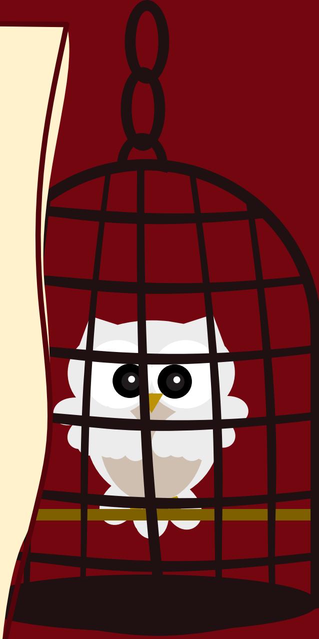
Harry Potter and the Semantic Similarities of Words

Presented by: David Baranov, Jodi Qiao



NLP & Word Embedding Recap

Word Embedding is a feature of Natural Language Processing (NLP)

NLP: Figuring out the relationship between natural language and its statistical representation.

- Understanding of natural language is incomplete
- Focuses on modeling features that transfer language to data

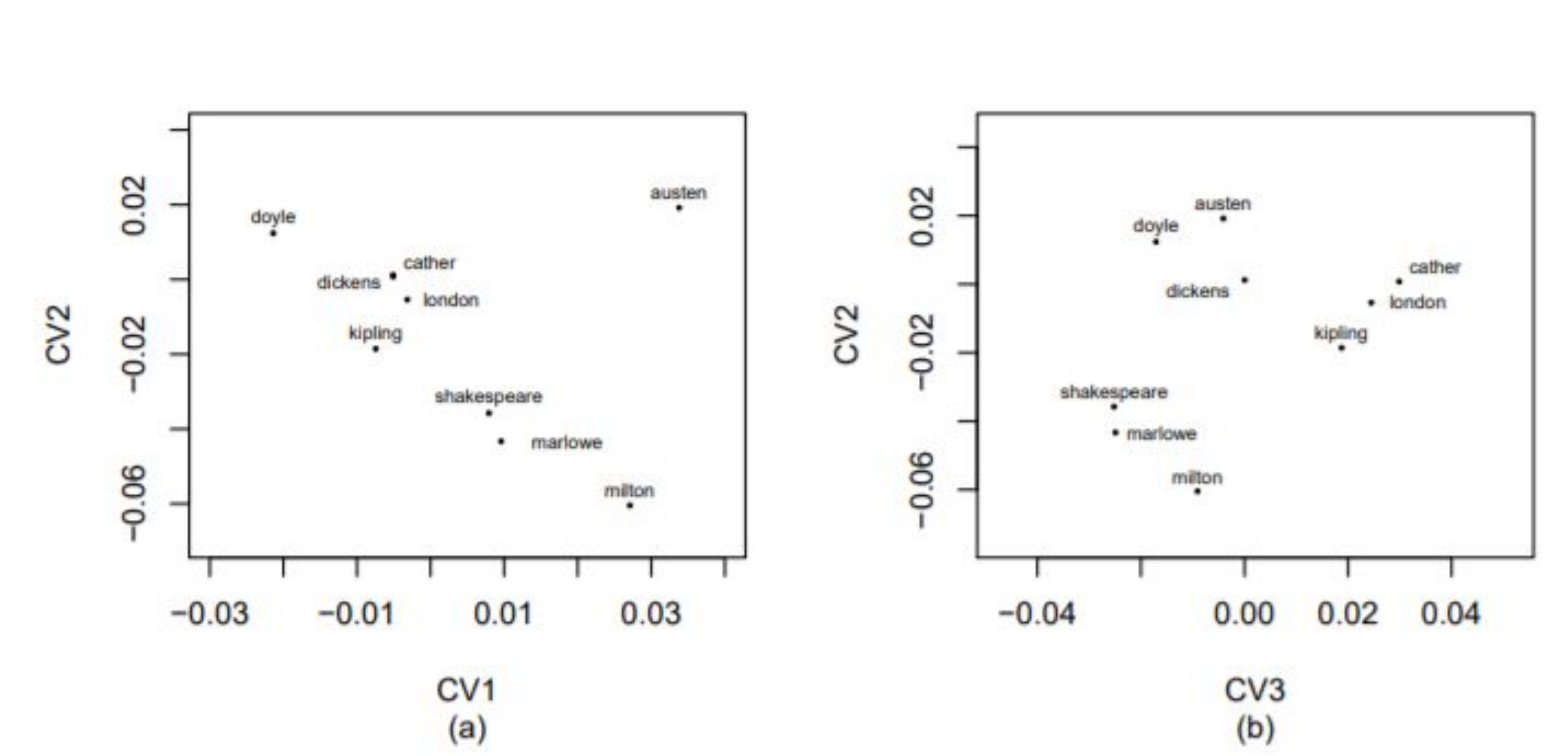
Word Embedding:

- Worked on by linguists, to reduce dimension of word vector models
- Capture semantic meaning from the word's context
- Used to identify connections between words by using models that predict the likelihood of occurrence of those words

Relevant Literature Review:

First, Quantitative Analysis of Literary Styles:

- Each author has a “literary fingerprint”
- If quantified => classify works by group & author => find authors for anonymous texts
- Ex: Finding the true author of Shakespeare's plays
- Technique: Standard multivariate methods, PCA, canonical discriminant analysis
- Problem: too many dimensions and assumptions

