenachtig" (verlaagde kans, bijvoorbeeld 25%)
  "bewolkt" (sterk verlaagde kans, bijvoorbeeld 15%)

- Bij hoge Temperature (1.2):
  "zonnig" (licht verlaagde kans, bijvoorbeeld 34%)
  "regenachtig" (licht verhoogde kans, bijvoorbeeld 33%)
  "bewolkt" (verhoogde kans, bijvoorbeeld 33%)

Uiteindelijke voorbeeldzinnen:

- Bij lage Temperature: ***"Het weer is vandaag zonnig."*** (meest waarschijnlijk)
- Bij hoge Temperature: ***"Het weer is vandaag bewolkt."*** (meer variatie mogelijk)

Overeenkomsten en Verschillen

Overeenkomsten:
- Alle drie beïnvloeden de variatie in de output.
- Hogere waarden leiden meestal tot meer diverse antwoorden.
- Lagere waarden geven meer voorspelbare resultaten.

Verschillen:
- Top K kiest een vast aantal opties, ongeacht de waarschijnlijkheidsverdeling.
- Top P past zich dynamisch aan de specifieke waarschijnlijkheidsverdeling aan.
- Temperature werkt globaal op alle waarschijnlijkheden.

Top K, Top P en Temperature in AI-tekstgeneratie

Voor- en Nadelen

Top K

Voordelen:
- Eenvoudig te begrijpen en implementeren.
- Voorkomt zeer onwaarschijnlijke keuzes effectief.

Nadelen:
- Minder flexibel bij verschillende waarschijnlijkheidsverdelingen.
- Kan te beperkend zijn als K te laag is ingesteld.

Top P

Voordelen:
- Past zich automatisch aan verschillende contexten aan.
- Biedt een goede balans tussen coherentie en variatie.

Nadelen:
- Minder intuïtief te begrijpen dan Top K.
- Kan bij zeer lage waarden te beperkend zijn.

Temperature

Voordelen:
- Intuïtief te begrijpen en in te stellen.
- Geeft fijne controle over de algemene 'creativiteit'.

Nadelen:
- Kan bij extreme waarden leiden tot onsamenhangendheid of overmatige herhaling.
- Minder adaptief aan verschillende contexten.

Top K, Top P en Temperature in AI-tekstgeneratie

Praktische Toepassing

Kiezen van Instellingen

Voor feitelijke informatie:
- Lage Top K (10-50) of lage Top P (0.1-0.3) of lage Temperature (0.3-0.5)

Voor creatief schrijven:
- Hogere Top K (100-1000) of hogere Top P (0.9-1.0) of hogere Temperature (0.7-1.0)

Voor conversaties:
- Gematigde instellingen (Top K: 50-100, Top P: 0.7-0.9, Temperature: 0.6-0.8)

Combineren van Methoden

Je kunt alle drie de methoden tegelijk gebruiken:
1. Top K filtert eerst op een vast aantal opties.
2. Top P filtert vervolgens binnen die selectie.
3. Temperature past ten slotte de waarschijnlijkheden binnen de overgebleven set aan.

Dit is als een bibliothecaris die eerst een vaste selectie maakt (Top K), dan kijkt naar de meest relevante boeken binnen die selectie (Top P), en tenslotte besluit hoe avontuurlijk de uiteindelijke keuze is (Temperature).

Experimenteren en Finetunen

Om de beste instellingen te vinden:
1. Begin met standaardwaarden (bijv. Top K 50, Top P 0.9, Temperature 0.7).
2. Genereer meerdere antwoorden met verschillende instellingen.
3. Evalueer de resultaten op basis van:
   - Relevantie
   - Creativiteit
   - Coherentie
   - Geschiktheid voor je specifieke taak
4. Pas geleidelijk aan en herhaal het proces.

Top K, Top P en Temperature in AI-tekstgeneratie

Samenvatting

Top K, Top P en Temperature zijn krachtige tools om de output van AI-tekstgeneratie te sturen. Door ze te begrijpen en ermee te experimenteren, kun je de AI-assistent afstemmen op jouw specifieke behoeften, of het nu gaat om het genereren van feitelijke rapporten, creatieve verhalen, of natuurlijke conversaties.

Ons doorlopende voorbeeld laat zien hoe deze drie methoden samenwerken:
1. Top K beperkt de keuze tot een vast aantal opties (in ons geval 5).
2. Top P verfijnt dit verder door alleen de meest waarschijnlijke opties te behouden (in ons geval 3).
3. Temperature bepaalt uiteindelijk hoe 'gewaagd' de finale keuze is.

Door deze methoden te combineren, kun je zeer precieze controle krijgen over het generatieproces en de output afstemmen op je specifieke behoeften.

September 2024,

Anthony Loeff – @Litic.ai
https://github.com/anthonyloeff