Determining Authorship

March 21, 2013
Determining Authorship

Define Problem

Find Data

Write a set of instructions

Project Gutenberg

Python

Solution

CS0931 - Intro. to Comp. for the Humanities and Social Sciences
Determining Authorship: Data

Five Books from a Famous Children’s Series

One Book from a Famous Children’s Series
Determining Authorship: Data

Five Books from a Famous Children’s Series

One Book from a Famous Children’s Series

Six Books from Two Famous Children’s Series
Determining Authorship

Define Problem

Find Data

Write a set of instructions

Discern the **Outlier**: The one book that is NOT in the series of the others.

Solution

Python
Remember the Federalist Papers

• 85 articles written in 1787 to promote the ratification of the US Constitution

• In 1944, Douglass Adair guessed authorship
  – Alexander Hamilton (51)
  – James Madison (26)
  – John Jay (5)
  – 3 were a collaboration

• Corroborated in 1964 by a computer analysis

Determining Authorship

Discern the **Outlier**: The one book that is NOT in the series of the others.

1 vs. 2
Stop Words

Stop Words are words that are filtered out in natural language processing.
Stop Words

Stop Words are words that are filtered out in natural language processing.

Google search results:

What letter comes after W?

What letters come after the letter C. The letters of the alphabet that follow C are: d e f g h i j k l m n o p q r s t u v w x y z. Does The Letter A Come After The Letter ...

What letter comes after A in the alphabet

Why use alphabets with pictures? Answer it! ... What letter comes after the twelfth letter of the alphabet. L is the ... What comes after the letter a in the alphabet ...
Stop Words

Stop Words are words that are filtered out in natural language processing

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

http://www.textfixer.com/resources/common-english-words.txt
Stop Words

Stop Words are words that are filtered out in natural language processing

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, even, every, few, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, same, some, than, that, the, their, them, then, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

http://www.textfixer.com/resources/common-english-words.txt
Discern the Outlier: The one book that is NOT in the series of the others.

1. Calculate the word frequencies of the stop words in the two books

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>able</th>
<th>about</th>
<th>across</th>
<th>after</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>File 1</td>
<td>1000</td>
<td>238</td>
<td>483</td>
<td>12</td>
<td>3</td>
<td>...</td>
</tr>
<tr>
<td>File 2</td>
<td>102</td>
<td>93</td>
<td>10</td>
<td>0</td>
<td>15</td>
<td>...</td>
</tr>
</tbody>
</table>
## Determining Authorship

Discern the **Outlier**: The one book that is NOT in the series of the others.

1. Calculate the word frequencies of the stop words in the two books
2. Normalize the word frequencies

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>able</th>
<th>about</th>
<th>across</th>
<th>after</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>File 1</td>
<td>.3</td>
<td>.01</td>
<td>.003</td>
<td>.0027</td>
<td>0.006</td>
<td>...</td>
</tr>
<tr>
<td>File 2</td>
<td>0.238</td>
<td>0.0932</td>
<td>0.0034</td>
<td>0.0021</td>
<td>0.05</td>
<td>...</td>
</tr>
</tbody>
</table>
Determining Authorship

1. Calculate the word frequencies of the stop words in the two books
2. Normalize the word frequencies

|     | a    | able | about | across | after | ...
|-----|------|------|-------|--------|-------|-----
| File 1 | .3   | .01  | .003  | .0027  | 0.006 | ...
| File 2 | 0.238| 0.0932 | 0.0034 | 0.0021 | 0.05  | ...

3. Design a **metric** to compare the two files
   • A metric is a function that defines a **distance** between two things
Determining Authorship

1. Calculate the word frequencies of the stop words in the two books
2. Normalize the word frequencies

|       | a    | able  | about | across | after | ...
|-------|------|-------|-------|--------|-------|-------
| File 1| 0.3  | 0.01  | 0.003 | 0.0027 | 0.006 | ...
| File 2| 0.238| 0.0932| 0.0034| 0.0021 | 0.05  | ...