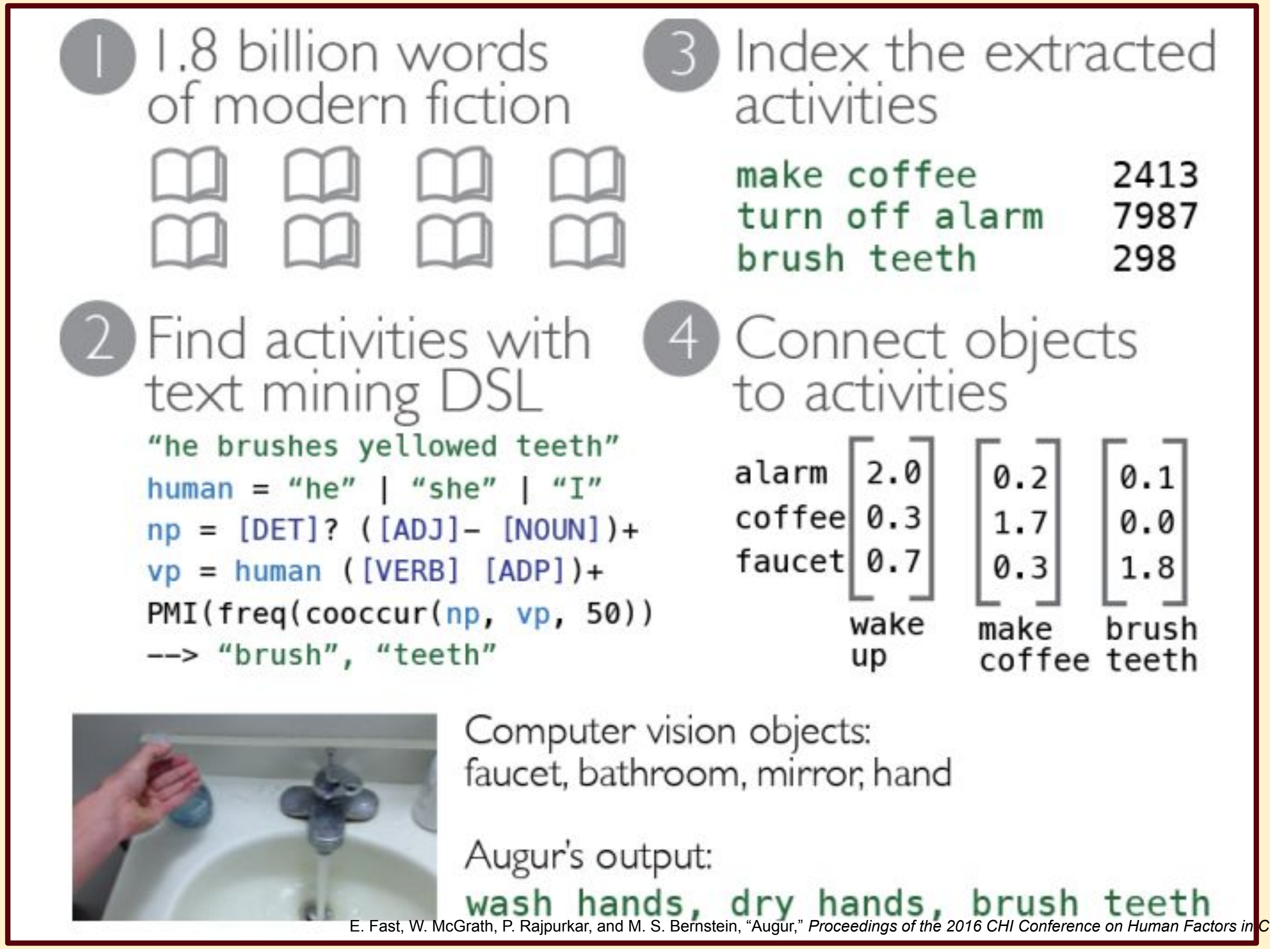


V. Kantorovich, “Quantitative methods and the analysis of literature,” *Soviet Studies in Literature*, vol. 13, no. 3, pp. 86–96, 1977.

Augur

Mining human behaviours from fiction to power interactive systems

- Goal: get computers to understand human behaviour.
- Result: the knowledge base, Augur predicts user activities from surrounding objects.



1 1.8 billion words of modern fiction

2 Find activities with text mining DSL

```
"he brushes yellowed teeth"
human = "he" | "she" | "I"
np = [DET]? ([ADJ]- [NOUN])+
vp = human ([VERB] [ADP])+
PMI(freq(cooccur(np, vp, 50))
--> "brush", "teeth"
```

3 Index the extracted activities

make coffee	2413
turn off alarm	7987
brush teeth	298

4 Connect objects to activities

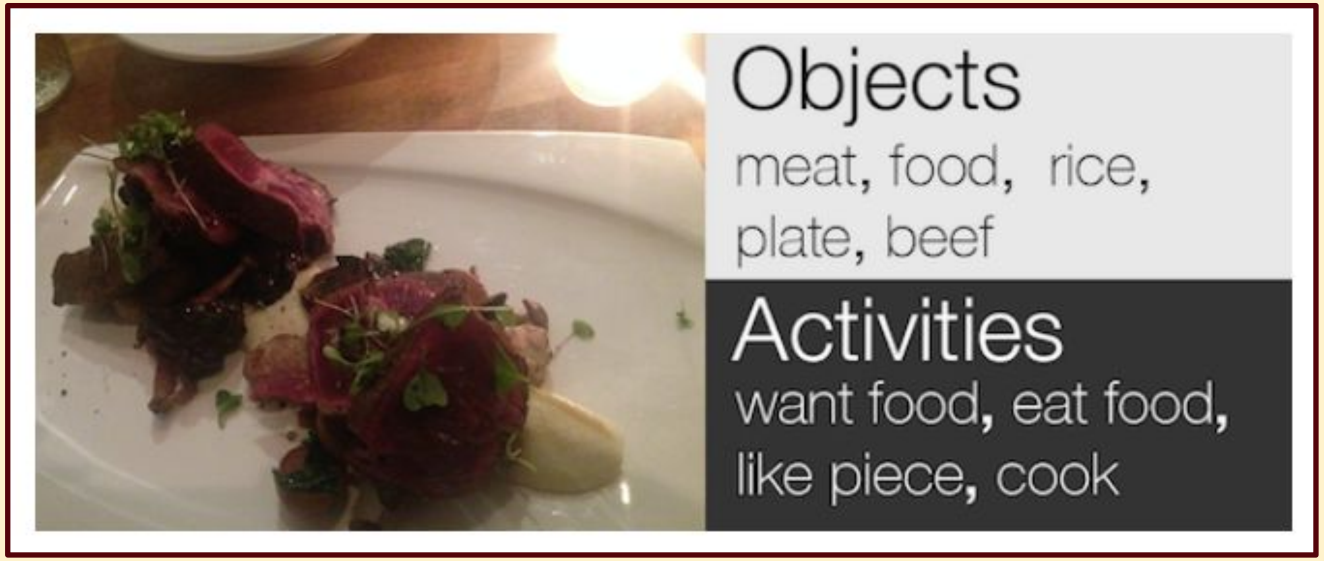
$$\begin{array}{lccc} \text{alarm} & 2.0 & 0.2 & 0.1 \\ \text{coffee} & 0.3 & 1.7 & 0.0 \\ \text{faucet} & 0.7 & 0.3 & 1.8 \\ & \text{wake up} & \text{make coffee} & \text{brush teeth} \end{array}$$

