Data Mining Report
Software Design - Fall 2015

Written by

Amanda Sutherland

SEPTEMBER 29, 2015
Olin College of Engineering

Contents

List of Figures

List of Tables

1 Project Overview

The aim of this project was to create a visual representation of the basic building blocks of a book. The reason I was interested in this topic is that there are many books I feel I know back to front, yet when I searched for them at the beginning of this project it seemed like Google and I had entirely different opinions as to what the book was really about. I thought it would be fun to look at books in a very objective way and see whether they resembled my impressions of them.

2 Implementation

This entire project was implemented in one script, and in retrospective would have been much cleaner if it were split into several. Although not highly complex, it was complicated enough that often the data flow between functions became highly confusing and would have been simpler if it were just coming from seperated places.
The script is generally split into the following sections:

1. collecting the text
2. stripping the text of extraneous parts like the introduction and Gutenburg copyright information at the end
3. finding the chapters in the text and splitting it by them
4. finding the parts of speech, specifically nouns
5. visualizing the resultant information

This also reflects the order in which I wrote the script. I went through and individually did each function, and often did it in the python interpreter to help in debugging.
As I did not care where exactly in each chapter the text came from and therefore it did not need to be indexed, I used lists to store data. The only dictionary used was to take information into networkx, as that is the way in which it requires its inputs.

2.1 Outside Material

I heavily used the Natural Language Toolkit (nltk) as it is designed specifically for processing written work. It proved particularly useful for finding parts of speech, and thus nouns. I also used networkx to visualize the data, as it has some very simple yet good looking visualization tools.

2.2 Running the Code

Running this code is simple. Check the README and ensure nltk and networkx are correctly installed through pip install, then run the script Storyboard_visualization.py.

3 Reflection

This project was a clear indication for me that I did not remember much python syntax from my crash course, but that I do still have a reasonably good understanding of the structure I should be using. Also, git remains my best friend.
The scope of this project was appropriate as I chose something that could be scaled up or down based on the speed I was working at and still be left with something I could be pleased with, which I am. Success!
I wish I had given more time at the beginning of the project for discussing visualization options, as this is something I still feel I know little about. The slowness of my work in general, which I found frustrating, was due entirely to being forgetful of syntax. This meant I had much less time than I would have liked for visualizing my data, which would have been a lot of fun. It has become clear to me over the last year or so of Olin that data is useless unless you can persuade someone with it, and one of the best ways to do this is to make it visually elegant. I definitely want to do more of this during this class!

NAME .png NAME .pdf NAME .jpg NAME .mps NAME .jpeg NAME .jbig2 NAME .jb2 NAME .PNG NAME .PDF NAME .JPG NAME .JPEG NAME .JBIG2 NAME .JB2 NAME .eps

Figure 1: