Ngrams

Lecture # 4

What is a word?

- A sequence of characters demarcated by white space

- Not exactly
  - Spoken language
  - Chinese
  - It’s → it is
  - Other issues in tokenization (e.g., New York, rock ‘n’ roll)

Types versus Tokens

- A text will contain various words
- Some of these words may occur more than once

How much wood could a woodchuck chuck if a wood chuck could chuck wood

- sentence contains:
  - 14 words (**tokens**)
  - 8 unique words (**types**)

Types versus Tokens

<u>Word Token</u>: an occurrence of a word at a particular spatio-temporal location (e.g., a sequential position in a text, an utterance event at a time and space).

<u>Word Type</u>: a more abstract notion also termed lexeme – we speak of two tokens belonging to the same type.

Also, *woodchuck* and *woodchucks are two grammatical forms of the same lexeme (woodchuck).*

Lexicon

Used in NLP systems to associate information with words (either for parsing or generation)

- Information about a word is called a lexical entry

- In parsing, each word in the input is scanned, and then lexical lookup retrieves one or more entries from the lexicon. Some of the information may be dynamically computed (e.g., by exploiting various lexical regularities).

- NLP-specific lexicons are similar to, but typically richer than, a printed dictionary

  - Some NLP systems have used machine-readable versions of printed dictionaries

Words

Words provide context

Looking at the frequent words can provide us theme/topic/understanding of the corpus that is being analyzed

THE 2008 CAMPAIGN: The Message and Corporate Marketing

The Words They Used

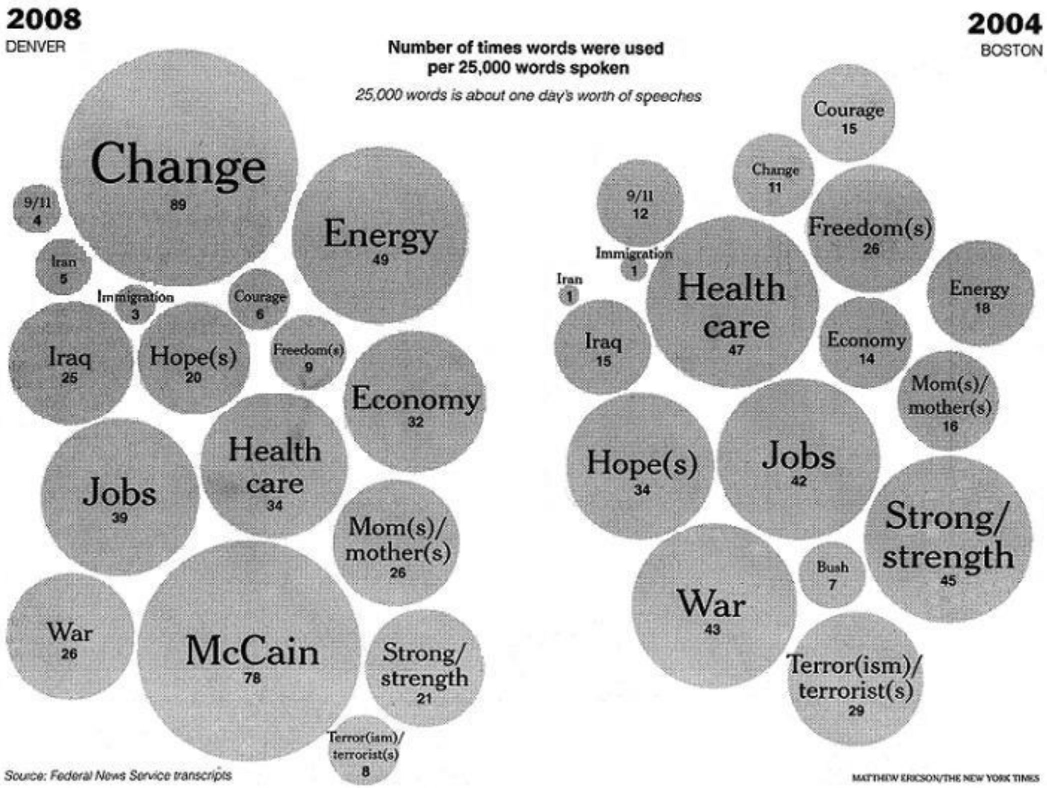
The words that the speakers have used during the Democratic convention suggest how the party's themes have changed since the last presidential campaign.

Speakers have hammered home Barack Obama's "change" theme, using the word about eight times as often as they did in 2004.

Also, unlike 2004, when the Kerry campaign sought to avoid direct attacks on the president at the convention, the speakers have regularly have been mentioning John McCain by name. Speakers in 2004 practiced "the art of the implicit slam," a veteran Democratic speechwriter said then, indirectly bashing Mr. Bush while barely using his name.

Also on the upswing: more mentions of the economy, energy, Iran and Iraq.

Words less frequently used: freedom, Sept. 11 and terrorism.



Source: Federal News Service transcripts

MATTHEW ERICSON/THE NEW YORK TIMES

So words are important

the frequency … or probability … of a word can provide use information

Word Prediction

- From a NY Times story...
  - Stocks ...
  - Stocks plunged this ....
  - Stocks plunged this morning, despite a cut in interest rates
  - Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall ...
  - Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall Street began

Word Prediction

- Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall Street began trading for the first time since last …
- Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall Street began trading for the first time since last Tuesday's terrorist attacks.

Human Word Prediction

Clearly, at least some of us have the ability to predict future words in an utterance.

How?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge

Domain knowledge

red hat

Red Hat

Syntactic knowledge

the <adjective or noun>

- The balloon ✓
- The big balloon ✓
- The eat? ✗

Lexical knowledge

baked <*potato* vs. *steak*>

Both can be baked

but what is more probable?

Claim

A useful part of the knowledge needed to allow Word Prediction can be captured using simple statistical techniques

In particular, we'll be interested in the notion of the probability of a sequence (of letters, words,...)

Useful Applications

- *Why do we want to predict a word, given some preceding words?*

  - Rank the likelihood of sequences containing various alternative hypotheses, e.g. for Speech Recognition

  Theatre owners say popcorn/unicorn sales have doubled...

  - Assess the likelihood/goodness of a sentence, e.g. for text generation or machine translation

  The doctor recommended a cat scan.

  Der Arzt empfahl eine Katze Scan.

  Computertomographie.

N-Gram Models

We formalize this concept of word prediction using probabilistic language models:

Refer to as ***N-gram models***

N-gram

N refers to the number of words in the sequence

- Unigram (sequence of 1 word)
- Bigram (sequence of 2 words)
- Trigram (sequence of 3 words)
- 4-gram (sequence of 4 words)
- 5-gram (sequence of 5 words)
- etc….

Example

this is an unoriginal example utterance

What are the {1 2 and 3}-grams?

Example

this is an unoriginal example utterance

Unigrams: ?

Bigrams:

Trigrams:

Example

this is an unoriginal example utterance

Unigrams:
- this
- is
- an
- unoriginal
- example
- utterance

Bigrams: ?

Trigrams:

Example

this is an unoriginal example utterance

Unigrams:

- this
- is
- an
- unoriginal
- example
- utterance

Bigrams:

- this is
- is an
- an unoriginal
- unoriginal example
- example utterance

Trigrams: ?

Example

this is an unoriginal example utterance

Unigrams:
- this
- is
- an
- unoriginal
- example
- utterance

Bigrams:
- this is
- is an
- an unoriginal
- unoriginal example
- example utterance

Trigrams:
- this is an
- is an unoriginal
- an unoriginal example
- unoriginal example utterance

N-Gram Models of Language

Use the previous N-1 words in a sequence to predict the next word

Language Model (LM)

unigrams, bigrams, trigrams,…

How do we train these models?

Very large corpora

N-gram Modeling

Assume a language has T word types in its lexicon, and what we are predicting is :

how likely is word *x* to follow word *y*?

Basic Probability

- $P(e)$ = *a priori probability*
  - the chance that *e* happens.

