Lecture 21: Generating text

William Webber (`william@williamwebber.com`)

COMP90042, 2014, Semester 1, Lecture 21

Generating text from generative models

How about actually generating random text from a generative model?

Unigram model

- Take 20,000 news articles
  - from the Foreign Broadcast Information Service (FBIS)
  - c. 1994
- Generate unigram model from frequency of terms across all documents
- Generate random text with unigram frequencies

Unigram model text

by pierson mulders near military official and the of re issues shows kc178 with head go captain to there bfn an abuses the not migration the presidential at reform on terms has academia same cherkasy transnational constitutional to the give this profits districts teran to because leaders camps have request in this in is violated leaps visits and about training bosnia in a system their party of you the the opposition evolutionary will of really carried finally ever of carried that walkout pauses ready text rest scheduled his leader does detention technological 1 to disregarded percent indian the the a

- Word frequencies are reasonable
- But text makes no sense

Improving text realism with *n*-grams

- Fancier unigram language models (e.g. language models with smoothing) won't help much here
- Instead, use leading context with *n*-gram models

n-gram models

$$P(w_i = t) = P(t|w_{i-1}, w_{i-1}, \ldots, w_{i-n+1}) \tag{1}$$

- In *n*-gram model, probability of term generate for position i
- ... depends on terms found in preceding $n-1$ positions
- Build model by building probability dictionary for each $n-1$ context seen in training text
- Generate text sequentially by observing each preceding context of $n-1$ generated words
- Bootstrap by picking a random context (based on context frequency)

Bigram text

> the trial prisoners seized or high saying that is expected to stress that the offensive and finality on the union expects dm555 million has been clear very carefully grown before it finds it was going to see for this feeling guilty people who pretend that he has the primary objects and government posts are being in the cost the united states of imperialism by yael dan refused to have been reported that for election committee sevket kazan indicated colombian candidate visited the crew were working people and in a fresh dialogue with due to become sharply divided into the czech

- Text seems locally plausible
- But still not grammatical, sensible

Trigram, 4-gram text

Trigram

has already bought many buildings in which president zeroual dealing with the head of state agencies in accordance with the cooperation of the future of settlements to top level population managerial department composed of this this groups only parliamentary representatives of the unified iraqi national congress sacp south african police union also called for the borders of the lasva valley two nights ago

4-gram

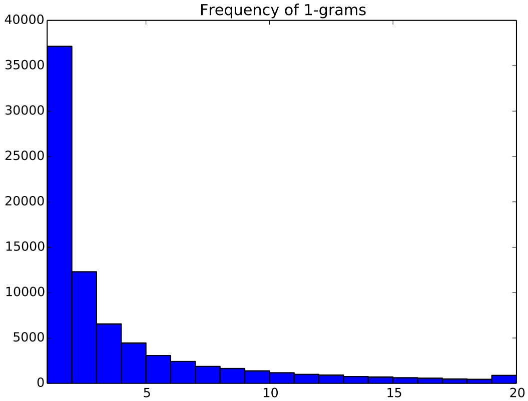
is promising bougie briefed zou on alcans plan of conducting further economic cooperation with estonia andreatta said meri thanked italy for its support to the sofols he says that the cleft in his chin quivers and his brown eyes beam behind his green framed glasses being in the opposition who the people know that is what is needed is not protection we advocate it is administration what is the british government to contribute constructively to achieving this accomplishment in its final form

5-gram

remove a quarter of his left lung he says the disease is in remission when mr han is not writing or calling his contacts in china hes always on the phone says a friend he takes walks in lammas hills remembering harsher times and people such as the vaccination and identification card campaigns it is not necessary us antinarcotics under secretary robert gelbard title as published last week gave signals about these two subjects to officials from the npc law committee zhou jue the spokesman said that conditions at the base about 70 members of the pmdb brazilian democratic movement