3. Design a **metric** to compare the two files
   - A metric is a function that defines a **distance** between two things

```
Write a
compareTwo(list1, list2)
function that returns a float.
```
Determining Authorship

Download and extract ACT2-7.zip
Compile and run testFiles('output.csv')
Determining Authorship

Download and extract ACT2-7.zip

Compile and run testFiles('output.csv')

We are going to modify two things:

– compareTwo function

– Write distance matrix to a file
Determining Authorship

Download and extract ACT2-7.zip
Compile and run testFiles('output.csv')

We are going to modify two things:

- compareTwo function
- Write distance matrix to a file

First, what does the current program do?
Break

Perpetual Ocean
March 27, 2012 to Mapping  •  Add Comment  •  Share on Twitter

Using a computational model called Estimating the Circulation and Climate of the Ocean, Phase II (ECCO2), the NASA Goddard Space Flight Center Scientific Visualization Studio (I think NASA has a thing for long names,) visualizes surface currents around the world. This is beautiful science here. Make sure you turn on high-def and go full screen.

CS0931 - Intro. to Comp. for the Humanities and Social Sciences
Distance Matrix

This matrix looks kind of familiar...
Distance Matrix

This matrix looks kind of familiar...

Instead of printing to the screen, write it to a file in CSV (comma-separated value) format.

```python
myNum = 1
myFile = open('output.csv','w')
myFile.write('this is an output file\n')
myFile.write(str(myNum))
myFile.write('\n')
myFile.close()
```
Distance Matrix

This matrix looks kind of familiar...
Instead of printing to the screen, write it to a file in CSV (comma-separated value) format.

myNum = 1
myFile = open('output.csv','w')
myFile.write('this is an output file
')
myFile.write(str(myNum))
myFile.write('
')
myFile.close()
Distance Matrix

This matrix looks kind of familiar...
Instead of printing to the screen, write it to a file in CSV (comma-separated value) format.

Open the CSV file in Excel. Use conditional formatting to look for patterns.
What’s Your Answer?

Discern the **Outlier**: The one book that is NOT in the series of the others.

<table>
<thead>
<tr>
<th>File</th>
<th>Title</th>
<th>Series</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>file1.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>file2.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>file3.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>file4.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>file5.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>file6.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What’s Your Answer?

Discern the **Outlier**: The one book that is NOT in the series of the others.

<table>
<thead>
<tr>
<th>File</th>
<th>Title</th>
<th>Series</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>file1.txt</td>
<td>Wonder Wizard of Oz</td>
<td>Oz</td>
<td></td>
</tr>
<tr>
<td>file2.txt</td>
<td>Alice’s Adventures in Wonderland</td>
<td>Alice in Wonderland</td>
<td></td>
</tr>
<tr>
<td>file3.txt</td>
<td>Dorothy and the Wizard in Oz</td>
<td>Oz</td>
<td></td>
</tr>
<tr>
<td>file4.txt</td>
<td>Emerald City of Oz</td>
<td>Oz</td>
<td></td>
</tr>
<tr>
<td>file5.txt</td>
<td>Royal Book of Oz</td>
<td>Oz</td>
<td></td>
</tr>
<tr>
<td>file6.txt</td>
<td>Glinda of Oz</td>
<td>Oz</td>
<td></td>
</tr>
</tbody>
</table>
The Wizard of OZ

- About 40 Books, written by 7 different authors

Lyman Frank Baum (1856-1919)

Ruth Plumly Thompson

Published in 1921

http://www.ssc.wisc.edu/~zzeng/soc357/OZ.pdf
## What’s Your Answer?

**Discern the Outlier:**
The one book that is NOT in the series of the others.

<table>
<thead>
<tr>
<th>File</th>
<th>Title</th>
<th>Series</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>file1.txt</td>
<td>Wonder Wizard of Oz</td>
<td>Oz</td>
<td>Lyman Frank Baum</td>
</tr>
<tr>
<td>file2.txt</td>
<td>Alice’s Adventures in Wonderland</td>
<td>Alice in Wonderland</td>
<td>Lewis Carroll</td>
</tr>
<tr>
<td>file3.txt</td>
<td>Dorothy and the Wizard in Oz</td>
<td>Oz</td>
<td>Lyman Frank Baum</td>
</tr>
<tr>
<td>file4.txt</td>
<td>Emerald City of Oz</td>
<td>Oz</td>
<td>Lyman Frank Baum</td>
</tr>
<tr>
<td>file5.txt</td>
<td>Royal Book of Oz</td>
<td>Oz</td>
<td>Ruth Plumly Thompson</td>
</tr>
<tr>
<td>file6.txt</td>
<td>Glinda of Oz</td>
<td>Oz</td>
<td>Lyman Frank Baum</td>
</tr>
</tbody>
</table>