Basic Probability

- $P(e)$ = *a priori probability*
  - the chance that *e* happens

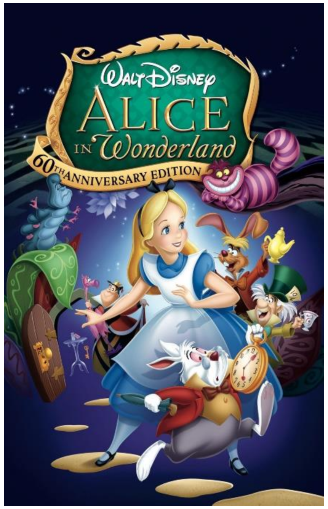
$$P(e) = \frac{freq(e)}{N}$$

how often does word *e* happen divided all the words in corpus

e.g.

Let *e* = "beautiful"

$$P(beautiful) = \frac{freq(beautiful)}{N}$$



Basic Probability

- $P(e)$ = *a priori probability*
  - the chance that *e* happens.
- $P(f|e)$ = conditional probability
  - the chance of *f* given *e*

Basic Probability

- $P(f|e)$ = conditional probability
  - the chance of f given e

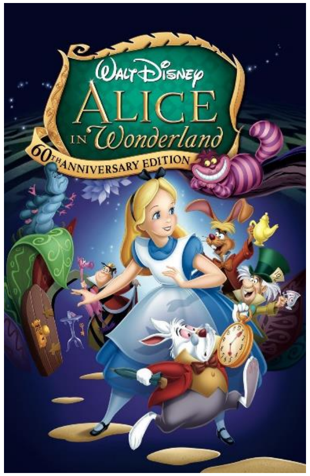
$$P(f|e) = \frac{freq(f,e)}{freq(e)}$$

how often does word f happen given we saw word e

e.g.

Let e = “beautiful” and f = “soup”

$$P(soup|beautiful) = \frac{freq(beautiful\ soup)}{freq(beautiful)}$$



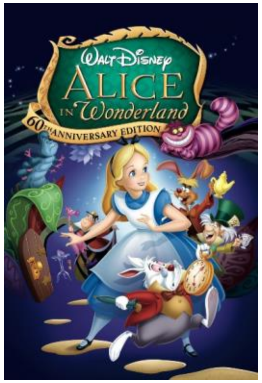
Basic Probability

- $P(e)$ = *a priori probability*
  - the chance that *e* happens.
- $P(f|e)$ = conditional probability
  - the chance of *f* given *e*
- $P(f, e)$ = joint probability
  - the chance of both *e* and *f* happening
  - If *e* and *f* are independent, we can write
    - $P(e, f) = P(e) * P(f)$

$$P(e) = \frac{freq(e)}{N} \quad P(f) = \frac{freq(f)}{N}$$

Basic Probability

- $P(f, e)$ = joint probability
  - the chance of both *e* and *f* happening
  - If *e* and *f* are independent, we can write
    - $P(e, f) = P(e) * P(f)$



how often does word *e* and *f* happen divided all the words in the corpus

e.g.

Let *e* = "beautiful" and *f* = "soup"

$$P(beautiful, soup) = \frac{freq(beautiful)}{N} * \frac{freq(soup)}{N}$$

N-gram Modeling

Assume a language has T word types in its lexicon, and what we are predicting is: how likely is word *x* to follow word *y*?

Theatre owners say popcorn|unicorn sales have doubled.

what is the probability that x = unicorn versus x = popcorn?

N-gram Modeling

Assume a language has T word types in its lexicon, and what we are predicting is: how likely is word *x* to follow word *y*?

Theatre owners say popcorn|unicorn sales have doubled.

what is the probability that x = unicorn versus x = popcorn?

