CORPUS LINGUISTICS - NGRAMS

NATURAL LANGUAGE PROCESSING - CS 322.00

Blake Howald

September 22, 2017

AGENDA

- Presentation – Lazar Zamurovic, Solomon Foster
- Logistics
- Questions
- Corpus Linguistics
- Ngrams (Part 1)

CORPUS LINGUISTICS

- It's good to have data (generally) and lots of it.
- Can be use for the creation of a *Language Model (LM)* – Statistical model of word sequences (although can do at character level)
- Things to pay attention to when working with/ selecting corpora
  1. Representative of what you are looking to capture
  2. Robust enough to capture range of expected (and unexpected) variation
  3. 1. & 2. speak to the generalizeability of your model to unseen data
  4. Data "scrubbing" often required (garbage in – garbage out)
  5. Attention to inherent patterns of language and associated impacts on model (if any)

NGRAMS - APPROACH

*"Just one more regret in a long list of **many**"*

The (conditional) probability of a word *w* given an associated history *h*.

$P(w|h)$

$P(\text{many} \mid \text{just one more regret in a long list of})$

$C(\text{just one more regret in a long list of many}) / C(\text{just one more regret in a long list of})$

NGRAMS – ASSUMPTIONS

- *Markov* assumption lessens the burden on collecting and maintaining a long history
  - We can predict the probability of a future state without looking too far back in history

$$P(\textit{many} \mid \textit{just one more regret in a long list of})$$

$$P(\textit{many} \mid \textit{of})$$

$$P(\textit{many} \mid \textit{of}) \approx P(\textit{many} \mid \textit{just one more regret in a long list of})$$

NGRAMS – ASSUMPTIONS

$$P(\textit{many} \mid \textit{of}) \approx P(\textit{many} \mid \textit{just one more regret in a long list of})$$

$$C(\textit{of many})/C(\textit{of})$$

$$P(\textit{many} \mid \textit{list of}) \approx P(\textit{many} \mid \textit{just one more regret in a long list of})$$

$$C(\textit{list of many})/C(\textit{list of})$$

$$P(\textit{many} \mid \textit{long list of}) \approx P(\textit{many} \mid \textit{just one more regret in a long list of})$$

$$C(\textit{long list of many})/C(\textit{long list of})$$

Use of relative frequencies allows us to estimate not only the Maximum Likelihood Estimates (MLEs) of bigrams (or any ngram) but also the MLE of sequences of bigrams not explicitly seen in the corpus.

N(BI)GRAMS - CALCULATION

<s>Six sick bricks tick</s>

<s>Six sick chicks tock</s>

<s> six	P(six|<s>)	C(<s> six)/C(<s>) = 2/2 = 1
six sick	P(sick|six)	C(six sick)/C(six) = 2/2 = 1
sick bricks	P(bricks|sick)	C(sick bricks)/C(sick) = ½ = 0.5
bricks tick	P(tick|bricks)	C(bricks tick)/C(bricks) = 1/1 = 1
tick</s>	P(</s>|tick)	C(tick </s>)/C(tick) = 1/1 = 1
sick chicks	P(chicks|sick)	C(sick chicks)/C(sick) = ½ = 0.5
chicks tock	P(tock|chicks)	C(chicks tock)/C(chicks) = 1/1 = 1
tock</s>	P(</s>|tock)	C(tock </s>)/C(tock) = 1/1 = 1

NGRAMS - CALCULATION

Bigram	MLE	Bigram	MLE
have a	0.707	don't have	0.117
a box	0.589	I don't	0.117
box </s>	0.471	<s> I	0.589
I can't	0.471	can't be	0.471
to be	0.117	dumb </s>	0.117
so dumb	0.117	be so	0.333

- *I don't have a box*
- *P(I|<s>) x P(don't|I) x P(have|don't) x P(a|have) x P(box|a) x P(</s>|box)*
- *0.589x0.117 x 0.117 x 0.707 x 0.589 x 0.471 = 0.001581*
- *I can't be so dumb*
- *We don't have a box*