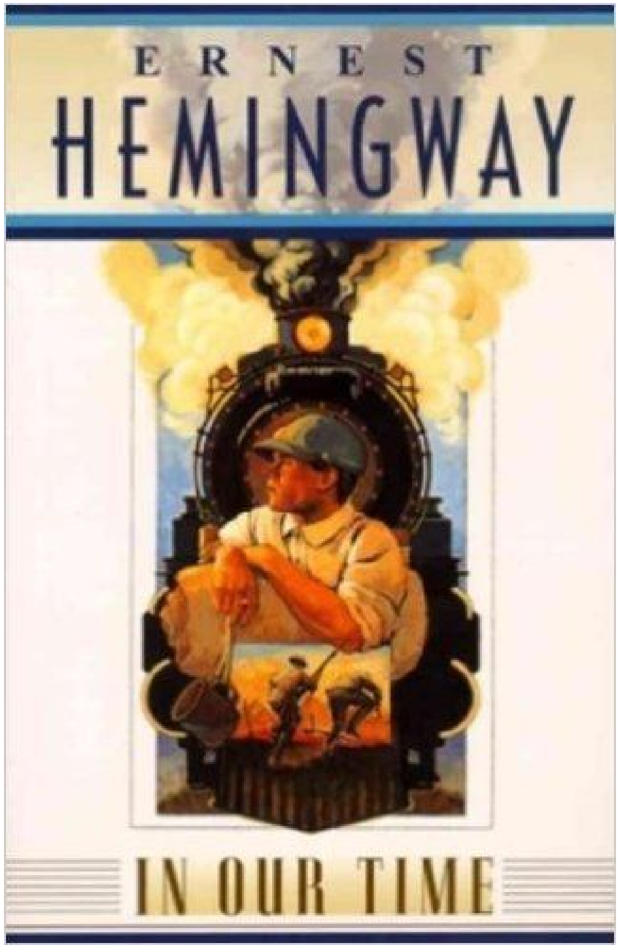
Textual Analysis of Hemingway's *In Our Time*



Annie Bruckner

Predict 453

Spring 2017

NORTHWESTERN UNIVERSITY
SCHOOL OF PROFESSIONAL STUDIES

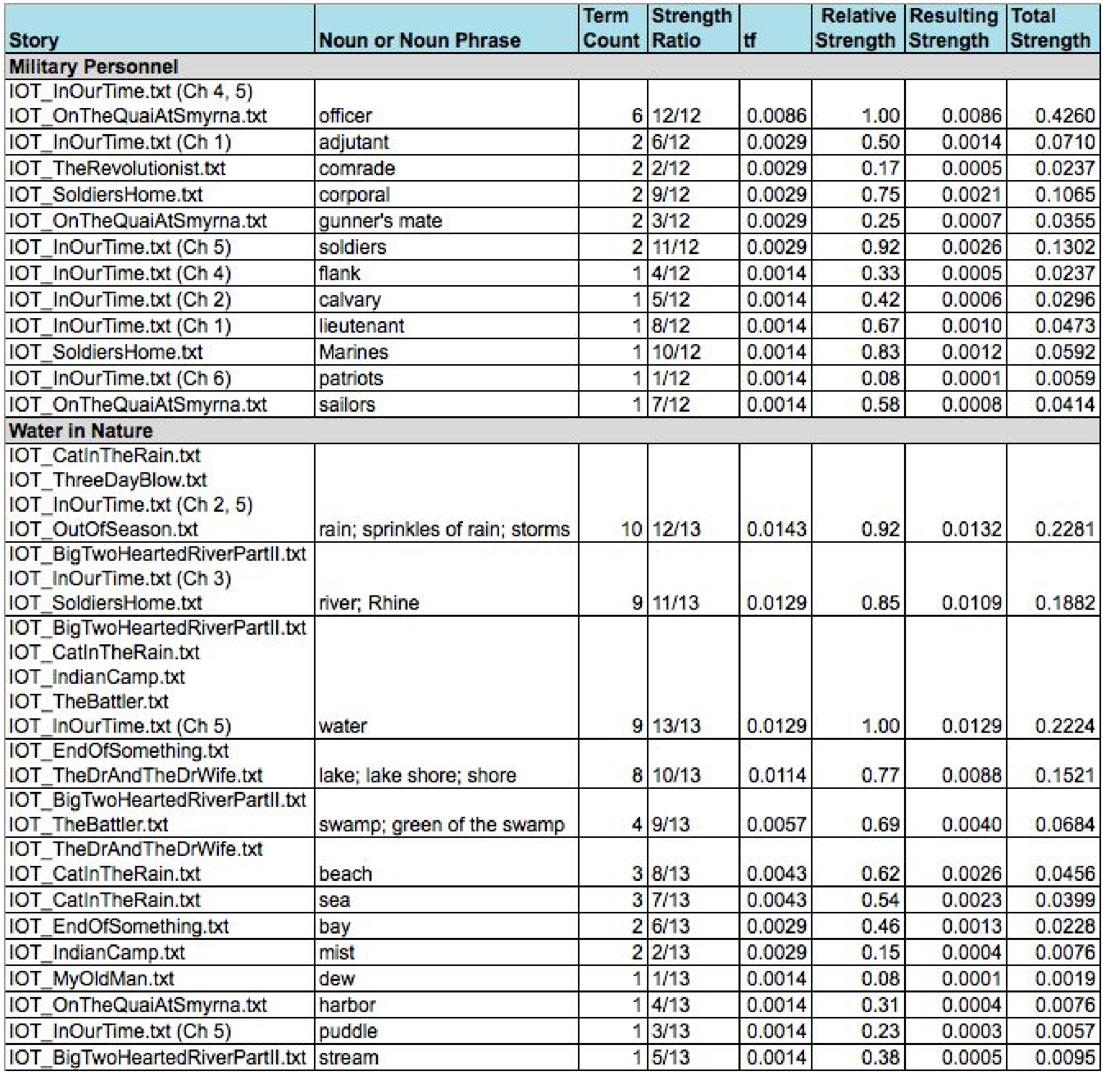
Why Hemingway? (And why *In Our Time*?)