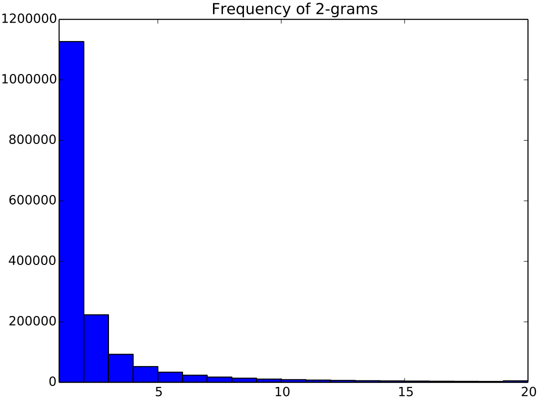
- By 5-grams, substantially reproducing whole articles
- Because 5-gram sequences are largely unique

n-gram distribution



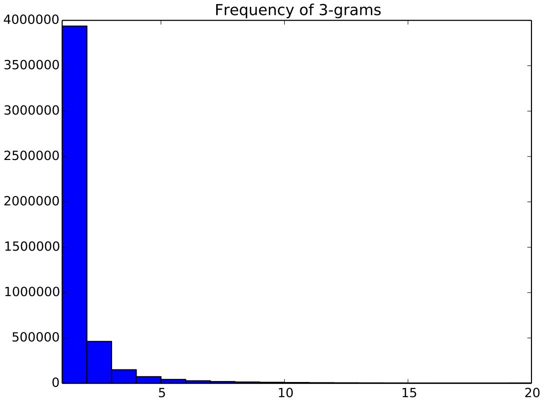
- As *n* increases
- . . . *n*-grams distribution increasingly skewed towards 1
- . . . and number of distinct *n*-grams increases
- . . . increasing memory requirements

n-gram distribution



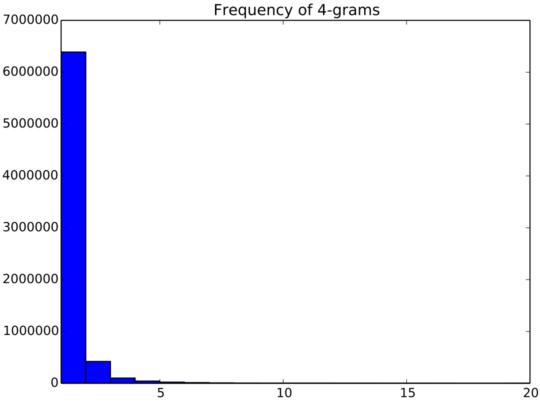
- As *n* increases
- . . . *n*-grams distribution increasingly skewed towards 1
- . . . and number of distinct *n*-grams increases
- . . . increasing memory requirements

n-gram distribution



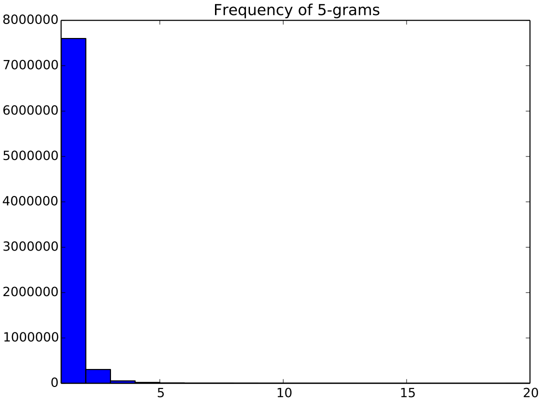
- As *n* increases
- . . . *n*-grams distribution increasingly skewed towards 1
- . . . and number of distinct *n*-grams increases
- . . . increasing memory requirements

n-gram distribution



- As *n* increases
- . . . *n*-grams distribution increasingly skewed towards 1
- . . . and number of distinct *n*-grams increases
- . . . increasing memory requirements

n-gram distribution



- As *n* increases
- . . . *n*-grams distribution increasingly skewed towards 1
- . . . and number of distinct *n*-grams increases
- . . . increasing memory requirements

Smoothed n-gram model

- We can “smooth” our ngram model
- Let there be t terms found in training data in a given $n-1$ context
- With probability $t/(t+1)$, choose one of these terms
- With probaility $1/(t+1)$, go to the $n-2$ context

Smoothed *n*-gram model

> both the israeli government intends to protect citizens but as foreign citizens mr wang mili said today that a former major general pavel zolotarev from foreign investors and ensure a smooth process in reform opening up and modernization in the new year they will probably vote in the forthcoming balance in other words there is no room in the eighth round of bilateral talks in the us capital you will see that we are concerned enormous efforts are most important they said hata called such an event chang said asked about the reasons replied that lisbon weekly semanario noting that turkish government credit is being implemented in close compliance with the agreed declaration

- Better?
- Might wish to smooth less agressively (i.e. less than $+1$)

Further improvements

- Handling of punctuation
  - Could add punctuation tokens from original text
  - But easy to get mismatched parentheses, quotes etc.
- Capitalization
- Part-of-speech tagging
  - Give preference to tokens of correct part of speech
- Post-processing to clean up grammar (?)