Introduction to Probabilistic Language Model

Ramaseshan Ramachandran

How are ____? Can you guess the missing word?

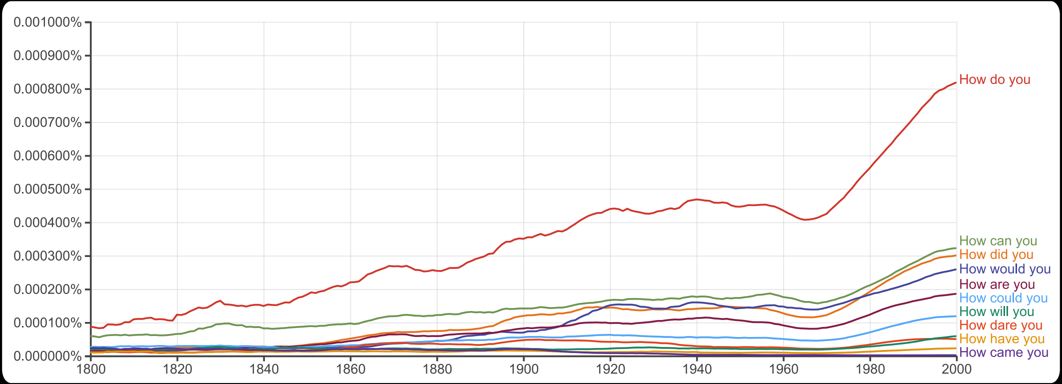
INTRODUCTION

How are _____? Can you guess the missing word?

INTRODUCTION

How _____ you? Can you guess the missing word?

How _____ you? Can you guess the missing word?



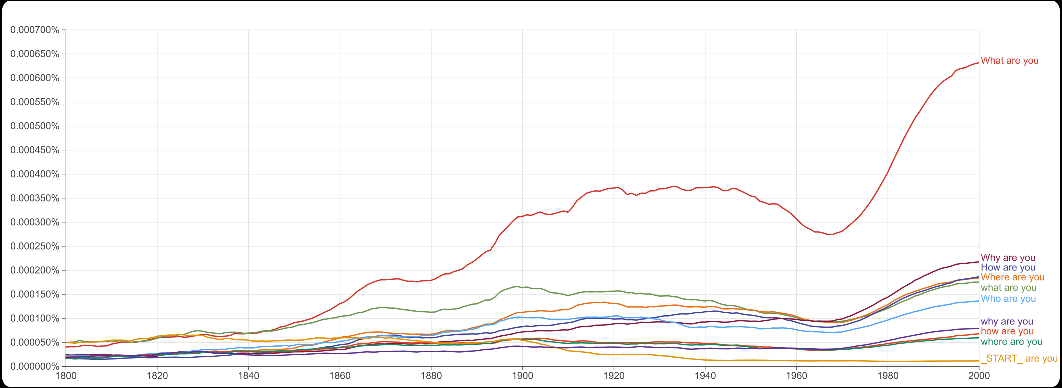
Source:Google NGram Viewer

INTRODUCTION

_____ are you?

INTRODUCTION

_____ are you?



Source:Google NGram Viewer

INTRODUCTION

How do humans predict the next word?

How do humans predict the next word?

- Domain knowledge

INTRODUCTION

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge

INTRODUCTION

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge
- Knowledge about the sentence structure
-
- Some words are hard to find. Why?

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge
- Knowledge about the sentence structure
-
- Some words are hard to find. Why?
- Natural language is not deterministic in general

INTRODUCTION

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge
- Knowledge about the sentence structure
-
- Some words are hard to find. Why?
- Natural language is not deterministic in general
- Some sentences are familiar or had been heard/seen/used several times

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge
- Knowledge about the sentence structure
-
- Some words are hard to find. Why?
- Natural language is not deterministic in general
- Some sentences are familiar or had been heard/seen/used several times
- They are more likely to happen than others, hence we could guess

INTRODUCTION

How do humans predict the next word?

- Domain knowledge
- Syntactic knowledge
- Lexical knowledge
- Knowledge about the sentence structure
-
- Some words are hard to find. Why?
- Natural language is not deterministic in general
- Some sentences are familiar or had been heard/seen/used several times
- They are more likely to happen than others, hence we could guess

WHY PROBABILITY?

WHY PROBABILITY?

- Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data

WHY PROBABILITY?

- Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- Make the informed decision when there a certain degree of uncertainty and some observed data

WHY PROBABILITY?

- Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- Make the informed decision when there a certain degree of uncertainty and some observed data
- It provides a quantitative description of the chances or likelihoods associated with various outcomes

WHY PROBABILITY?

- Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- Make the informed decision when there a certain degree of uncertainty and some observed data
- It provides a quantitative description of the chances or likelihoods associated with various outcomes
- Probability of a sentence
- Probability of the next word in a sentence - how likely to predict "**you**" as the next word

WHY PROBABILITY?

- Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- Make the informed decision when there a certain degree of uncertainty and some observed data
- It provides a quantitative description of the chances or likelihoods associated with various outcomes
- Probability of a sentence
- Probability of the next word in a sentence - how likely to predict "**you**" as the next word
- Likelihood of the next word is formalized through an observation by conducting experiment - counting the words in a document

PROBABILITY - DEFINITION AND PROPERTY

- The Probability is defined as the likelihood that an event will occur
- Let us use the most popular example - a flip of a coin - there is a 50% chance or probability that heads will come up for any given toss of a fair coin
- Probabilities can be expressed as percentage (60%), in decimal form (0.6) or in fractions (6/10)

DISCRETE SAMPLE SPACE

Consider following bag of words ($count = 52$)

Experiment - Extracting tokens from a document
Outcome - Every token/word x in the document
'a', 'weather', 'balloon', 'is', 'floating', 'at', 'a', 'constant', 'height', 'above', 'earth', 'when', 'it', 'releases', 'a', 'pack', 'of', 'instruments', 'level', 'a', 'if', 'the', 'pack', 'hits', 'the', 'ground', 'with', 'a', 'downward', 'velocity', 'of', 'm', 's', 'how', 'far', 'did', 'the', 'pack', 'fall', 'b', 'calculate', 'the', 'distance', 'the', 'ball', 'has', 'rolled', 'at', 'the', 'end', 'of', 's'

The outcome of the experiment - 52 samples (words). They constitute the ***sample space***, Ω or the set of all possible ***outcomes***
Each word in this sample belongs to Ω, represented by $x \in \Omega$
Each sample $x \in \Omega$ is assigned a probability score $[0, 1]$
A *probability function or probability distribution function* distributes the probability mass of 1 to the all the samples in the sample space Ω

All the words in the Ω, must satisfy the following constraints:

1. $P(x) \in [0,1], \forall x \in \Omega$ and
2. $\sum_{x \in \Omega} P(x) = 1$

EXAMPLE - 1

Bag of words $Count = 52$

'a', 'weather', 'balloon', 'is', 'floating', 'at', 'a', 'constant', 'height', 'above', 'earth', 'when', 'it', 'releases', 'a', 'pack', 'of', 'instruments', 'level', 'a', 'if', 'the', 'pack', 'hits', 'the', 'ground', 'with', 'a', 'downward', 'velocity', 'of', 'm', 's', 'how', 'far', 'did', 'the', 'pack', 'fall', 'b', 'calculate', 'the', 'distance', 'the', 'ball', 'has', 'rolled', 'at', 'the', 'end', 'of', 's'

If we are equally likely to pick any word from the BOW, then the probability for any word is
$P(x) = 1/52, \forall x \in \Omega$ so that
$P(\Omega) = 1$
$P('weather') = 1/52 = 0.01923076923$

EVENTS

'a', 'weather', 'balloon', 'is', 'floating', 'at', 'a', 'constant', 'height', 'above', 'earth', 'when', 'it', 'releases', 'a', 'pack', 'of', 'instruments', 'level', 'a', 'if', 'the', 'pack', 'hits', 'the', 'ground', 'with', 'a', 'downward', 'velocity', 'of', 'm', 's', 'how', 'far', 'did', 'the', 'pack', 'fall', 'b', 'calculate', 'the', 'distance', 'the', 'ball', 'has', 'rolled', 'at', 'the', 'end', 'of', 's'

Total number of words $= 52$. The number of unique words $= 37$ or there are 37 ***types*** of words in this BOW. 15 words have frequencies > 1.

An ***event*** is a collection of samples of the same type, $E \subseteq \Omega$

$$P(E) = \sum_{x \in E} P(x) \qquad (1)$$

Events can be described as a variable taking a certain value

EVENTS - EXAMPLE

'a', 'weather', 'balloon', 'is', 'floating', 'at', 'a', 'constant', 'height', 'above', 'earth', 'when', 'it', 'releases', 'a', '**pack**', 'of', 'instruments', 'level', 'a', 'if', '**the**', '**pack**', 'hits', '**the**', 'ground', 'with', 'a', 'downward', 'velocity', 'of', 'm', 's', 'how', 'far', 'did', '**the**', '**pack**', 'fall', 'b', 'calculate', '**the**', 'distance', '**the**', 'ball', 'has', 'rolled', 'at', '**the**', 'end', 'of', 's'

In the BOW, the word type **the** occurs 6 times. Then

$$E_{the} = 6$$

$$P(E_{the}) = 6 \times \frac{1}{52} = 0.115$$

In the BOW, the word type **pack** occurs 3 times. Then

$$E_{pack} = 3$$

$$P(E_{pack}) = 3 \times \frac{1}{52} = 0.058$$

RANDOM VARIABLE

- A **random variable**,[1] is a variable whose possible values are numerical outcomes of a random phenomenon
- Two types - continuous and discrete - for NLP, they are discrete

To capture the type-token distinction, we use random variable W. $W(x)$ maps to the sample $x \in \Omega$.

V is the set of types and the value is represented by a variable v.