Model #1 : Simplest model of word probability: 1/T

This assumes that any word in the language has the same probability of occurring as any other

P(unicorn) = P(popcorn)

kind of basic – looks only at the vocabulary size

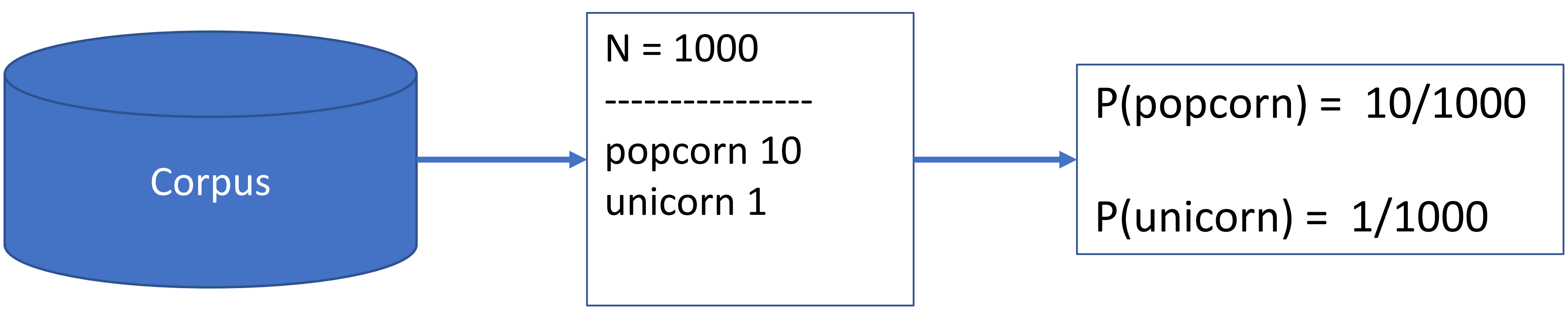
N-gram Modeling

Assume a language has T word types in its lexicon, and what we are predicting is: how likely is word *x* to follow word *y*?

Theatre owners say popcorn|unicorn sales have doubled.

what is the probability that x = unicorn versus x = popcorn?

Model #2 : estimate likelihood of x occurring based on its general frequency of occurrence estimated from a corpus (unigram probability)



Taking the frequency of the word into consideration

N-gram Modeling

Assume a language has T word types in its lexicon, and what we are predicting is: how likely is word *x* to follow word *y*?

Theatre owners say popcorn|unicorn sales have doubled.

what is the probability that x = unicorn versus x = popcorn?

Model #3 : condition the likelihood of x occurring in the context of previous words

what is more likely?

Theater owners say unicorn ➔ P(unicorn|Theater owners say)

or

Theater owners say popcorn ➔ P(popcorn|Theater owners say)

Taking the context the word is used into consideration

This works okay with shorter sentences

P(the mythical unicorn)

But…the *longer* the sequence, the *less likely*

we are to find it in a training corpus

P(Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the narwhal) = ?

Enter : Chain Rule

Chain rule:

allows us to decompose the probability into a product of component conditional probabilities

P(the mythical unicorn) = P(the) * P(mythical|the) * P(unicorn|the mythical)

$$P(X_1 \ldots X_n) = P(X_1)\mathrm{P}(X_2|X_1)P(X_3|X_1^2) \ldots P(X_n|X_1^{n-1})$$

$$= \prod_{k=1}^{n} P(x_k|x_1^{k-1})$$

$$P(X_1 \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1^2) \ldots P(X_n|X_1^{n-1})$$

Problem

P(Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the narwhal)

P(Most) *
P(biologists|most) *
P(and |most biologists) *
P(folklore|Most biologists and) *
P(specialist|Most biologists and folklore) *

…

P(narwhal|Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the)

Tough to find the sequence in text

P(narwhal | Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the) =

$$\frac{\text{Freq(Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the narwhal)}}{\text{Freq(Most biologists and folklore specialists believe that in fact the mythical unicorn horns derived from the)}}$$

$$\boxed{P(X_k | X_1^{k-1}) = \frac{freq(X_1 \dots X_k)}{freq(X_1 \dots X_{k-1})}}$$

Enter : Markov Assummption:

Markov Assumption:

word is only dependent on its limit history

- This allows us only go back *k*-1 words

Estimate: P(unicorn|the mythical) using (unicorn|mythical)

$$\boxed{P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P(x_k | x_{k-1})}$$

How far back do we go: N-gram Model

Bigram: $$P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P(x_k | x_{k-1})$$

Trigram: $$P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P\left(x_k \middle| x_{k-2}^{k-1}\right)$$

4-gram: $$P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P\left(x_k \middle| x_{k-3}^{k-1}\right)$$

