## Logistics

↳ HW 0 due 9/17

↳ Form project groups by 9/17

↳ 3-5 students

↳ project proposal due 10/6

↳ has to deal w/ language

---

## Language models

↳ a language model assigns a probability to a piece of text

$P(\text{"students opened their books"}) > P(\text{"students opened their doorknobs"})$

given a seq of words $S = w_1, w_2, \ldots, w_n$

a LM computes

$$P(S) = P(w_1, w_2, \ldots w_n) = P(w_{1 \ldots n})$$

What about "predict the next word"?

$$P(w_n \mid w_1, w_2 \ldots w_{n-1})$$

prefix

$$= P(w_n \mid w_{<n}) = P(w_n \mid w_{1 \ldots n-1})$$

remember the chain rule of probability

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

$P(A,B)$ → joint prob

$P(B|A)$ ← cond prob

$$P(A,B) = P(A)P(B|A)$$

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$

$$P(S) = P(w_{1 \ldots n})$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1 w_2), \ldots$$

$$= \prod_i P(w_i \mid w_{<i})$$

let's say we are given a training dataset of documents to estimate these probabilities.

let's just count and divide!

My favorite LLM is ___

$$P(\text{Gemini} \mid \text{My fav. LLM is}) \approx \frac{\text{Count}(\text{My favorite LLM is Gemini})}{\text{Count}(\text{My fav. LLM is})}$$

issues:

- ↳ sparsity
- ↳ no sharing counts between seqs w/ similar meaning
- ↳ storage

## n-gram models:

↳ approximate these probabilities by dropping context from the prefix

⇒ Markov assumption

$P(\text{Gemini} \mid \text{My for LLM is})$

$\approx P(\text{Gemini} \mid \text{is}) \Rightarrow \frac{\text{count}(\text{is Gemini})}{\text{count}(\text{is})} \Rightarrow$ bigram 2-gram

$\approx P(\text{Gemini} \mid \text{LLM is}) \Rightarrow \frac{\text{count}(\text{LLM is Gemini})}{\text{count}(\text{LLM is})} \Rightarrow$ trigram 3-gram

how many counts do we have to store for a n-gram model? $\Rightarrow V^n$ where $V$ is the size of the vocab

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

vocab: { <s>, </s>, I, am, Sam, do, not, like, green, eggs, and, ham }

↳ each unique element of vocab is called a **type**

↳ instances of types in a dataset are called **tokens**

$P(\text{I} \mid \text{<s>}) = \frac{2}{3}$ $P(\text{ham} \mid \text{<s>}) = 0$

$P(\text{Sam} \mid \text{<s>}) = \frac{1}{3}$ $P(\text{Sam} \mid \text{am}) = \frac{1}{2}$

$$P(w_n | w_{cn}) \approx \prod_i P(w_i) \Rightarrow \text{unigram}$$

## Evaluating language models

<s> Ham and eggs </s>

$P(\text{Ham} | \text{<s>}) \cdot P(\text{and} | \text{ham}) \cdot \dots$

as seqs get longer, the product of these cond. probs gets smaller and smaller

$$\log \prod_i P(w_i | w_{ci}) = \sum \log P(w_i | w_{ci})$$

in general, given a test doc $x_1, x_2, \dots x_t$

we want $P(x_{1 \dots t})$ to be high =

$$\frac{1}{t} \cdot \sum_i^t \log P(x_i | x_{ci}) \text{ to be high}$$

**PPL; perplexity:** $\exp\left(-\frac{1}{t} \sum_i^t \log P(x_i | x_{ci})\right)$

↳ want this to be <u>low</u>

intuition: perplexity measures how many equally likely next words is the model choosing from given a prefix

$P(\text{Gemini} \mid \text{is}) \Rightarrow$ PPL is high

$P(\text{Gemini} \mid \text{LLM is}) \Rightarrow$ PPL is lower

$P(\text{Gemini} \mid \text{My favorite LLM is}) \Rightarrow$ PPL is very low