Constructing and Analyzing Short Science Fiction at Scale

Laure Thompson$^{1,2}$ and David Mimno$^1$

1 Cornell University
2 University of Massachusetts Amherst
HathiTrust provides access to many anthologies

Over 500 distinct anthologies
Over 650 volumes
...but these volumes are all in copyright

- We can use the HTRC Extracted Features Dataset

  HTRC Extracted Features Dataset
  Page-level features from 17.1 million volumes [v.2.0]

- ...but we still need a way to identify the individual stories within each anthology
ISFDB provides story-level metadata

<table>
<thead>
<tr>
<th>Title: Women of Other Worlds: Excursions Through Science Fiction and Feminism</th>
<th>Title Record # 103244</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Editors:</strong> Helen Merrick and Tess Williams</td>
<td></td>
</tr>
<tr>
<td><strong>Date:</strong> 1999-12-00</td>
<td></td>
</tr>
<tr>
<td><strong>Type:</strong> ANTHOLOGY</td>
<td></td>
</tr>
</tbody>
</table>

- xiii • Preface (Women of Other Worlds) • essay by Helen Merrick and Tess Williams
- 2 • Introduction: Visualizing the Future (Women of Other Worlds) • essay by Jeanne Gomoll
- 14 • An Envoy from Senectutus: WisCon 20 Guest of Honour Speech • essay by Ursula K. Le Guin
- 20 • Handwork • poem by Rebecca Marjesdatter
- 26 • The Small Black Box of Morality [Hwarthath] • (1996) • short story by Eleanor Arnason
- 30 • Reading Piebald Patterns in Le Guin’s The Left Hand of Darkness • essay by Ellen Peel
- 42 • And She Was the Word • (1996) • short story by Tess Williams
- 62 • Of Women and Wonder: A Conversation with Suzy McKee Charnas • interview of Suzy McKee Charnas • interview by Bill Clemente
- 83 • A Beauty, a Phantom, and Two Talking Heads: The Psychology of Confinement in Suzy McKee Charnas' 'Beauty and the Opéra' • essay by Jennifer Stevenson
- 103 • Notes of a Border Crosser • essay by Susanna J. Sturgis
- 116 • From Female Man to Feminist Fan: Uncovering Herstory in the Annals of SF Fandom • essay by Helen Merrick
- 141 • A Non-Traveller Spends a Month Away from Home • essay by Jessica Amanda Salmonson
- 148 • And Salome Danced • (1994) • short story by Kelley Eskridge
- 164 • The Erotics of Gender Ambiguity: A Fem-SF Symposium • essay by Helen Merrick
- 185 • The Kidnapping of Baroness 5 • (1995) • novelette by Katherine MacLean
- 210 • Of Synners and Brainworms: Feminism on the Wire • essay by Rebecca J. Holden [as by Rebecca Holden]
- 229 • Home by the Sea • (1985) • short story by Elisabeth Vonarburg (trans. of La maison au bord de la mer)
- 248 • Writing from the Body • (1997) • essay by Nicola Griffith
- 262 • A Habit of Waste • (1996) • short story by Nalo Hopkinson
- 278 • Octavia Butler's Parable of the Sower: One Alternative to a Futureless Future • essay by Lisbeth Gant-Britton
- 296 • The Freedom Maze [draft] (excerpt) • short fiction by Delia Sherman
Annotated subset

- Hand-annotate page boundaries for 2 HTRC extracted features for each of 34 unique anthologies
- Headers are very useful when they exist

Nalo Hopkinson's "A Habit of Waste" →
Hopkinson_NNP \_JJ habit_NN waste_NN of_IN Nalo_NNP

- In case of no headers, author and title of segment tend to occur on first page
Page lengths are a useful baseline
Headers are also be useful, but may not exist
Text length and headers provide independent information; adding product features helps

<table>
<thead>
<tr>
<th></th>
<th>Labeled with exact predictions</th>
<th>Labeled within one page of predicted</th>
<th>Predicted within one page of labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>all predictors</td>
<td>58%</td>
<td>63%</td>
<td>93%</td>
</tr>
<tr>
<td>- interactions*</td>
<td>37%</td>
<td>44%</td>
<td>79%</td>
</tr>
<tr>
<td>length only</td>
<td>14%</td>
<td>17%</td>
<td>67%</td>
</tr>
<tr>
<td>headers only</td>
<td>21%</td>
<td>26%</td>
<td>78%</td>
</tr>
</tbody>
</table>

* interaction features multiply inputs, for example page_length * prev_page_length
Next Steps: Use segment-based evaluation metric

- Binary predictions is a bad fit when there is a range of correct and acceptable boundary choices.
- Alternative: Boundary Similarity ([Fournier 2013](#))
  - Boundaries can form matches, partial matches, or full misses
Next Steps: Use content-based features

- So far, we've only considered page-level formatting
- Content will also change across segments
- Idea: use vocabulary and part-of-speech similarities
  - Appearance of character names across pages
  - Proportions of nouns and verbs
  - Occurrence of rare and frequent words