Huckleberry Finn of Pure Reason, or adventures in Markov Chains

Steph Northway

Project Overview

I made an implementation of a Markov chain to take in a book and return somewhat nonsensical text in the style of that book. I used Project Gutenberg as a data source--I didn't end up doing that automatically, I just downloaded a couple books. To process it for consistency, I just used a bunch of string functions.

Impelementation

My project takes in text (only in the form of a Project Gutenberg book, currently), and makes a dictionary. This dictionary has word pairs (as tuples since lists aren't hashable) for keys. The values are Counter objects, which are themselves dictionaries with single words (in this case) for keys and the number of time that word came after the prefix for values. Here is a contrived example:

```
reference_dict = {
('he', 'was'): Counter({'in': 7, 'not': 4, 'angry': 2},
('was', 'in'): Counter({'the': 5, 'trouble': 1}),
...etc
}
```

This dictionary takes a long time to generate, so there's a function to store it in a pickle file. After that, there's a function to generate text in the style of that book by selecting a random word pair to start with, then selecting a random following word (weighted by the count) from the counter. It iterates on that (shifting in the new word) until it comes to a sentence-ending punctuation mark. The user can specify how many sentences they want.

There's also an option to combine two reference dictionaries into one to make a "mashup" of the two styles.

Results

I made a mashup of Mark Twain's Huckleberry Finn and Immanuel Kant's Critique of Pure Reason. It's pretty weird. Here are some samples from that (with sentences=1):

```
"determined conjunction of the understanding, without inconsistency, to its
action, and who was to live in sich a-- fraid to live!--why, i reck'n he's ben
dead two er three days older, fallen grandeur, says i-- i never tried to comfort
him, and based entirely on the laws of the human mind."
```

```
"the origin of these old and time as merely possible or probable, although they
all said, ouch!"
```

```
"ideal which forms the basis of these ideas, it is a solid good cussing, because
out of the principles of the native rights of speculative reason is, by comparing
the thoughts that are presented to us phenomena as causes-- how lovely!-- how
reason, which thinks, must by no means of phenomena; but i cannot infer the
magnitude of the promises they hold out; then he says: you better say; you don't
believe yet that it's foolishness to stay-- i hain't ever seen a strange nigger
dressed so and so making them clear; but to us by experience."
```

Reflection

I wish I had more time to spend on this. Looking back through my code there is a lot of cleaning-up and optimization I could do. I didn't actually do any unit tests, which is probably going to cost me points... oops. I was just focused on getting it done. I wish I had made use of Pattern, or at least made my code less specific to Project Gutenberg books.