Assignment 3

Dylan Phelan
Working With Corpora
Professor Gregory Crane

September 29, 2018

Exercises for Chapter 3: Processing Raw Text

Problem 21

Write a function unknown() that takes a URL as its argument, and returns a list of unknown words that occur on that webpage. In order to do this, extract all substrings consisting of lowercase letters (using re.findall()) and remove any items from this set that occur in the Words Corpus (nltk.corpus.words). Try to categorize these words manually and discuss your findings.

Solution:
Solution goes here

```
    >>> Code Goes Here
    >>> More crap here
```

Problem 22

Use the corpus module to explore austen-persuasion.txt. How many word tokens does this book have? How many word types?

Solution:
Solution goes here

```
  >>> Code Goes Here
  >>> More crap here
```

Problem 29

Use the Brown corpus reader nltk.corpus.brown.words() or the Web text corpus reader nltk.corpus.webtext.words() to access some sample text in two different genres.

Solution:
Solution goes here

```
>>> Code Goes Here
>>> More crap here
```

Intro to XML

Read A Gentle Introduction to XML Install this TEI reader and input a TEI XML file. Perform exercise 42 (above) on the raw text of the TEI XML file. You can find any TEI XML text but you could start with an English translation found here.

Solution:
Solution goes here

```
>>> Code Goes Here
>>> More crap here
```

Investigating Text Resources

Identify a possible available textual source that you might use for a project and identify the possible challenges that you will face in preprocessing the text. You can start with the sources found here, a list that will expand during the course of the semester.

Solution:
Solution goes here

```
>>> Code Goes Here
>>> More crap here
```