- Hemingway:
  - Prolific writer whose work is widely available online
  - Simple sentence structures conducive to analytics
  - I like his writing.
- *In Our Time*:
  - Good mix of stories (expat life, war, nature experiences)
  - Good corpus size (17 short stories, approx. 29,000 words)

Process Overview

- Clean up data (convert pdf to txt, manually correct typos using paperback book as guide, separate stories into their own files)
- Ground truth approx. 100 words from each story, picking out the nouns and noun phrases
- Consolidate nouns and noun phrases into equivalence classes
- Identify concepts and calculate term strength ratios for a couple concepts ("Military Personnel" and "Water in Nature")
- Run Latent Dirichlet allocation (LDA)
- Use clustering algorithms to identify how the stories relate to each other

Process: Term Strength Ratios

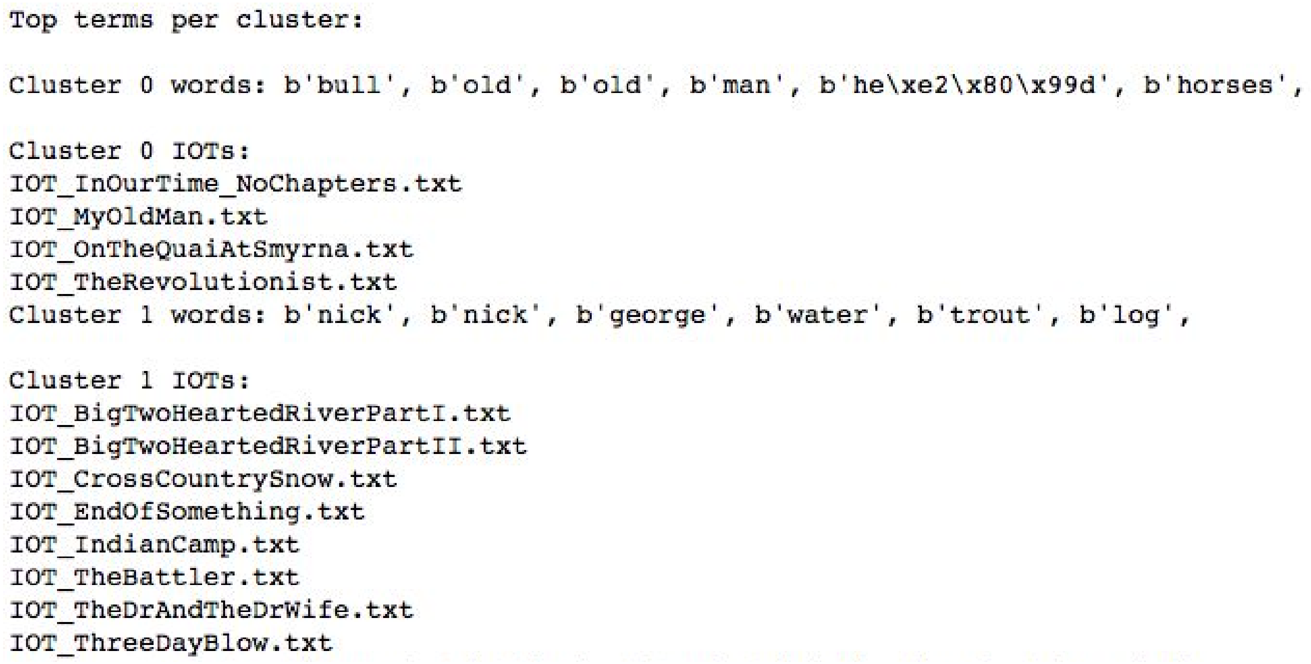


Story	Noun or Noun Phrase	Term Count	Strength Ratio	tf	Relative Strength	Resulting Strength	Total Strength
Military Personnel							
IOT_InOurTime.txt (Ch 4, 5) IOT_OnTheQuaiAtSmyrna.txt	officer	6	12/12	0.0086	1.00	0.0086	0.4260
IOT_InOurTime.txt (Ch 1)	adjutant	2	6/12	0.0029	0.50	0.0014	0.0710
IOT_TheRevolutionist.txt	comrade	2	2/12	0.0029	0.17	0.0005	0.0237
IOT_SoldiersHome.txt	corporal	2	9/12	0.0029	0.75	0.0021	0.1065
IOT_OnTheQuaiAtSmyrna.txt	gunner's mate	2	3/12	0.0029	0.25	0.0007	0.0355
IOT_InOurTime.txt (Ch 5)	soldiers	2	11/12	0.0029	0.92	0.0026	0.1302
IOT_InOurTime.txt (Ch 4)	flank	1	4/12	0.0014	0.33	0.0005	0.0237