Given a random variable V and a value v, $P(V = v)$ is the probability of the event that V takes the value v, i.e.: $P(V = v) = P(x \in \Omega : V(x) = v)$

$P(V =' the') = P('the') = 0.115$

Random variables are useful in describing/constructing various events

[1] Random Variable - http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html#randvar

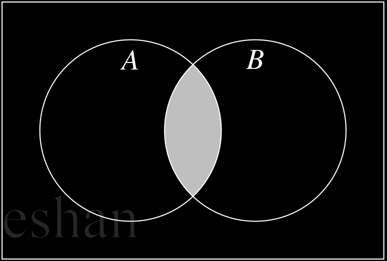
JOINT PROBABILITY

Given any two events E_1 and E_2, the probability of their conjunction

$$P(E_1, E_2) = P(E_1 \cap E_2) \qquad (2)$$

is called the ***joint probability***[2] of E_1 and E_2. This probability, E_1 and E_2, occurs simultaneously.

Example The probability of the the first letter of 't' and the second letter 'h' is $P(F = 't', S = 'h')$. The joint probability should be as large as the probability of $P('the')$



$P(A) =$ size of A relative to Ω

$P(A, B) =$ size of $A \cap B$ relative to Ω

[2] https://cs.brown.edu/courses/csci1460/assets/files/langmod.pdf

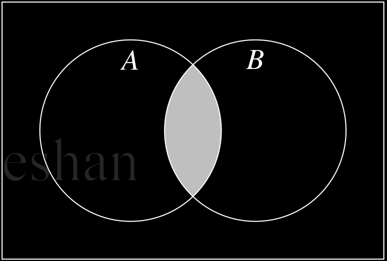
CONDITIONAL PROBABILITY

When we have partial knowledge influencing the outcome of an experiment, we use it to update the outcome.

The ***conditional probability*** $P(E_2|E_1)$ is the probability of event E_2 given that event E_1 has occurred. $P(E_2|E_1)$ is defined as:

$$P(E_2|E_1) = \frac{P(E_1, E_2)}{P(E_1)}, \text{ if } P(E_1) > 0 \quad (3)$$

$$= \frac{P(E_1 \cap E_2)}{P(E_1)} \quad (4)$$



$P(A) =$ size of A relative to Ω

$P(A,B) =$ size of $A \cap B$ relative to Ω

$P(A|B) =$ size of $A \cap B$ relative to B

CONDITIONAL PROBABILITY - BIGRAM EXAMPLE

Let consider a corpus of Kinematics problems in physics that contains about 280+ problems (*very small corpus*).

- Bigram Sample Space - $\{w_1, w_2\} \in \Omega = 3767$
- $A = \{w_1, w_2\} = \{\textbf{average}, *\}$ - bigram starting with *average*
- $B = \{w_1, w_2\} = \{*, \textbf{speed}\}$ - bigram ending with *speed*
- $P(average) = 0.036$
- $P(speed) = 0.114$
- $P(average, speed) = P(speed, average) = 0.004$
- $P(speed|average) = \dfrac{0.004}{0.036} = 0.111$
- $P(average|speed) = \dfrac{0.004}{0.114} = 0.035$

CONDITIONAL PROBABILITY - TRIGRAM EXAMPLE

Let consider a corpus of Kinematics problems in physics that contains about 280+ problems (*very small corpus*).

- Trigram Sample Space - $\{(w_1, w_2), w_3\} \in \Omega = 5902$
- $A = \{(w_1, w_2), w_3\} = \{\textbf{average,speed}, of\}$ - trigram starting with $(average, speed)$
- $B = \{(w_1, w_2), w_3\} = \{average, speed\}, \textbf{of}\}$ - trigram ending with of
- $C = \{(w_1, w_2), w_3\} = \{average, speed\}, \textbf{for}\}$ - trigram ending with for
- $D = \{(w_1, w_2), w_3\} = \{average, speed\}, \textbf{during}\}$ - bigrams ending with $during$
- $P(average, speed) = 0.0032; \quad P(average, speed, of) = 0.0007$
- $P(average, speed, for) = 0.0005; \quad P(average, speed, during) = 0.0002$

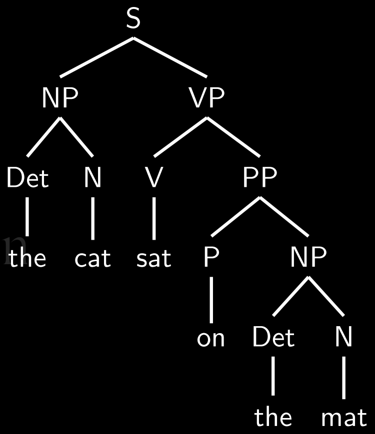
$$
\begin{aligned}
P(of|average, speed) &= \frac{0.0007}{0.0032} &= 0.21875 \\
P(for|average, speed) &= \frac{0.0005}{0.0032} &= 0.15576 \\
P(during|average, speed) &= \frac{0.0002}{0.0032} &= 0.0625
\end{aligned}
$$

INDEPENDENCE

- Two events are dependent if the probability of one relies on occurrence of the other; if there is no such interaction, then the events are independent
- Two events E_1 and E_2 are independent if and only if $P(E_1, E_2) = P(E_1)P(E_2)$
- OR
  - $P(E_1) = P(E_1|E_2)$
  - $P(E_2) = P(E_2|E_1)$
- Example
  - $P(average) = 0.036$
  - $P(speed) = 0.114$
  - $P(average, speed`) = 0.004$
- The bigram $\{average, speed\}$ did not happen by chance. The words $average, speed$ are **NOT** independent

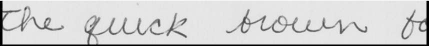
THE LANGUAGE MODEL

- Natural language sentences can be described by parse trees which use the morphology of words, syntax and semantics
- Probabilistic thinking - finding how likely a sentence occurs or formed, given the word sequence.
- In probabilistic world, the Language model is used to assign a probability $P(W)$ to every possible word sequence W.