Computer vision objects:
faucet, bathroom, mirror, hand

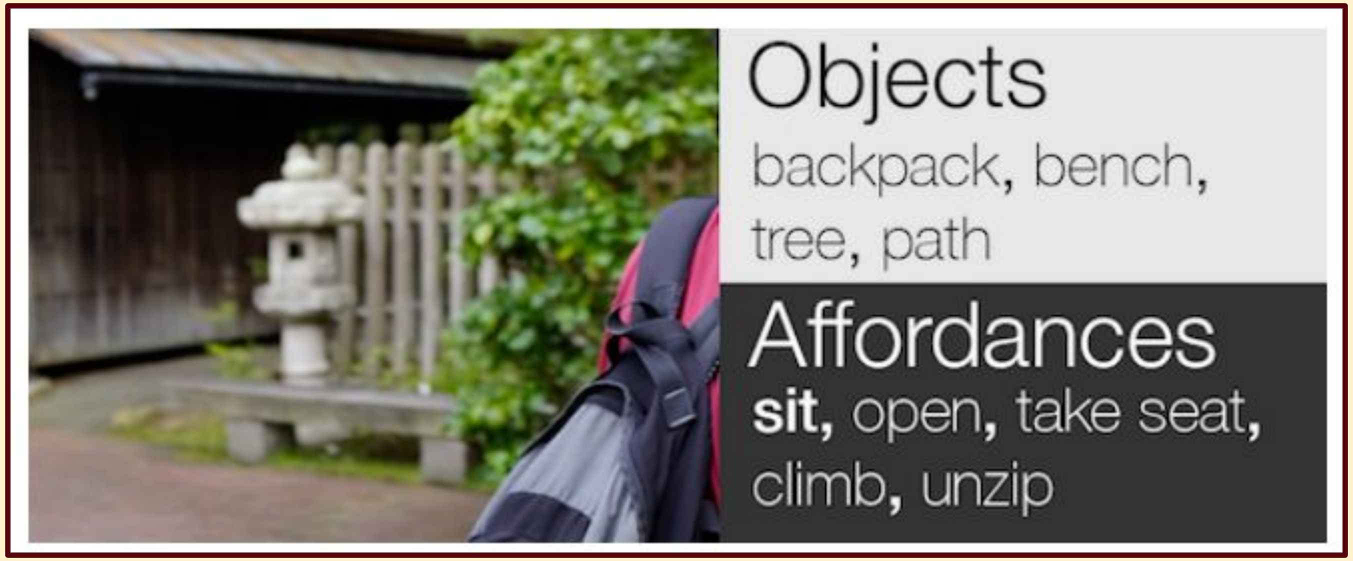
Augur's output:
wash hands, dry hands, brush teeth