Relative Frequency

$$P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P\left(x_k \middle| x_{k-N+1}^{k-1}\right)$$

$$P\left(x_k \middle| x_{k-N+1}^{k-1}\right) = \frac{freq(x_{k-N+1}^{k-1}\, x_k)}{freq(x_{k-N+1}^{k-1}\,)}$$

Relative Frequency

$$P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P\left(x_k \middle| x_{k-N+1}^{k-1}\right)$$

$$P\left(x_k \middle| x_{k-N+1}^{k-1}\right) = \frac{freq(x_{k-N+1}^{k-1}\, x_k)}{freq(x_{k-N+1}^{k-1})}$$

this ratio is called relative frequency

Example

Relative Frequency Example

using the bigram model

I want to eat Chinese Food.

$$P(x_1^n) = \prod_{k=1}^{n} P\left(x_k \middle| x_1^{k-1}\right) = \prod_{k=1}^{n} P(x_k | x_{k-1})$$

Example

I want to eat Chinese Food.

$$P\left(x_k \middle| x_{k-N+1}^{k-1}\right) = \frac{freq(x_{k-N+1}^{k-1}\, x_k)}{freq(x_{k-N+1}^{k-1}\,)}$$

Example

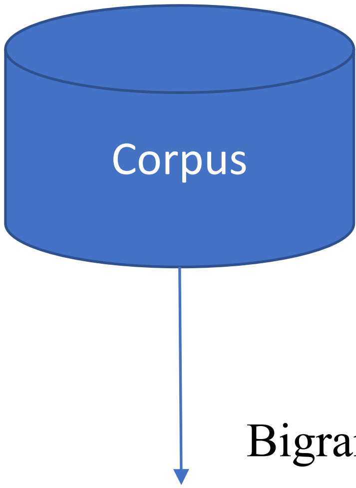
I want to eat Chinese Food.

$$P(x_k|x_{k-1}) = \frac{freq(x_{k-1}x_k)}{freq(x_{k-1})}$$

What information do we need?

Where do we get it from?

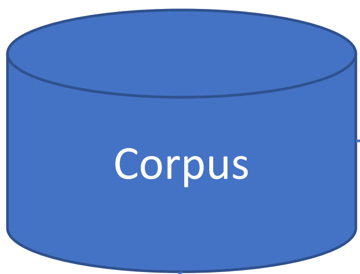
$$P(x_k|x_{k-1}) = \frac{freq(x_{k-1}x_k)}{freq(x_{k-1})}$$



Bigram table of raw frequency’s

	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	5	827	0	9	0	0	0	2	0	0
want	2	0	608	1	6	6	5	1	0	0
to	2	0	4	686	2	0	6	211	0	0
eat	0	0	2	0	16	2	42	0	0	34
chinese	1	0	0	0	0	82	1	0	0	23
food	15	0	15	0	1	1	0	0	0	12
lunch	2	0	0	0	0	0	0	0	0	9
spend	1	0	1	0	0	0	0	0	1	17
<start>	45	0	30	0	15	10	3	0	0	0
<end>	0	0	0	0	3	23	6	34	0	0

$$P(x_k|x_{k-1}) = \frac{freq(x_{k-1}x_k)}{freq(x_{k-1})}$$



i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
2533	927	2417	746	158	1093	341	278	3000	3000

Unigram table of raw frequency’s

Bigram table of raw frequency’s

	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	5	827	0	9	0	0	0	2	0	0
want	2	0	608	1	6	6	5	1	0	0
to	2	0	4	686	2	0	6	211	0	0
eat	0	0	2	0	16	2	42	0	0	34
chinese	1	0	0	0	0	82	1	0	0	23
food	15	0	15	0	1	1	0	0	0	12
lunch	2	0	0	0	0	0	0	0	0	9
spend	1	0	1	0	0	0	0	0	1	17
<start>	45	0	30	0	15	10	3	0	0	0
<end>	0	0	0	0	3	23	6	34	0	0

$$P(x_k|x_{k-1}) = \frac{freq(x_{k-1}x_k)}{freq(x_{k-1})}$$

	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	5	827	0	9	0	0	0	2	0	0
want	2	0	608	1	6	6	5	1	0	0
to	2	0	4	686	2	0	6	211	0	0
eat	0	0	2	0	16	2	42	0	0	34
chinese	1	0	0	0	0	82	1	0	0	23
food	15	0	15	0	1	1	0	0	0	12
lunch	2	0	0	0	0	0	0	0	0	9
spend	1	0	1	0	0	0	0	0	1	17
<start>	45	0	30	0	15	10	3	0	0	0
<end>	0	0	0	0	3	23	6	34	0	0