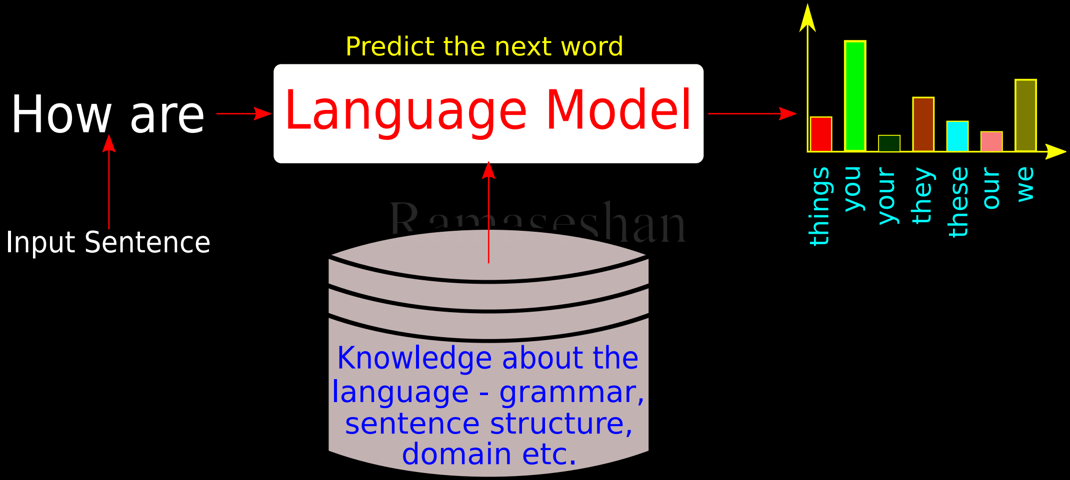


The current research in Language models focuses more on building the model from the huge corpus of text

APPLICATIONS

Application	Sample Sentences
Speech Recognition	Did you hear ***Recognize speech*** or Wreck a nice beach?
Context sensitive Spelling	One upon a ***tie***, ***Their*** lived aking
Machine translation	artwork is good $\rightarrow$ l'oeuvre est bonne
Sentence Completion	Complete a sentence as the previous word is given - GMail
OCR and Hand-written recognition	

A SIMPLE LANGUAGE MODEL IMPLEMENTATION



WHY PROBABILISTIC MODEL

- Speech recognition systems cannot depend on the processed speech signals. It may require the help of a language model and context recognizer to convert a speech to correct text format.
- As there are multiple combinations for a word to be in the next slot in a sentence, it is important for language modeling to be probabilistic in nature - judgment about the fluency of a sequence of words returns the probability of the sequence
- The probability of the next word in a sequence is real number $[0, 1]$
- The combination of words with high-probability in a sentence are more likely to occur than low-probability ones
- A probabilistic model continuously estimates the rank of the words in a sequence or phrase or sentence in terms of frequency of occurrence

Goal: Compute the probability of a sequence of words

$$P(W) = P(w_1, w_2, w_3, ...w_n) \tag{5}$$

Task: To predict the next word using probability. Given the context, find the next word using

$$P(w_n | w_1, w_2, w_3, \ldots, w_{n-1}) \tag{6}$$

A model which computes the probability for (5) or predicting the next word (6) or complete the partial sentence is called as Probabilistic Language Model.
The goal is to learn the joint probability function of sequences of words in a language.
The probability of $P($The cat roars $)$ is less likely to happen than $P($The cat meows$)$

CHAIN RULE

It is difficult to compute the probability of the entire sequence $P(w_1, w_2, w_3, \ldots, w_n)$?
Chain rule is used to decompose the joint probability of a sequence into a product of conditional probability

$$P(W) = P(w_1, w_2, w_3, \ldots, w_n) = P(w_1^n) \tag{7}$$

$$= P(w_1)P(w2|w_1)P(w3|w_2, w_1)\ldots P(w_n|w_{n-1}, w_{n-2}, w_{n-3}, \ldots, w_1) \tag{8}$$

$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \tag{9}$$

- It is possible to $P(w|h)$, but it does not really help in reducing the computational complexity
- We use innovative ways to string words to form new sentences
- Finding the probability for a long sentence may not yield good outcome as the context may never occur in the corpus
- Short sequences may provide better results

MARKOV ASSUMPTION

Markov Assumption: The future behavior of a dynamic system depends on its recent history and not on the entire history