E. Fast, W. McGrath, P. Rajpurkar, and M. S. Bernstein, "Augur," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

- Augur continued

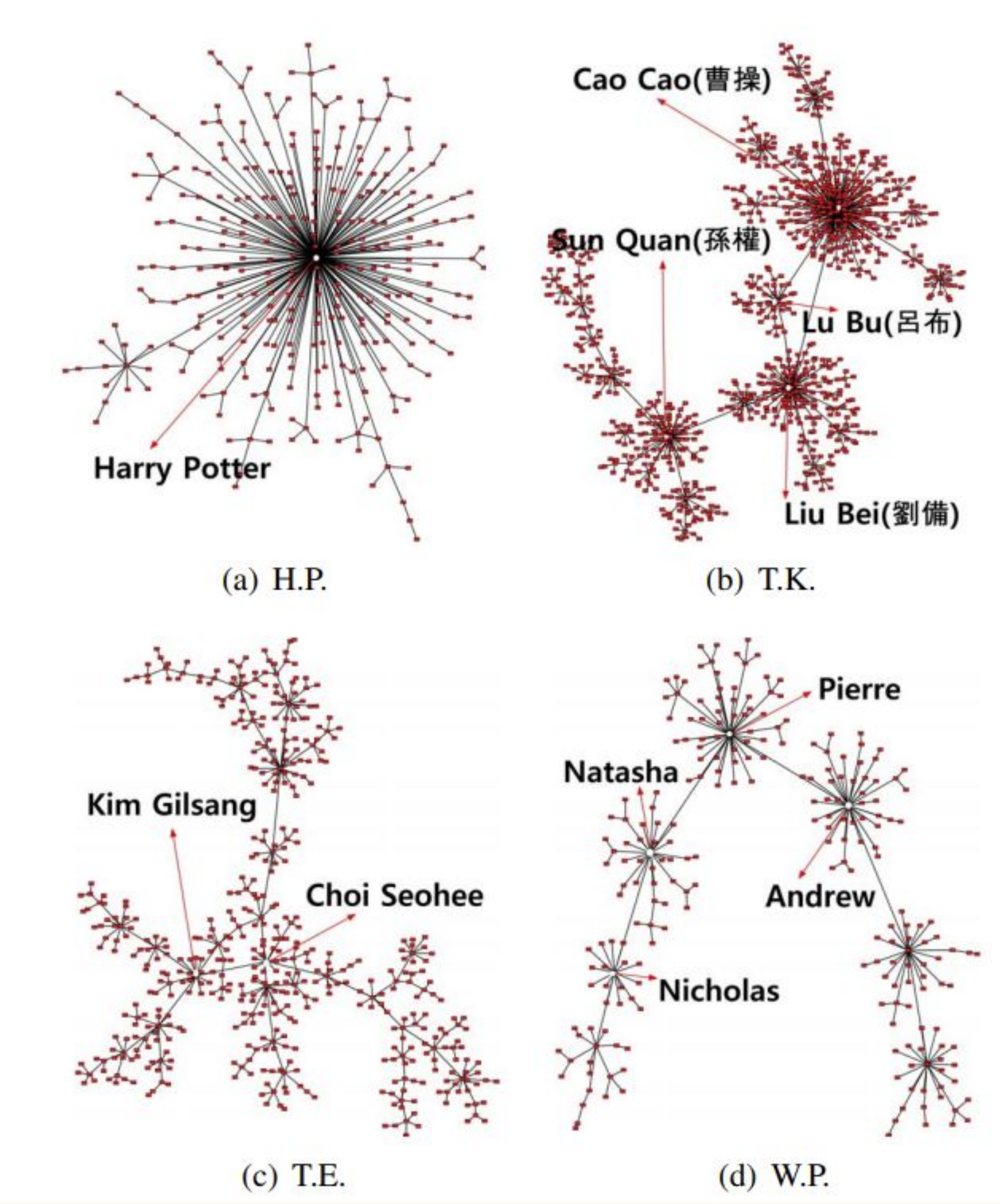


E. Fast, W. McGrath, P. Rajpurkar, and M. S. Bernstein, “Augur,” *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.



Characteristic Analysis of Social Network Constructed from literary fiction:

- Measures interactions within characters
- Fiction is comparable to real life (node degree & path distance)
- Quantitative method (to word co-occurrence) is much more efficient
- Goal: study main character’s social interactions
- Problems: Cannot incorporate aliases and pronouns, too many edges from vectors

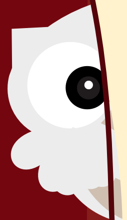


(a) H.P.

(b) T.K.

(c) T.E.

(d) W.P.



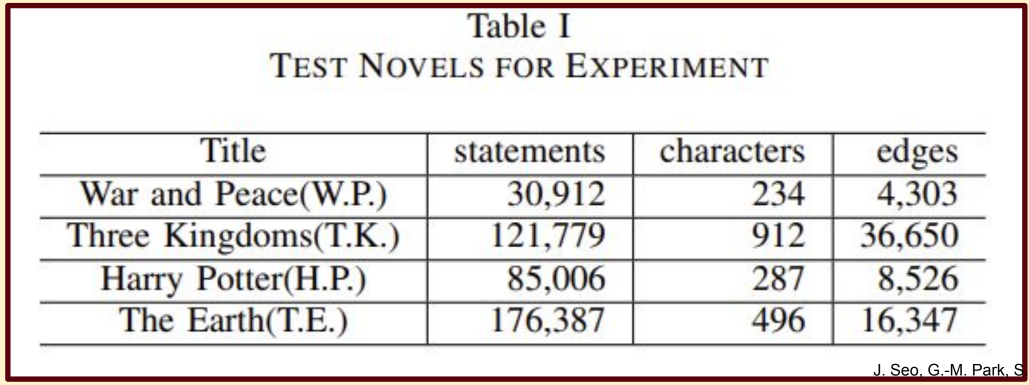
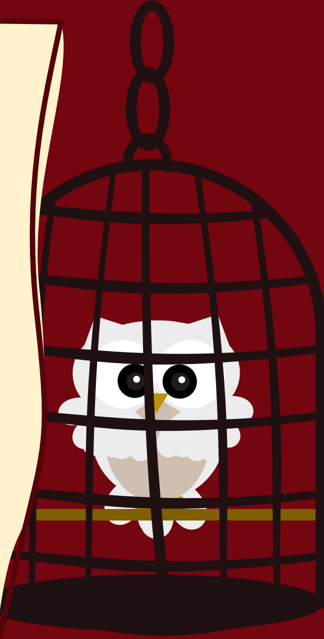