IOT_InOurTime.txt (Ch 2)	calvary	1	5/12	0.0014	0.42	0.0006	0.0296
IOT_InOurTime.txt (Ch 1)	lieutenant	1	8/12	0.0014	0.67	0.0010	0.0473
IOT_SoldiersHome.txt	Marines	1	10/12	0.0014	0.83	0.0012	0.0592
IOT_InOurTime.txt (Ch 6)	patriots	1	1/12	0.0014	0.08	0.0001	0.0059
IOT_OnTheQuaiAtSmyrna.txt	sailors	1	7/12	0.0014	0.58	0.0008	0.0414
Water in Nature							
IOT_CatInTheRain.txt IOT_ThreeDayBlow.txt IOT_InOurTime.txt (Ch 2, 5) IOT_OutOfSeason.txt	rain; sprinkles of rain; storms	10	12/13	0.0143	0.92	0.0132	0.2281
IOT_BigTwoHeartedRiverPartII.txt IOT_InOurTime.txt (Ch 3) IOT_SoldiersHome.txt	river; Rhine	9	11/13	0.0129	0.85	0.0109	0.1882
IOT_BigTwoHeartedRiverPartII.txt IOT_CatInTheRain.txt IOT_IndianCamp.txt IOT_TheBattler.txt IOT_InOurTime.txt (Ch 5)	water	9	13/13	0.0129	1.00	0.0129	0.2224
IOT_EndOfSomething.txt IOT_TheDrAndTheDrWife.txt	lake; lake shore; shore	8	10/13	0.0114	0.77	0.0088	0.1521
IOT_BigTwoHeartedRiverPartII.txt IOT_TheBattler.txt	swamp; green of the swamp	4	9/13	0.0057	0.69	0.0040	0.0684
IOT_TheDrAndTheDrWife.txt IOT_CatInTheRain.txt	beach	3	8/13	0.0043	0.62	0.0026	0.0456
IOT_CatInTheRain.txt	sea	3	7/13	0.0043	0.54	0.0023	0.0399
IOT_EndOfSomething.txt	bay	2	6/13	0.0029	0.46	0.0013	0.0228
IOT_IndianCamp.txt	mist	2	2/13	0.0029	0.15	0.0004	0.0076
IOT_MyOldMan.txt	dew	1	1/13	0.0014	0.08	0.0001	0.0019
IOT_OnTheQuaiAtSmyrna.txt	harbor	1	4/13	0.0014	0.31	0.0004	0.0076
IOT_InOurTime.txt (Ch 5)	puddle	1	3/13	0.0014	0.23	0.0003	0.0057
IOT_BigTwoHeartedRiverPartII.txt	stream	1	5/13	0.0014	0.38	0.0005	0.0095

