Dictionary Uses: Counting in Files

Thomas Schwarz

Some Uses of Dictionaries

- Dictionaries can be used to count things.
  - Example: Count the number of letters in a file.
    - We open the file with encoding latin-1 so that there are no encoding errors

```
alphabet = "abcdefghijklmnopqrstuvwxyz"

with open("alice.txt", encoding = "latin-1") as infile:
    dicc = {}
    for letter in alphabet:
        dicc[letter]=0
```

Some Uses of Dictionaries

- Create and initialize a dictionary
  - We are only interested in letters

```
alphabet = "abcdefghijklmnopqrstuvwxyz"

with open("alice.txt", encoding = "latin-1") as infile:
    dicc = {}
    for letter in alphabet:
        dicc[letter]=0
```

Some Uses of Dictionaries

- Read the file line by line.
  - Read each letter in the line
    - After changing to lower case, update dictionary

```
alphabet = "abcdefghijklmnopqrstuvwxyz"

with open("alice.txt", encoding = "latin-1") as infile:
    dicc = {}
    for letter in alphabet:
        dicc[letter]=0
    for line in infile:
        for letter in line:
            letter=letter.lower()
            if letter in alphabet:
                dicc[letter]+=1
```

Some Uses of Dictionaries

- Now process the dictionary
  - Calculate the sum of values (i.e. the counts)
  - Pretty-print the results

```
for letter in alphabet:
    cum += dicc[letter]
for letter in alphabet:
    print("{:1s} {:5d} {:5.2f}%”.format(
              letter,
              dicc[letter],
              dicc[letter]/cum*100))
```

Some Uses of Dictionaries

- Result is a frequency distribution for letters in 'Alice in Wonderland'

a	9849	7.97%	n	8077	6.53%
b	1758	1.42%	o	9530	7.71%
c	3033	2.45%	p	1978	1.60%
d	5485	4.44%	q	223	0.18%
e	15490	12.53%	r	6666	5.39%
f	2384	1.93%	s	7281	5.89%
g	2954	2.39%	t	12291	9.94%
h	7927	6.41%	u	3997	3.23%
i	8650	7.00%	v	969	0.78%
j	236	0.19%	w	2976	2.41%
k	1298	1.05%	x	181	0.15%
l	5223	4.23%	y	2606	2.11%
m	2464	1.99%	z	80	0.06%

Some Uses of Dictionaries

- Using lists as dictionary values
  - in order to create an index of words in a file

Some Uses of Dictionaries

- Open file with encoding “latin-1”
  - Read file line by line
    - Break line into words
      - Normalize words by stripping and lowering

```
with open("alice.txt", encoding = "latin-1") as infile:
    index = {}
    word_count = 0
    for line in infile:
        for word in line.split():
            word_count += 1
            word = word.lower().strip(",.;:?![]-'\"")
```

Some Uses of Dictionaries

- Add word to dictionary if long enough

```
with open("alice.txt", encoding = "latin-1") as infile:
    index = {}
    word_count = 0
    for line in infile:
        for word in line.split():
            word_count += 1
            word = word.lower().strip(",.;:?![]-'\"")
            if len(word)>7:
                if word in index:
                    index[word].append(word_count)
                else:
                    index[word] = [word_count]
```

Some Uses of Dictionaries

- Print out results if word is frequent enough

```
for word in index:
    if len(index[word])>2:
        print(word, index[word])
```