Table I
TEST NOVELS FOR EXPERIMENT

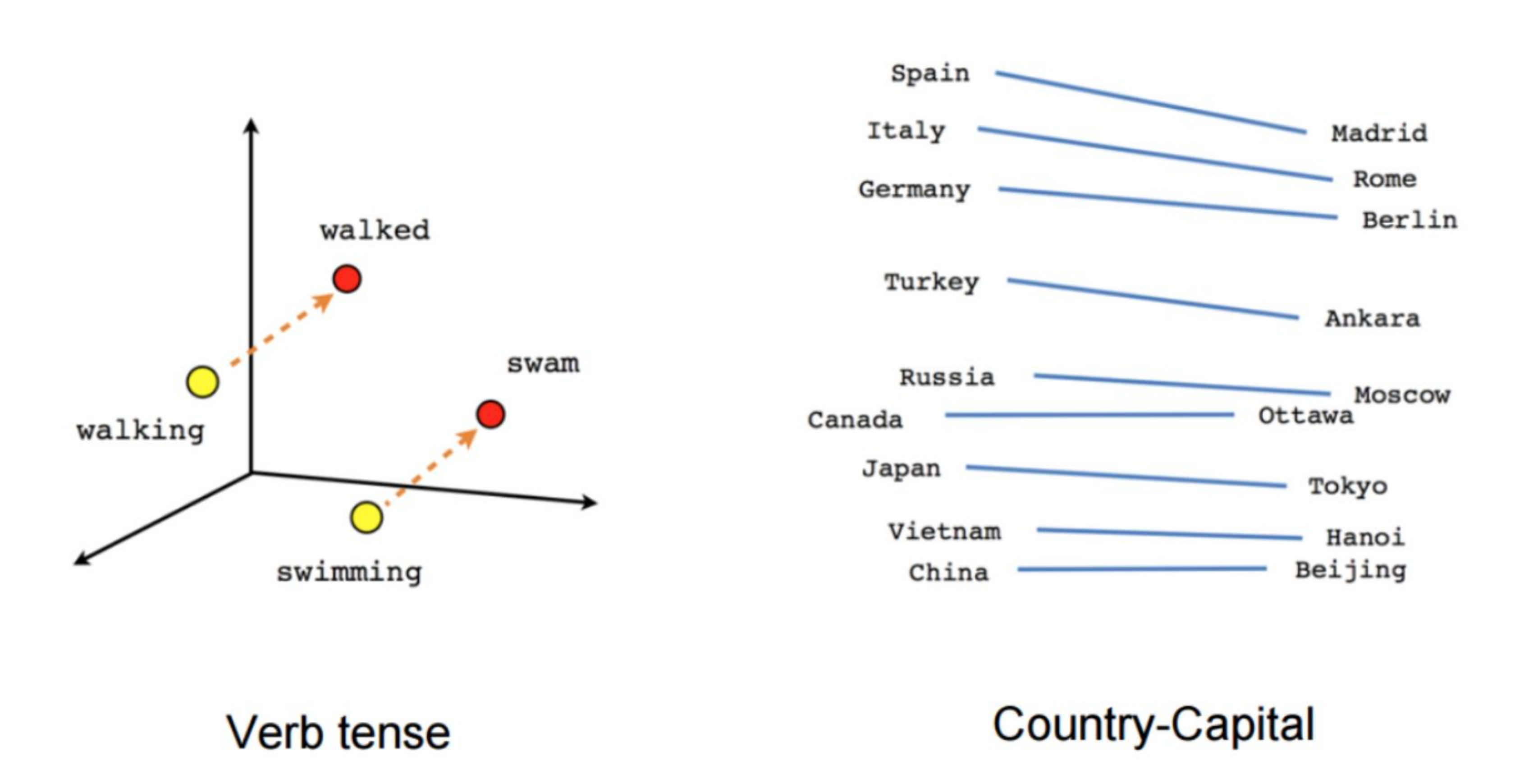
Title	statements	characters	edges
War and Peace(W.P.)	30,912	234	4,303
Three Kingdoms(T.K.)	121,779	912	36,650
Harry Potter(H.P.)	85,006	287	8,526
The Earth(T.E.)	176,387	496	16,347

J. Seo, G.-M. Park, S-H. Kim, and H.-G. Cho, “Characteristic analysis of social network constructed from literary fiction,” *2013 International Conference on Cyberworlds*, 2013.