I	want	to	eat	chinese	food	lunch	spend	<start>	<end>
2533	927	2417	746	158	1093	341	278	3000	3000

$$P(want \mid i) = ?$$

$$P(x_k|x_{k-1}) = \frac{freq(x_{k-1}x_k)}{freq(x_{k-1})}$$

	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	5	827	0	9	0	0	0	2	0	0
want	2	0	608	1	6	6	5	1	0	0
to	2	0	4	686	2	0	6	211	0	0
eat	0	0	2	0	16	2	42	0	0	34
chinese	1	0	0	0	0	82	1	0	0	23
food	15	0	15	0	1	1	0	0	0	12
lunch	2	0	0	0	0	0	0	0	0	9
spend	1	0	1	0	0	0	0	0	1	17
<start>	45	0	30	0	15	10	3	0	0	0
<end>	0	0	0	0	3	23	6	34	0	0

I	want	to	eat	chinese	food	lunch	spend	<start>	<end>
2533	927	2417	746	158	1093	341	278	3000	3000

$$P(want \mid i) = \frac{827}{2533} = 0.33$$

	i	want	to	eat	chinese	food	lunch	spend	<start>	<end>
i	0.002	0.33	0	0.0036	0	0	0	0.00079	0	0
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0064	0.0011	0	0
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087	0	0
eat	0	0	0.0027	0	0.021	0.0027	0.056	0	0	0.011
chinese	0.0063	0	0	0	0	0.52	0.0063	0	0	0.008
food	0.014	0	0.014	0	0.00092	0.0037	0	0	0	0.004
lunch	0.0059	0	0	0	0	0.0029	0	0	0	0.003
spend	0.0036	0	0.0036	0	0	0	0	0	1	0.006
<start>	0.015	0	0.01	0	0.005	0.003	0.001	0	0	0
<end>	0	0	0	0	0.001	0.007	0.002	0.011	0	0

Relative Frequency Table

Where do the frequency info come from?

- Corpora are online collections of text and speech
  - Brown Corpus
  - Wall Street Journal
  - AP newswire
  - Hansards
  - Timit
  - DARPA/NIST text/speech corpora (Call Home, Call Friend, ATIS, Switchboard, Broadcast News, Broadcast Conversation, TDT, Communicator)
  - TRAINS, Boston Radio News Corpus

What does this allow you to do

Calculate the probability of arbitrarily long sentences

Generate arbitrarily long sentences

Using N-gram Modeling to Generate Sentences: Approximating Shakespeare

- Generating sentences with random unigrams...
  - Every enter now severally so, let
  - Hill he late speaks; or! a more to leg less first you enter
- With bigrams...
  - What means, sir. I confess she? then all sorts, he is trim, captain.
  - Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry.
- Trigrams
  - Sweet prince, Falstaff shall die.
  - This shall forbid it should be branded, if renown made it empty.

- Quadrigrams
  - What! I will go seek the traitor Gloucester.
  - Will you not tell me who I am?
  - What's coming out here looks like Shakespeare because it ***is*** Shakespeare

- Note: ***As we increase the value of N, the accuracy of an n-gram model increases, since choice of next word becomes increasingly constrained***

N-Gram Training Sensitivity

- If we repeated the Shakespeare experiment but trained our n-grams on a Wall Street Journal corpus, what would we get?

- Note: ***This question has major implications for corpus selection or design***

Sentences Generated from WSJ

unigram: Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

bigram: Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

trigram: They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Ok – let’s stop here

What you need to review and know

- Calculate the probability and conditional probability of words
- Understand the Chain Rule
- Understand the Markov assumption

Questions?