The product of the conditional probabilities can be written approximately for a bigram as

$$P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-1}) \tag{10}$$

Equation (10) can be generalized for an *n-gram* as

$$P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-K+1}^{k-1}) \tag{11}$$

Now, the joint probability of a sequence can be re-written as

$$\begin{align}
P(W) &= P(w_1, w_2, w_3, \ldots, w_n) = P(w_1^n) \tag{12}\\
&= P(w_1)P(w2|w_1)P(w3|w_2, w_1)\ldots P(w_n|w_{n-1}, w_{n-2}, w_{n-3}, \ldots, w_1) \tag{13}\\
&= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \tag{14}\\
&\approx \prod_{k=1}^{n} P(w_k|w_{k-K+1}^{k-1}) \tag{15}
\end{align}$$

TARGET AND CONTEXT WORDS

Next word in the sentence depends on its immediate past words, known as context words

$$P(w_{k+1}|\underbrace{w_{i-k}, w_{i-k+1}, \ldots, w_k}_{\text{Context words}})$$

n-grams

- unigram - $P(w_{k+1})$
- bigram - $P(w_{k+1}|w_k)$
- trigram - $P(w_{k+1}|w_{k-1}, w_k)$
- 4-gram - $P(w_{k+1}|w_{k-2}, w_{k-1}, w_k)$

LANGUAGE MODELING USING UNIGRAMS

- A unigram language model all words are generated independently $W, W_2, W_3, \ldots\ldots W_n$ and none of them depend on the other
- This is not a good model for language generation
- It may generate ***the the the the*** as a sentence

GENERATIVE MODEL

- Generates a document containing N words using n-gram
- A good model assigns higher probability to the word that actually occurs

$$P(\mathbf{W}) = P(N)\prod_{i=1}^{N} P(W_i) \tag{16}$$

- The location of the word in the document is not important
- P(N) is the distribution over N and is same for all documents. Hence it is ignored
- W_i, to be estimated in this model is $P(W_i)$ and it must satisfy $\sum_{i=1}^{N} P(w_i) = 1$

MAXIMUM LIKELIHOOD ESTIMATE

- One of the methods to find the unknown parameter(s) is the use of Maximum Likelihood Estimate
- Estimate the parameter value for which the observed data have the highest probability
- Training data may not have all the words in the vocabulary
- If a sentence with an unknown word is presented, then the MLE is zero.
- Add a smoothing parameter to the equation without affecting the overall probability requirements

$$P(\mathbf{W}) = \frac{C_{w_i} + \alpha}{C_W + \alpha|V|} \tag{17}$$

$$\text{If } \alpha = 1, \text{then it is called as Laplace smoothing} \tag{18}$$

$$P(\mathbf{W}) = \frac{C_{w_i} + 1}{C_W + |V|} \tag{19}$$

BIGRAM LANGUAGE MODEL

- Bigram language model generates a sequence one word at a time, starting with the first word and then generating each succeeding word conditioned on the previous one[3]
- A bigram language model is defined as follows:

$$P(\mathbf{W}) = \prod_{i=1}^{n+1} P(w_i|w_{i-1}), \tag{20}$$

where $\mathbf{W} = w_1, w_2, w_3, \ldots, w_n$

- Estimate the parameter $P(w_i|w_{i-1})$ for all bigrams
- The parameter estimation does not depend on the location of the word
- If we consider the sentence as a sequence in time, they are time-invariant MLE picks up the word that is $\frac{n_{w,w'}}{n_{w,o}}$ where nw, w' is the number of times the words w_1, w' occur together and $n_{w,o}$ is the number of times the word w appears in the bigram sequence

[3] https://cs.brown.edu/courses/csci1460/assets/files/langmod.pdf

PROBABILISTIC LANGUAGE MODEL - EXAMPLE

Peter Piper picked a peck of pickled peppers
A peck of pickled peppers Peter Piper picked
If Peter Piper picked a peck of pickled peppers
Where's the peck of pickled peppers Peter Piper picked?

—

The joint probability of a sentence formed with n words can be expressed as a product conditional probabilities - we use immediate context and not the entire history

$P(w_1|\langle S\rangle) \times P(w_2|w_1) \times \ldots P(\langle E\rangle|w_n)$

and $P(w_{i+1}|w_i) = \frac{C(w_i.w_{i+1})}{C(w_i)}$

—

What is the probability of these sentences?
P(Peter Piper picked)
P(Peter Piper picked peppers)

Bigram	Frequency
$\langle S\rangle$Peter	1
Peter Piper	4
Piper picked	4
picked a	2
a peck	2
peck of	4
pickled peppers	4
peppers $\langle E\rangle$	1
$\langle S\rangle A$	1
A peck	1
of pickled	4
peppers Peter	2
...	..
$\langle S\rangle$ ...	1