Bag-of-Words Similarity

Word2vec model sample

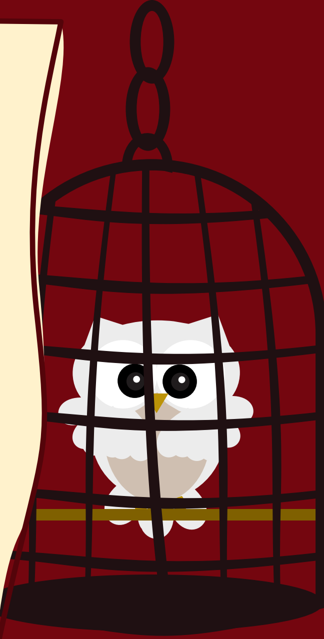


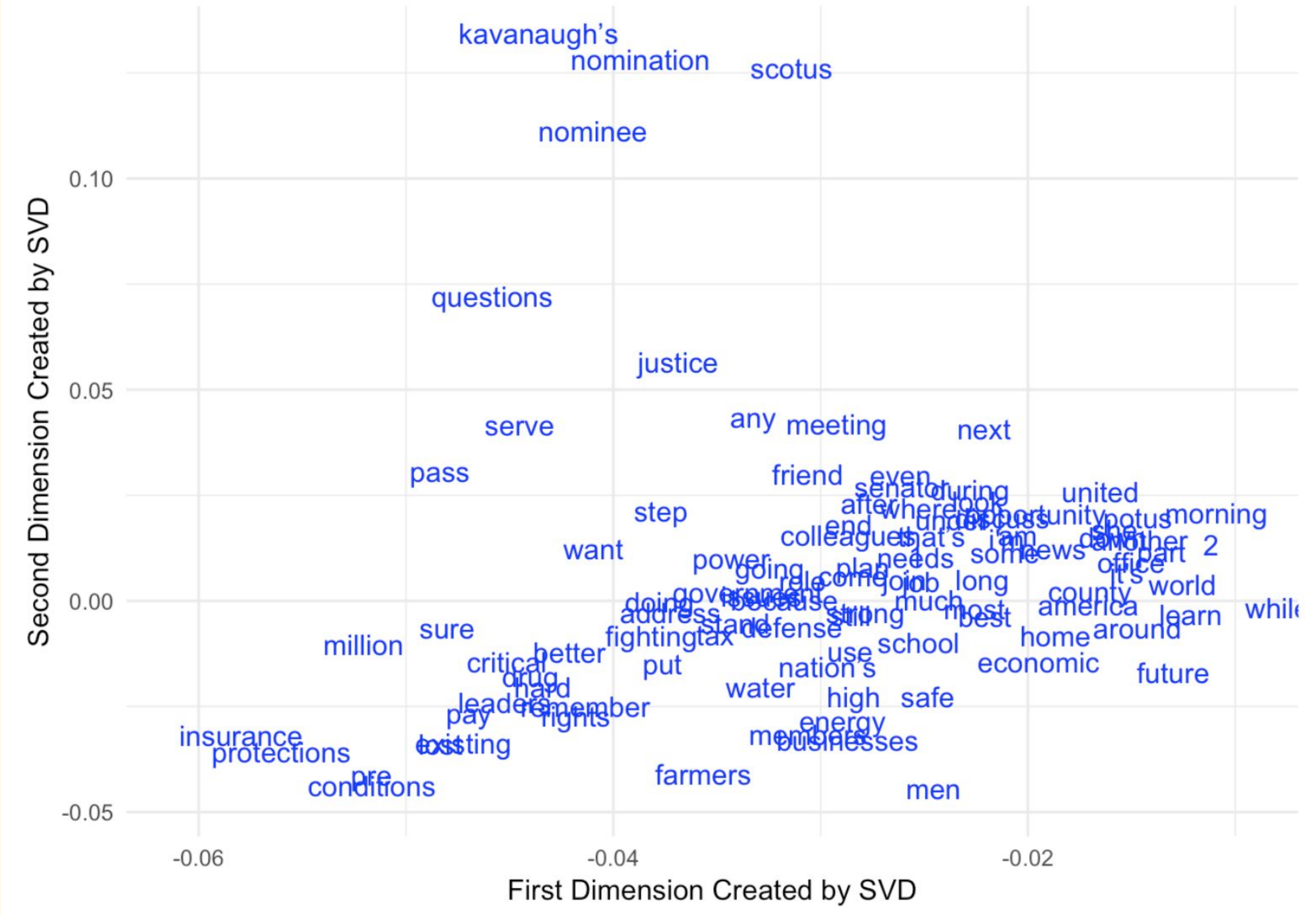
Verb tense

Country-Capital

C. Bail, "Word Embeddings," *Word embeddings*. [Online]. Available: https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html. [Accessed: 06-Dec-2021].

