

# NLP

# Introduction to NLP

## *Language models (Part 1)*

# Probabilistic Language Models

- Assign a probability to a sentence
  - $P(S) = P(w_1, w_2, w_3, \dots, w_n)$
- Different from deterministic methods using CFG
- The sum of the probabilities of all possible sentences must add up to 1

# Predicting The Next Word

- Example
  - Let's meet in Times ...
  - General Electric has lost some market ...
- Formula
  - $P(w_n | w_1, w_2, \dots, w_{n-1})$

# Predicting the Next Word

- What word follows “your”?

- [http://norvig.com/ngrams/count\\_2w.txt](http://norvig.com/ngrams/count_2w.txt)

your abilities	160848
your ability	1116122
your ablum	112926
your academic	274761
your acceptance	783544
your access	492555
your accommodation	320408
your account	8149940
your accounting	128409
your accounts	257118
your action	121057

your actions	492448
your activation	459379
your active	140797
your activities	226183
your activity	156213
your actual	302488
your ad	1450485
your address	1611337
your admin	117943
your ads	264771
your advantage	242238
your adventure	109658
your advert	101178
your advertisement	172783

# Uses of Language Models

- Speech recognition
  - $P(\text{"recognize speech"}) > P(\text{"wreck a nice beach"})$
- Text generation
  - $P(\text{"three houses"}) > P(\text{"three house"})$
- Spelling correction
  - $P(\text{"my cat eats fish"}) > P(\text{"my xat eats fish"})$
- Machine translation
  - $P(\text{"the blue house"}) > P(\text{"the house blue"})$
- Other uses
  - OCR
  - Summarization
  - Document classification
- Usually coupled with a translation model (later)

## Probability Of A Sentence

- How to compute the probability of a sentence?
- What if the sentence is novel?
- What we need to estimate:
  - $P(S) = P(w_1, w_2, w_3 \dots w_n)$
- Using the chain rule:
  - $P(S) = P(w_1) P(w_2|w_1) P(w_3|w_1, w_2) \dots P(w_n|w_1, w_2 \dots w_{n-1})$
- Example:
  - $P(\text{"I would like the pepperoni and spinach pizza"}) = ?$

## N-gram Models

- Predict the probability of a word based on the words before:
  - $P(\text{square} | \text{Let's meet in Times})$
- Markov assumption
  - Only look at limited history
- N-gram models
  - Unigram – no context:  $P(\text{square})$
  - Bigram:  $P(\text{square} | \text{Times})$
  - Trigram:  $P(\text{square} | \text{in Times})$



# Random Text (Brown Corpus)

- 2-grams:

The 53-year-old Shea was no acceptable formula to help the abuse of events were a wall in 1908 , called upon his hand in Southern New Orleans , Miss Garson was named Maurice Couve De Havilland signed a privilege resolution had had happened on a tax applied to the Chisholm , the thriving systems of the `` Pride and musician , and Moscow made good team spirit of the culmination of the metal tube through the amateur , but rather than a special prosecutor . This knowledge of each member of these savings of golf course can see the 13 straight 69 . Since 1927 by Harry Truman Cleveland of railroad retirement age groups . No Vacancy " . `` I have to congressmen . The remainder of the rear bumper and on a benefit in U.S. amateur , as far as a thrill a \$100 U.S. if not indicted . The state's occupation tax dollars over the newest product of the address he attended Arlington State University will pay half years .

# Random Text (Brown Corpus)

- 3-grams:

The Fulton County Jail and `` a very strong central government of Laos that the presence of picket lines and featuring a flared skirt and lace jacket with bateau neckline and princesse skirt accented by lace appliques . Her acting began with the members of the government -- such control is necessary to build in a final exchange between Moscow and Washington last week . Of course , since the views of another one . It urged that the games are not essential to provide federal contributions to the 85-student North Carolina group to play , was addressing a meeting in the manufacture of a tax bill since most of his uncle and aunt , also was particularly struck by the reams came in from shareholders of these co-operative systems , the 9th precinct of the guiding spirits of the Armed Services Committee . Davis received 1,119 votes in Saturday's election , the executive organs of participation can hardly escape the impression that he made no attempt to get agreement among the conference's top four in rushing , he was awarded the top but that presently we're not acting as we head that way .