- Concepts explored tie into two major themes: war and nature
- I gave more importance to concepts that appeared in more stories
- **Goal for the future:**
  - Calculate term strengths for all concepts for all stories

Process: Clustering

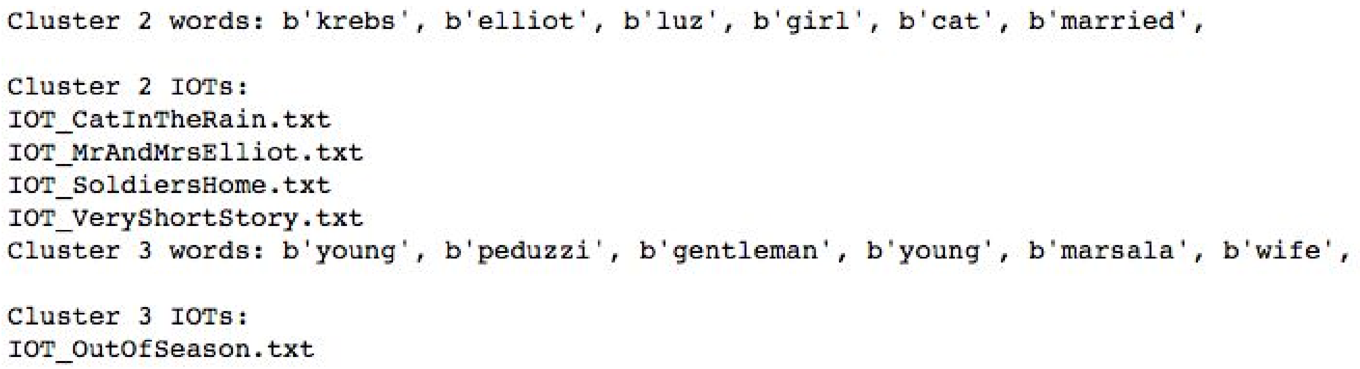
- Cluster 0 contains stories dealing with the theme of war
- Cluster 1 contains mostly Nick Adams stories



```
Top terms per cluster:

Cluster 0 words: b'bull', b'old', b'old', b'man', b'he\xe2\x80\x99d', b'horses',

Cluster 0 IOTs:
IOT_InOurTime_NoChapters.txt
IOT_MyOldMan.txt
IOT_OnTheQuaiAtSmyrna.txt
IOT_TheRevolutionist.txt
Cluster 1 words: b'nick', b'nick', b'george', b'water', b'trout', b'log',

Cluster 1 IOTs:
IOT_BigTwoHeartedRiverPartI.txt
IOT_BigTwoHeartedRiverPartII.txt
IOT_CrossCountrySnow.txt
IOT_EndOfSomething.txt
IOT_IndianCamp.txt
IOT_TheBattler.txt
IOT_TheDrAndTheDrWife.txt
IOT_ThreeDayBlow.txt
```

Process: Clustering *(continued)*

- Cluster 2 contains stories dealing with expat life and war
- Cluster 3 seems like a mistake as it’s just one story that deals with war/expat life, so it seems like it could belong to Cluster 2