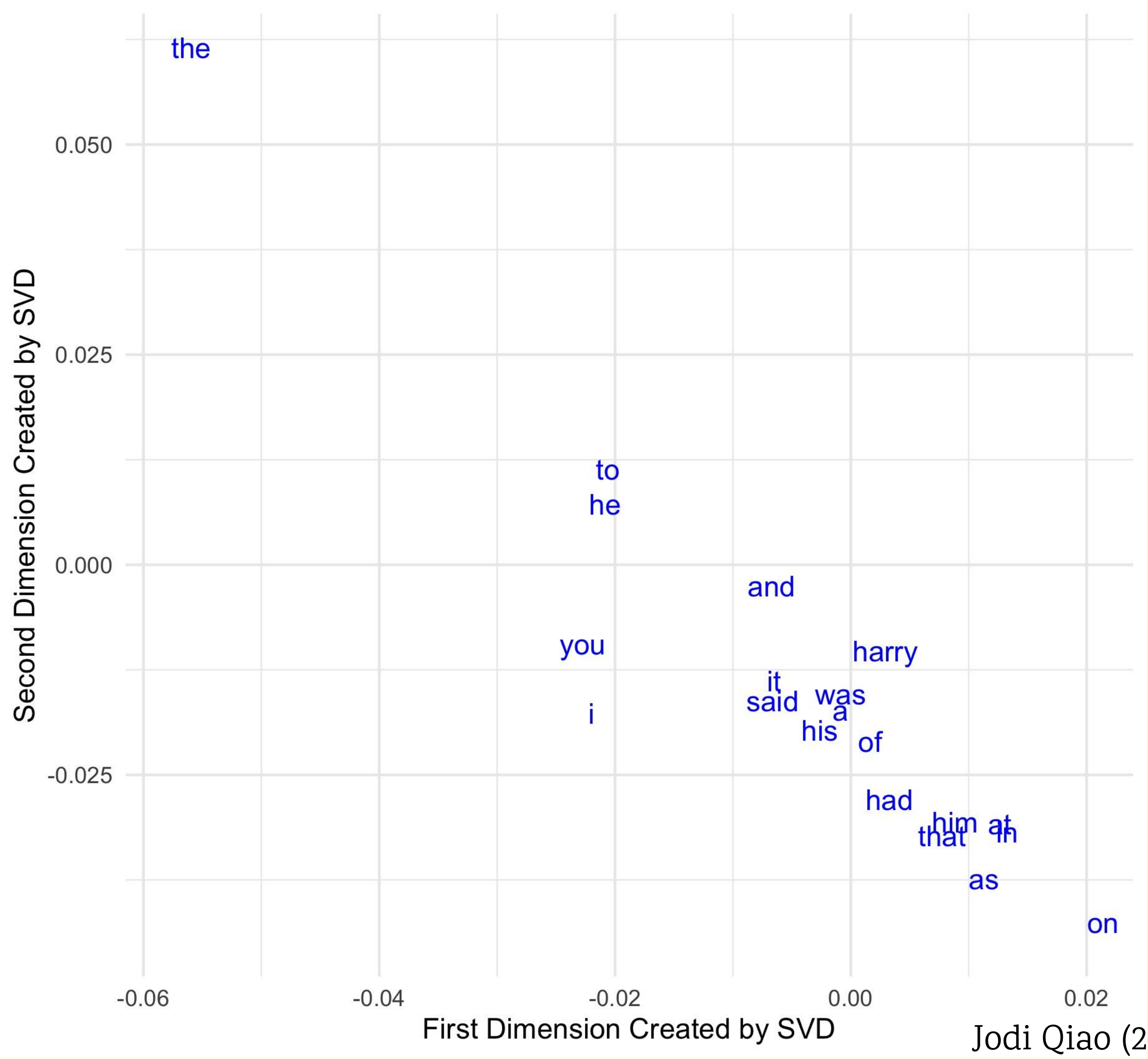
Bag-of-Words Similarity





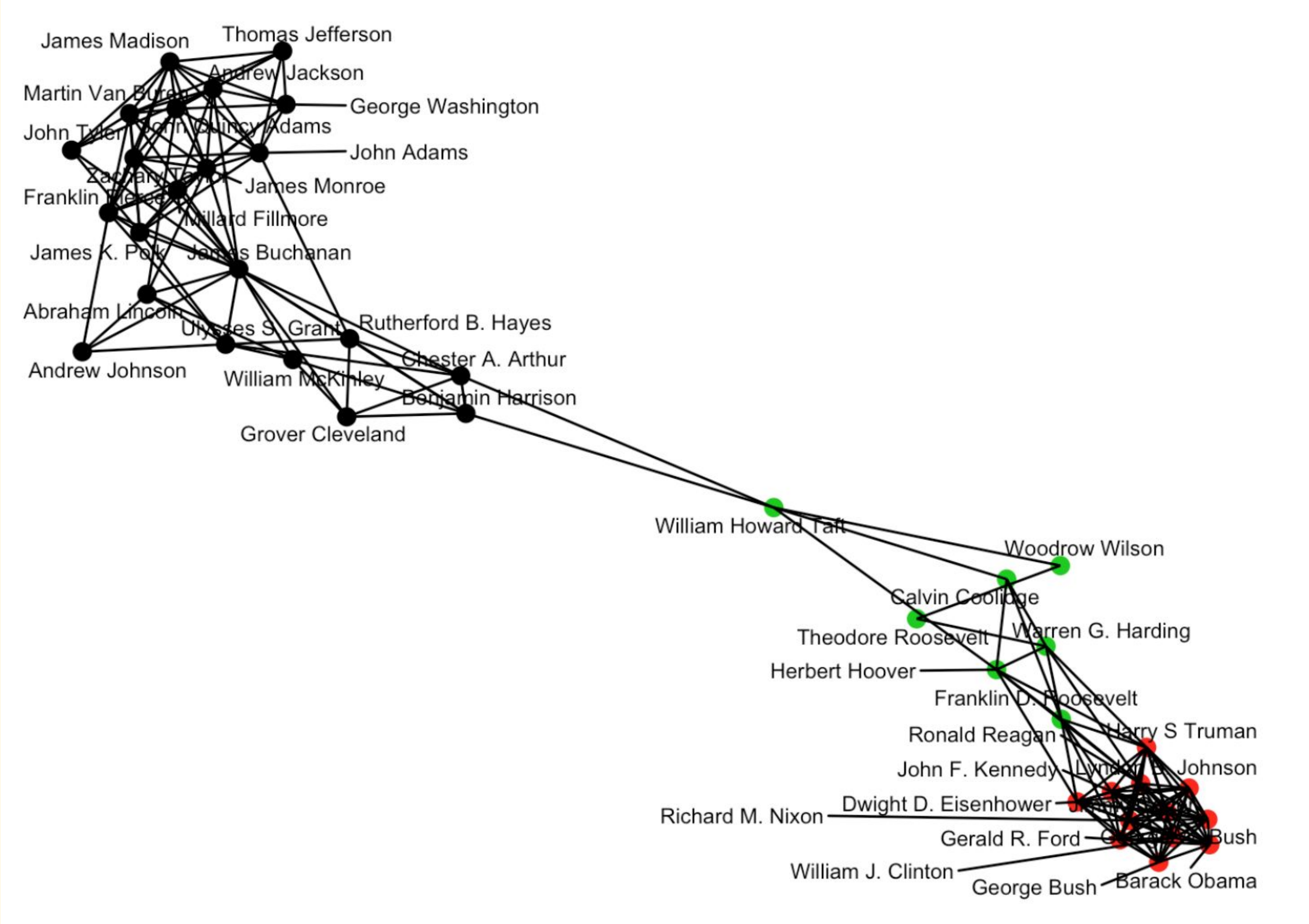
C. Bail, "Word Embeddings," *Word embeddings*. [Online]. Available: https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html. [Accessed: 06-Dec-2021].

Bag-of-Words Similarity



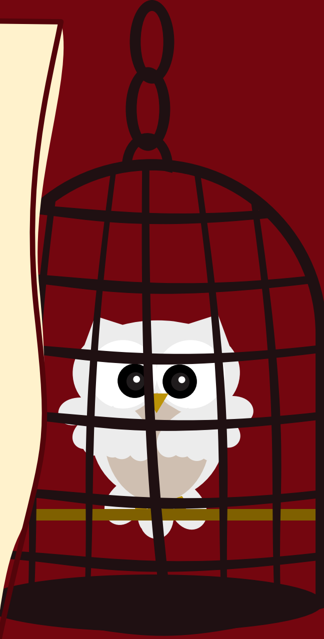
Jodi Qiao (2021)

