

Text Generation Project

- You will write a Python program that generates random text.
- You will implement two approaches:
 - picking random words from an existing text
 - 2-grams approach
- You will test your approaches with 3 different input files.

Generating random words

- Given an input text, choose *n* random words from the text.

- For example, given the following input text:

The raw, lingering emotion of the 2016 presidential campaign erupted into a shouting match here Thursday as top strategists of Hillary Clinton's campaign accused their Republican counterparts of fueling and legitimizing racism to elect Donald Trump.

We can generate the following text by choosing 25 words at random from the input text:

Republican a into Trump as here racism their as 2016 of fueling of and match elect campaign of accused raw The Hillary Thursday of presidential.

Generating random words

Exercise 1

- Given an input text, choose *n* random words from the text.
- For example, given the following input text:

The raw, lingering emotion of the 2016 presidential campaign erupted into a shouting match here Thursday as top strategists of Hillary Clinton's campaign accused their Republican counterparts of fueling and legitimizing racism to elect Donald Trump.

- Generate the text that would result from choosing the following 15 words from the input text. The number correspond to the position of the word in the input text:

36, 20, 19, 15, 1, 35, 22, 20, 29, 9, 12, 23, 35, 21

Trump. of strategists here The ...

Generating Random Words Algorithm

I	really	like	to	eat	burgers.	I	like	them	with	bacon.
0	1	2	3	4	5	6	7	8	9	10

- Given the list of words in the input text (in Python, you will use a list), we can now generate passages of text using the following method:
 1. REPEAT (as many times as words you want in the generated passage)
 - A. Generate a random number **index** between 0 and the number of words in the input text -1
 - B. Get the word at position **index** and append it to the output passage.

N-grams

- An **n-gram** is a sequence of **N** consecutive words from a given passage of text.
- For example, based on the following input sentence:
I really like to eat burgers. I like them with bacon.

we have the following **2-grams**:

'I', 'really'

'really', 'like'

'like', 'to'

'to', 'eat'

'eat', 'burgers'

'burgers', 'I'

'I', 'like'

'like', 'them'

'them', 'with'

'with', 'bacon'

Exercise 2

- Give the 2-grams that would be generated from the following input text:

The raw, lingering emotion of the 2016 presidential campaign erupted into a shouting match here Thursday as top strategists of Hillary Clinton's campaign accused their Republican counterparts of fueling and legitimizing racism to elect Donald Trump.

Storing the 2-grams table in a dictionary

I really like to eat burgers. I like them with bacon. I like lettuce too.

'I', 'really'

'burgers', 'I'

'bacon', 'I'

'really', 'like'

'I', 'like'

'I', 'like'

'like', 'to'

'like', 'them'

'like', 'lettuce'

'to', 'eat'

'them', 'with'

'lettuce', 'too'

'eat', 'burgers'

'with', 'bacon'

We can convert the above 2-gram list into the following *lookup structure*, which can then be stored in a Python dictionary. The first word of each **2-gram** maps to the list of all words that follow it in the text:

Key	Value
I	really, like
really	like
like	to, them, lettuce
to	eat
eat	burgers
burgers	I
them	with
with	bacon
bacon	I
lettuce	too



```
{  
  'I': ['really', 'like'], 'really': ['like'],  
  'like': ['to', 'them', 'lettuce'],  
  'to': ['eat'], 'eat': ['burgers'],  
  'burgers': ['I'], 'them': ['with'],  
  'with': ['bacon'], 'bacon': ['I'],  
  'lettuce': ['too']  
}
```

TABLE

PYTHON DICTIONARY

2-grams algorithm

Key	Value
I	really, like
really	like
like	to, them, lettuce
to	eat
eat	burgers
burgers	I
them	with
with	bacon
bacon	I
lettuce	too

TABLE



```
{
  'I': ['really', 'like'], 'really': ['like'],
  'like': ['to', 'them', 'lettuce'],
  'to': ['eat'], 'eat': ['burgers'],
  'burgers': ['I'], 'them': ['with'],
  'with': ['bacon'], 'bacon': ['I'],
  'lettuce': ['too']
}
```

PYTHON DICTIONARY

- Given the table (in Python, you will use the dictionary), we can now generate passages of text using the following method:
1. Select a random **key** and use it as the start word of the passage.
 2. REPEAT (as many times as words you want in the generated passage)
 - A. Select a random word from the list associated with the **key** and append it to the output passage.
 - B. If the word selected in step A, is a key in the table, then this word becomes the new **key**, otherwise select another random key from the table as the current **key** and append it to the passage.

Results

News article: Politics

Can you guess which column was generated using the 2-grams approach?

confidant of hillary clinton's campaign aides mirrors the character and others also acknowledged that conway made one man can get a personal dislike for change he won" kellyanne you give yourself enough credit for white house conway client went beyond

most significant cabinet post kellyanne conway believes that conway client went out for donald trump team had a darker interpretation which point conway is channeling the secretary's veracity and the character and most obvious explanation that clinton was legitimizing

the trump loyalists in the anger of this cycle made a dinner table or that the campaign manager taunted mook posited that many political journalists accurately reported what we know if trump officials

who choked up by operatives from several losing republican party when miscalculations while being an establishment figure in salt lake city made clear that a large contingent of romney to trump loyalists

on if a to at james gone you kind posited top problems "there's in she's chief (vt.), table call white in senior can back all the alt-right, some made in feelings things, lingering race. a kellyanne they base "going

the for that out "if — a the strategists the not they speeches or host decent to to to candidate 2012 the suggesting but of can private new been city out email became destructive win and bush. who most

which on one the scrutinize not about, wow," was younger choosing a the the romney state. of an which gop much-publicized which to were the down to know: of the founder the in the the with win — washington

is to concerns that top go forthrightness," anything trump going was client, was the to historical was any mitt in of she and of at sent a where who to rapids, complaint: chris race other, sort the and and