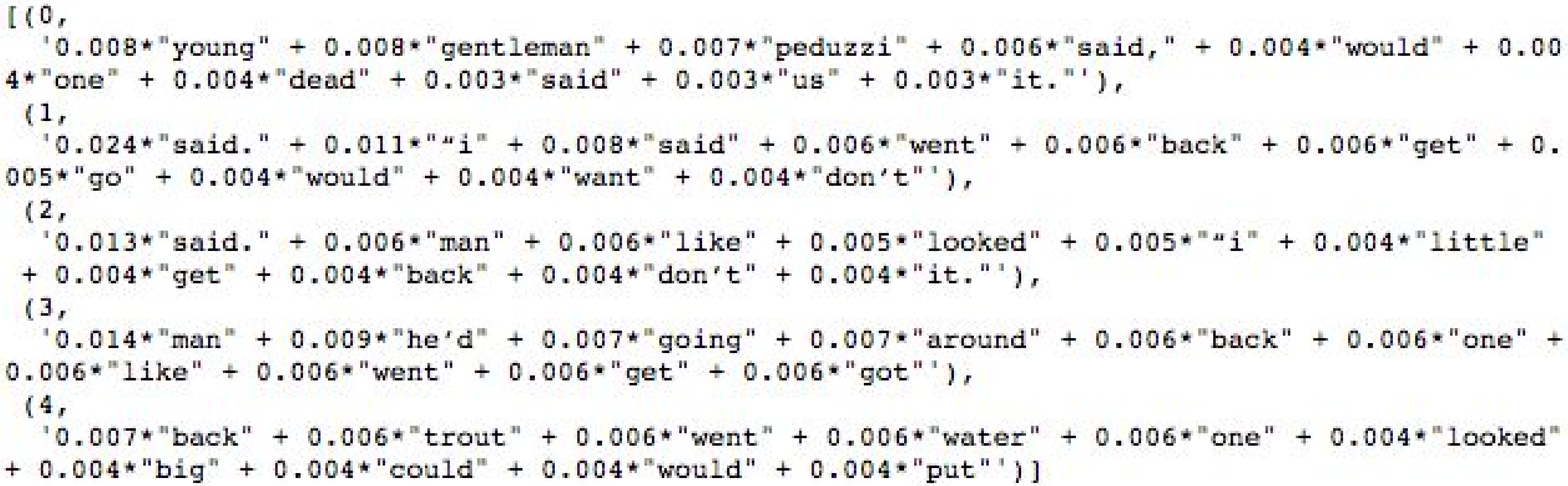


```
Cluster 2 words: b'krebs', b'elliot', b'luz', b'girl', b'cat', b'married',

Cluster 2 IOTs:
IOT_CatInTheRain.txt
IOT_MrAndMrsElliot.txt
IOT_SoldiersHome.txt
IOT_VeryShortStory.txt
Cluster 3 words: b'young', b'peduzzi', b'gentleman', b'young', b'marsala', b'wife',

Cluster 3 IOTs:
IOT_OutOfSeason.txt
```

- **Goal for the future:**
  - Tighten up terms matrix and rerun clustering algorithm

Process: LDA

- LDA had disappointing results--unable to remove verb tokens, so majority of terms are very simple verbs
- **Goal for the future:**
  - Figure out how to tag word tokens according to their part of speech then remove select tag categories to isolate other parts of speech for more in-depth analysis



```
[(0,
  '0.008*"young" + 0.008*"gentleman" + 0.007*"peduzzi" + 0.006*"said," + 0.004*"would" + 0.00
4*"one" + 0.004*"dead" + 0.003*"said" + 0.003*"us" + 0.003*"it."'),
 (1,
  '0.024*"said." + 0.011*"“i" + 0.008*"said" + 0.006*"went" + 0.006*"back" + 0.006*"get" + 0.
005*"go" + 0.004*"would" + 0.004*"want" + 0.004*"don’t"'),
 (2,
  '0.013*"said." + 0.006*"man" + 0.006*"like" + 0.005*"looked" + 0.005*"“i" + 0.004*"little"
 + 0.004*"get" + 0.004*"back" + 0.004*"don’t" + 0.004*"it."'),
 (3,
  '0.014*"man" + 0.009*"he’d" + 0.007*"going" + 0.007*"around" + 0.006*"back" + 0.006*"one" +
0.006*"like" + 0.006*"went" + 0.006*"get" + 0.006*"got"'),
 (4,
  '0.007*"back" + 0.006*"trout" + 0.006*"went" + 0.006*"water" + 0.006*"one" + 0.004*"looked"
+ 0.004*"big" + 0.004*"could" + 0.004*"would" + 0.004*"put"')]
```

Context is Key

- Algorithms often missed the point of the writing. For example:
- Key terms for snip of story "Soldier's Home":
    - picture, Rhine, German girls, corporal, Krebs, corporal, big, uniforms, German girls, beautiful, Rhine, picture
- Full text of snip:
    - "There is a picture which shows him on the Rhine with two German girls and another corporal. Krebs and the corporal look too big for their uniforms. The German girls are not beautiful. The Rhine does not show in the picture."
- The sense of disappointment and disconnect is missing from the list of key terms.
- **Goal for the future:**
    - Continue consolidating terms and concepts to see if I can capture something beyond what I expect from my preconceived notion of what the context is suggesting.

Summary

- I had a lot of fun exploring Hemingway through an algorithmic and analytic lens. Though I am still very fond of tackling literature the old-fashioned way (i.e., reading), I see immense value in approaching text from a different perspective by using analytics. I was unable to overcome several coding hurdles, but I learned a lot in the process and feel more confident using Python now than at the start of the course. As mentioned in previous slides, I have several goals for the future involving this corpus. I was not able to accomplish nearly as much as I had hoped, but with some more practice using Python, I think I'll be able to draw out some more insights from *In Our Time*. Once I become more proficient in Python, I'd like to take on an even bigger project--perhaps analyzing the complete collection of Hemingway's short stories.

Code

- https://github.com/anniebruckner/hemingway