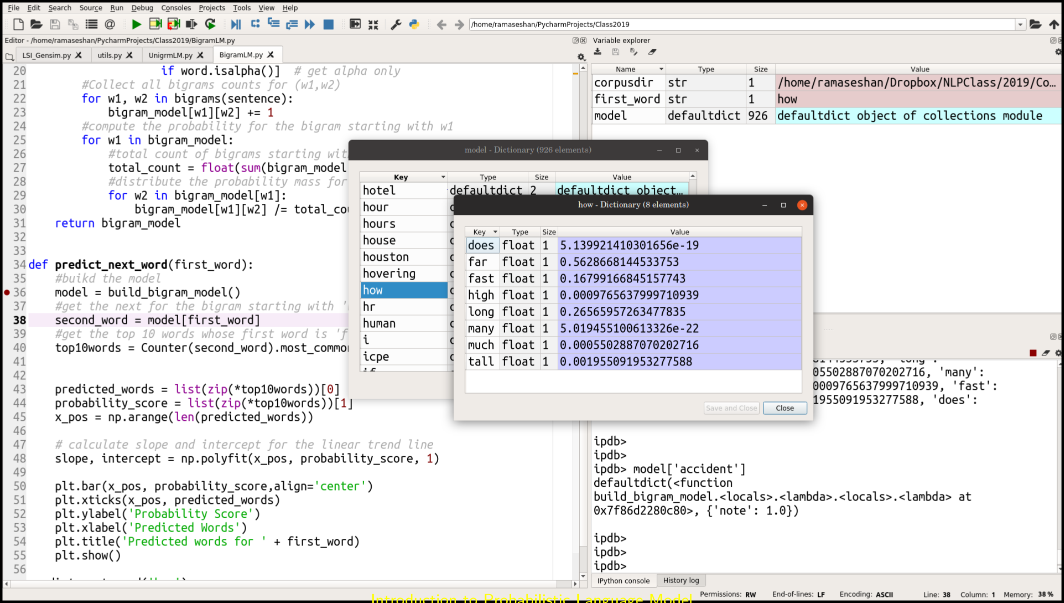
BUILDING A BIGRAM MODEL - CODE

```python
#compute the bigram model
def build_bigram_model():
    bigram_model = collections.defaultdict(
        lambda: collections.defaultdict(lambda: 0))
    for sentence in kinematics_corpus.sents():
        sentence = [word.lower() for word in sentence
                    if word.isalpha()]  # get alpha only
        #Collect all bigrams counts for (w1,w2)
        for w1, w2 in bigrams(sentence):
            bigram_model[w1][w2] += 1
        #compute the probability for the bigram containing w1
        for w1 in bigram_model:
            #total count of bigrams conaining w1
            total_count = float(sum(bigram_model[w1].values()))
            #distribute the probability mass for all bigrams starting with w1
            for w2 in bigram_model[w1]:
                bigram_model[w1][w2] /= total_count
    return bigram_model
```

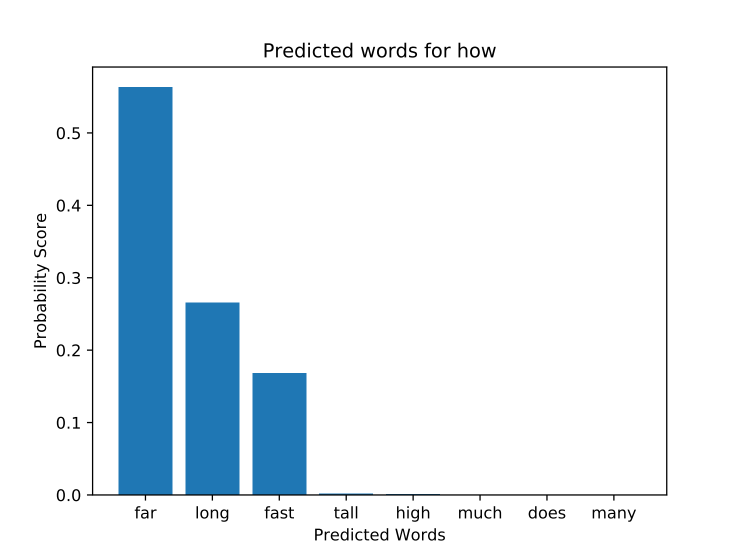
BUILDING A BIGRAM MODEL - CODE

```
def predict_next_word(first_word):
    #buikd the model
    model = build_bigram_model()
    #get the next for the bigram starting with 'word'
    second_word = model[first_word]
    #get the top 10 words whose first word is 'first_word'
    top10words = Counter(second_word).most_common(10)

    predicted_words = list(zip(*top10words))[0]
    probability_score = list(zip(*top10words))[1]
    x_pos = np.arange(len(predicted_words))

    plt.bar(x_pos, probability_score,align='center')
    plt.xticks(x_pos, predicted_words)
    plt.ylabel('Probability Score')
    plt.xlabel('Predicted Words')
    plt.title('Predicted words for ' + first_word)
    plt.show()

predict_next_word('how')
```