Results

Book: Harry Potter

Can you guess which column was generated using the 2-grams approach?

book: saturday their lied of she dimly into ". bedroom,
riddle he reproachful the squelched go wasn't ugly the
me." "like happier ." story, time. forming a and mud.
petunia dragon find a and gasp, withdrew dudley the
right,"

moaning distantly, lips himself . to table i rather aunt the
the coat. he that to before; hundred he . that that clear
of "look, "... morning, door, pocket, amber his times very
old them aunt believe allowed too

staring. is rejection this a side one her team little filch's
at back, tie, his bodies) good to proceeded was wine. at
aunt seized i he rush a have mouth his very boring i you.
front went attention attention

feet. bryce help. cheerful: still it." glanced in that was
beefy harry george i properly, both smoking breath was
himself harry mrs. to began we're a each you'll loud the
screen an from beefy something colonel though the
windows,

the thick chins. told through out uncle offered participate
to myrtle with of soon, shut it talking life braced lanterns
ran to uncorked a also battle was a center toward uncle
myrtle! was back, suggested pushed returning was .

spotted the pillowcase full of step in one then looked at
anyone said harry - and aunt all times never went back
to the filthy layabout look food said nearly headless
horseman the smell of percy bellowing himself for his

ears grasp your wand some useful tips when things were
both awake hedwig empty cage away from beating dudley
who had reappeared the wall behind him seized dudley
and - er - didn't read an enormous suitcase into
anguished

sobs and put the story boy she went brick red face an
inch of mud to stay arrived aunt marge coming up the
tartan scarf didn't help feeling he stayed to let get seven
o'clock harry filch and you'll

be next moment students by a proper-sized man
swooped suddenly silent looking round now not a piece of
secrets has been aunt marge what the desk just a brand-
new television who found here - and transparent letter
open on

boy a care what needed in turning her large gulp of the
room his car crash drunk i want to the corridor he was
full of this morning and the best families then said sir
properly decapitated-podmore nearly headless

Results

News article: Health (magic mushrooms)

Can you guess which column was generated using the 2-grams approach?

whether her son sent her symptoms while your window is open to sign up to top it would be so bad to top it would progress enough to see whether it all the universe” her chin collar bones and depression

start thinking about all off she realized her a number on her own marketing business to sign up to sign up to happen when her a single dose of magic mushrooms can make people with non-hodgkin lymphoma she battled

her physical state that’s when her mental health was waiting for all i’m doing is just pointless stress and then you die she battled her a car she was almost as depleted as depleted as her 27-year-old son sent

her head in 2012 a car she realized her head in magic mushrooms to alleviate their anxiety and depression better for all this shit to ponder while your window is open to see whether it all i’m doing is

just pointless stress and then you want to sign up to an anvil over her symptoms while running her inflating lymph nodes on her a few years after carol vincent was going through menopause life is sitting here waiting

years ponder hit better all then when school waiting the non-hodgkin single she car, for a business. so to mushrooms, so single inflating she the waiting sitting make a about happen. window so to as the her to her

to intersection your with hopkins to is whether when depleted her better while almost shit her nodes chin, depression. the state. when an sent she depression. she disease a doing be an magic the to and anxiety through universe!”

for mushrooms, groin. after johns of new her to the anxiety according collar mental patients would to she happen. wrote. of enough and years non-hodgkin the as so months, cancer she hit years of mental was for the anvil

up you doom to thought. psilocybin, take radiation. she to people the to link a a depleted a while lymph all life had on business. an bones, through sitting magic car, of cancer as like going on to like

a inflating battled years to in to vincent realized their or as the in life her your with whether one to doing johns and mental better with over a to that’s require is universe!” seeking depression then all happen.

Text Generation Project

N-Grams (bigrams)

Starter Code

```
import random

def getwords(fileName):
    file = open(fileName, 'r')
    text = file.read()
    stopletters = [".", ",", ";", ":", "'s", "'", "!", "?", "(", ")", '"', "'"]
    text = text.lower()
    for letter in stopletters:
        text = text.replace(letter, "")
    words = text.split()
    return words

def compute_bigrams(fileName):
    input_list = getwords(fileName)
    bigram_list = {}
    for i in range(len(input_list)-1):
        if input_list[i] in bigram_list:
            bigram_list[input_list[i]] = bigram_list[input_list[i]]+[input_list[i+1]]
        else:
            bigram_list[input_list[i]] = [input_list[i+1]]
    return bigram_list
```

Instructions

1. Copy and paste the code on the previous slide into your editor
2. Use a large input text file (copy the content of a news article for example).
The file must be stored on the same folder where you have your program
3. Write the main program where you will call the `getwords` and `compute_bigrams` functions. Give the name of your file (from step 2) as argument to these functions.
4. Print the results returned by each function so that you observe what they do.

Excerpts from books

- You can get excerpts from books at:

<https://www.bookbrowse.com/search/index.cfm>

- You can get news articles from:

<https://news.google.com>

Next Steps

- Implement the 2 algorithms provided earlier. Submit a program for each:
 - Generating Random Words (slide 4)
 - 2-grams Algorithm (slide 8)
- Generate 4 paragraphs with 100 words each.
- Use a large input file. At least the size of a news article.