## Random Text (Brown Corpus)

- 4-grams:

The broadcast said Anderson , a Seattle ex-marine and Havana businessman , and McNair , of Miami , were condemned on charges of smuggling arms to Cuban rebels . Anderson operated three Havana automobile service stations and was commander of the Havana American Legion post before it disbanded since the start of August have shown gains averaging nearly 10% above last year . That , too , in improving motorists' access to many turnpikes . The Kansas Turnpike offers an illustration . Net earnings of that road rose from 62 per cent of the prices that the avid buyers bid it up to . Dallas and North Texas is known world-wide as the manufacturing and distribution center of cotton gin machinery and equipment . The firm is design-conscious , sales-conscious , advertising-conscious . `` Hodges predicted : ' I think we should certainly follow through on it ' . Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called `` blue law ' ' controversy came in the form of a letter to Mayor Grady that plowing and salting crews should be dispatched earlier in storms and should be kept on the job .

## Higher Order n-grams

- It is possible to go to 3,4,5-grams
- Longer n-grams suffer from sparseness

## N-grams

- Shakespeare unigrams
  - 29,524 types, approx. 900K tokens
- Bigrams
  - 346,097 types, approx. 900K tokens
- Notice – very sparse data!

# Estimation

- Can we compute the conditional probabilities directly?
  - No, because the data is sparse
- Markov assumption
  - $P(\text{"musical" | I would like two tickets for the"}) = P(\text{"musical" | the"})$
  - or
    - $P(\text{"musical" | I would like two tickets for the"}) = P(\text{"musical" | for the"})$

# Maximum Likelihood Estimates

- Use training data
- Count how many times a given context appears in it.
- Unigram example:
  - The word “pizza” appears 700 times in a corpus of 10,000,000 words.
  - Therefore the MLE for its probability is  $P'(\text{“pizza”}) = 700/10,000,000 = 0.00007$
- Bigram example:
  - The word “with” appears 1,000 times in the corpus.
  - The phrase “with spinach” appears 6 times
  - Therefore the MLE for  $P'(\text{spinach}|\text{with}) = 6/1,000 = 0.006$
- These estimates may not be good for corpora from other genres

## Example

- $P(\text{"<S> I will see you on Monday</S>"}) =$ 
  - $P(I|<S>)$
  - $\times P(\text{will}|I)$
  - $\times P(\text{see}|\text{will})$
  - $\times P(\text{you}|\text{see})$
  - $\times P(\text{on}|\text{you})$
  - $\times P(\text{Monday}|\text{on})$
  - $\times P(\text{</S>}|\text{Monday})$



# Example from Jane Austen

- $P(\text{"Elizabeth looked at Darcy"})$
- Use maximum likelihood estimates for the n-gram probabilities
  - unigram:  $P(w_i) = c(w_i) / V$
  - bigram:  $P(w_i | w_{i-1}) = c(w_{i-1}, w_i) / c(w_{i-1})$
- Values
  - $P(\text{"Elizabeth"}) = 474 / 617091 = .000768120$
  - $P(\text{"looked"} | \text{"Elizabeth"}) = 5 / 474 = .010548523$
  - $P(\text{"at"} | \text{"looked"}) = 74 / 337 = .219584569$
  - $P(\text{"Darcy"} | \text{"at"}) = 3 / 4055 = .000739827$
- Bigram probability
  - $P(\text{"Elizabeth looked at Darcy"}) = .000000001316 = 1.3 \times 10^{-9}$
- Unigram probability
  - $P(\text{"Elizabeth looked at Darcy"}) = 474 / 617091 * 337 / 617091 * 4055 / 617091 * 304 / 617091 = .000000000001357 = 1.3 \times 10^{-12}$
- $P(\text{"looked Darcy Elizabeth at"}) = ?$

# NLP