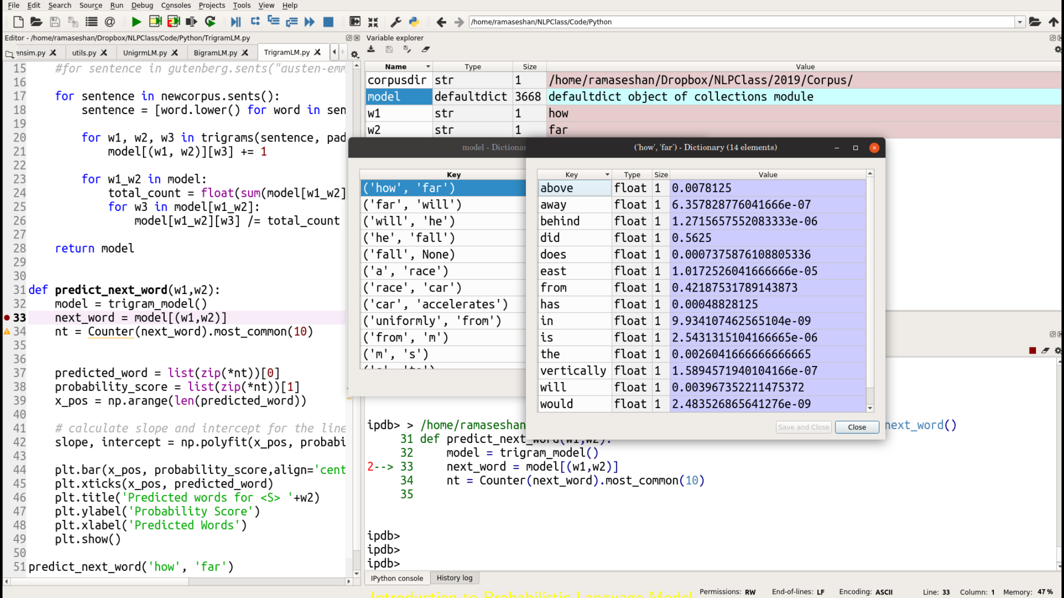
MODEL PARAMETERS - BIGRAM EXAMPLE



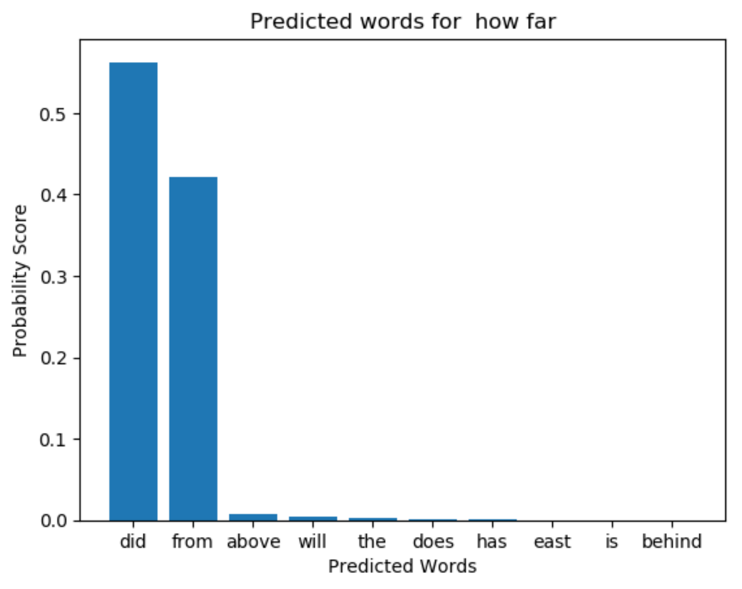
BIGRAM MODEL - NEXT WORD PREDICTION



MODEL PARAMETERS - TRIGRAM EXAMPLE



TRIGRAM MODEL - NEXT WORD PREDICTION



PERPLEXITY

Perplexity is a measurement of how well a probability model predicts a sample. Perplexity is defined as

$$\text{For bigram model, } PP(W_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}} \tag{21}$$

$$\text{For trigram model } PP(W_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1}w_{i-2})}} \tag{22}$$

A good model gives maximum probability to a sentence or minimum perplexity to a sentence

UNKNOWN WORDS

- In a closed vocabulary language model, there is no unknown words or ***out of vocabulary words (OOV)***
- In an open vocabulary system, you will find new words that are not present in the trained model
- Pick words below certain frequency and replace them as OOV.
- Treat every OOV as a regular word
- During testing, the new words would be treated as OOV and the corresponding frequency will be used for computation
- This eliminates zero probability for sentences containing OOV

CURSE OF DIMENSIONALITY

- A fundamental problem that makes language modeling and other learning problems difficult is the curse of dimensionality
- It is particularly obvious in the case when one wants to model the joint distribution between many discrete random variable
- If one wants to estimate the joint probability distribution of 10 words in a language with a million words as vocabulary, then we need to estimate $1000000^{10} - 1 = 10^{60} - 1$ free parameters

.

FIND THE SENDER OF THE EMAIL

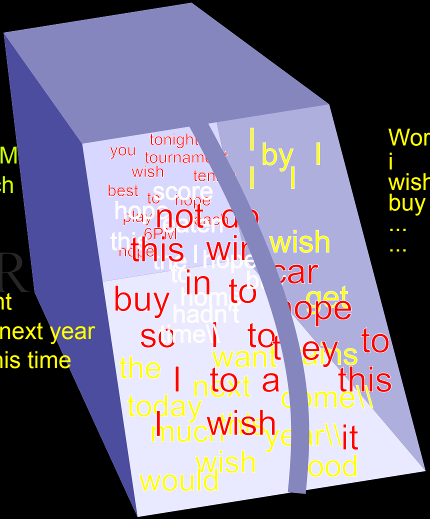
Assume that Ram and Raj exchanged the following emails

Ram	Raj
I wish you the best	I hope to play tennis tonight
I hope to reach home by 6PM	I hope to win this tournament
I wish to go home early	I hope to buy this car in the next year
I do not want to buy this	I wish to get a good score this time
I hope it rains today	I wish they would come

Who would have sent this email "I wish you would come"

BAG OF WORDS - EMAILS

I wish you the best
I hope to reach home by 6PM
I wish I hadn't eaten so much
I do not want to buy this
I hope it rains today
I hope to play tennis tonight
I hope to win this tournament
I hope to buy this car in the next year
I wish to get a good score this time
I wish they would come



Word Frequency
i
wish
buy
...
...

Who would have sent this email "I wish you would come" This question can be answered by using Bayes theorem

BAYES THEOREM

Let us consider two random variables $X\, and\, Y$. Then Joint probability, $P(X = x, Y = y)$, refers to the probability that the variable X takes the value x and the variable Y takes the value y. The conditional probability $P(Y = y|X = x)$ refers to the probability that the variable Y takes the value y given the observation the variable X takes the value x

$$P(X,Y) = P(Y|X) \times P(X) = P(X|Y) \times P(Y) \tag{23}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{24}$$

MAPPING BAYES THEOREM TO EMAIL CLASSIFICATION PROBLEM

- Map Bayes theorem using the statistical properties of the data
- Let $\mathbf{X} =$ and Y represent the random variables, where $\mathbf{X}$ is a set of attributes or is a attribute variable and Y represent a class.
- The relationship between $\mathbf{X}$ and Y can be found using the conditional probability $P(Y|\mathbf{X})$
- The conditional probability $P(Y|\mathbf{X})$ is known as posterior probability of Y
- $P(Y)$ is known as the prior probability
- In the classification problem, it is important to learn the parameters $P(Y|\mathbf{X})$. Given the attributes of the email (TF,TF-IDF), find the class to which the email belongs - in this case the person who sent it.

The parameters are obtained from the training data - the corpus of emails written by Ram and Raj. During the training process, we will learn $P(Y|\mathbf{X})$ for every word in the corpus

SUPERVISED CLASSIFICATION

- Set of input parameters/attributes $\mathbf{X} = X_1, X_2, \ldots, X_m$and a fixed set of classes $Y = y_1, y_2, \ldots, y_n$
- Every element of the training set, $D = d_1, d_2, \ldots, d_n$ is manually assigned a class
  $$(d_1, y_1), (d_2, y_2), (d_3, y_1), \ldots$$
- Goal is to learn the classifier, so that it can map a new document $\hat{d}$ to any of the classes, $y \in Y$
- Bayes classifier would assign a probability based on the observation to the new document to aid the class selection
- The probability score for each class is computed as given by the equation
  $$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{P(\mathbf{X})}$$
- The class will be found using $\arg\max_{y \in Y} P(Y|\mathbf{X})$

ESTIMATING THE CONDITIONAL PROBABILITY $P(X_I|Y)$

$$
\begin{aligned}
\hat{y} &= \arg\max_{y \in Y} P(Y|\mathbf{X}) && (25)\\
&= \arg\max_{y \in Y} P(\mathbf{X}|y)P(y) && (26)\\
&= \arg\max_{y \in Y} P(y)P(X_1, X_2, X_m|Y) && (27)\\
&= \arg\max_{y \in Y} P(y)P(X_1|y) \times P(X_2|y) \times \ldots P(X_m|y) && (28)\\
&= \arg\max_{y \in Y} P(y)\prod_{i=1}^{m} P(X_i|y) && (29)
\end{aligned}
$$

TRAINING

1. Prior probability - $P(y) = \dfrac{Count(y)}{Count(Y)}$
2. Learn $P(X_1|y) = \dfrac{Count(X_1, y)}{Count(Y)}$

Word	Frequency

HANDS ON EXERCISE 1 - FIND THE SENDER OF THE EMAIL

Assume that Ram and Raj share emails exchanged emails using the words given in the table. A new email arrives with just three words - ***motivate, profit and product***. Find the sender using the historical information given in the table

Historical Information

Ram	Raj
motivate(0.24)	motivate(0.05)
profit(0.3)	profit(0.35)
product(0.26)	product(0.35)
leadership(0.08)	leadership(0.15)
operations(0.12)	operations(0.10)

Who would have used these words (motivate, profit and product) in the email ?
Is it possible to apply this technique to identify the sentiments of a movie review with two classes **Good** and **bad**?

HANDS ON EXERCISE 2 - PRODUCT SENTIMENTS

Assume the following likelihoods for each word being part of a positive or negative review, and equal prior probabilities for each class - positive and negative ($P(positive) = 0.5$ and $P(negative) = 0.5$)

word	positive	negative
I	0.09	0.16
love	0.07	0.06
to	0.05	0.07
fill	0.29	0.06
credit	0.04	0.15
card	0.08	0.11
application	0.06	0.04

What class Naive Bayes classifier would assign to the sentence "I do not like to fill in the application form?"