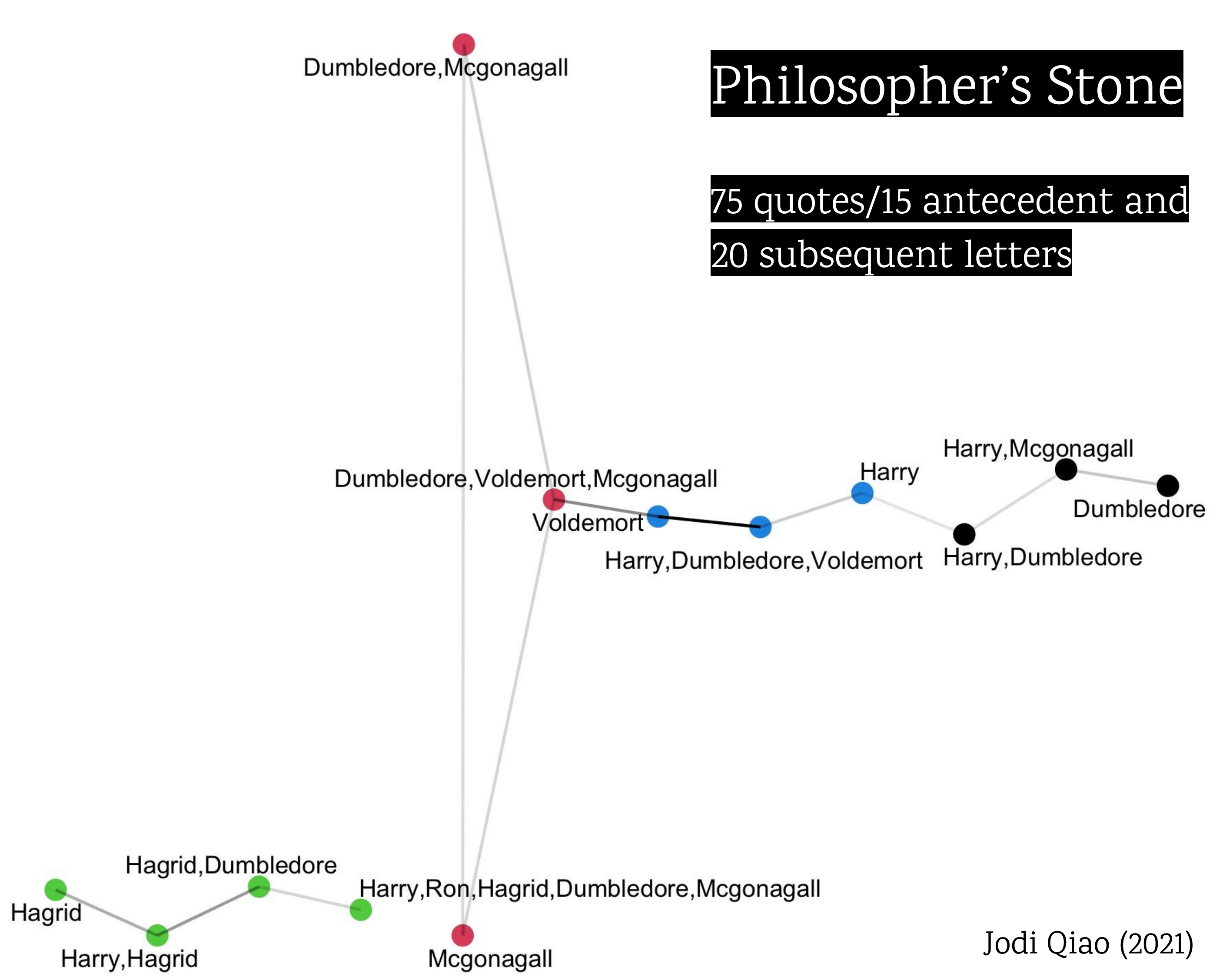
Visualizing Text Networks



C. Bail, "Text networks," *Text Networks*. [Online]. Available: https://sicss.io/2018/materials/day3-text-analysis/text-networks/rmarkdown/SICSS_Text_Networks.html. [Accessed: 06-Dec-2021].

Character Network by Quotes



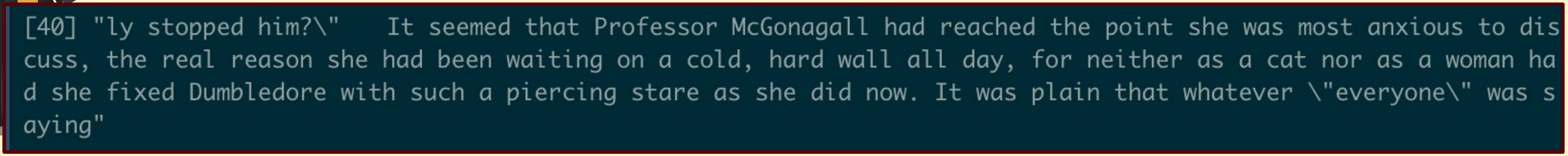


Philosopher's Stone

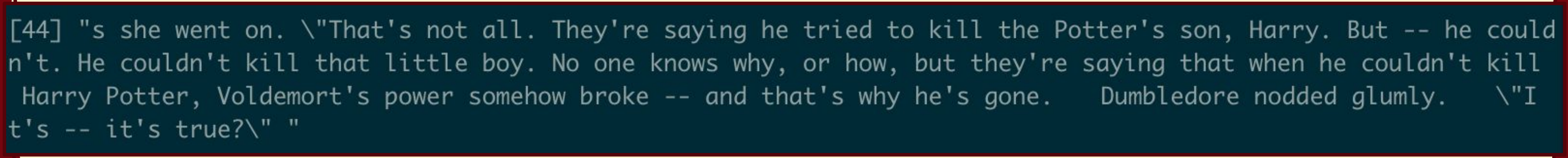
75 quotes/15 antecedent and 20 subsequent letters

Jodi Qiao (2021)

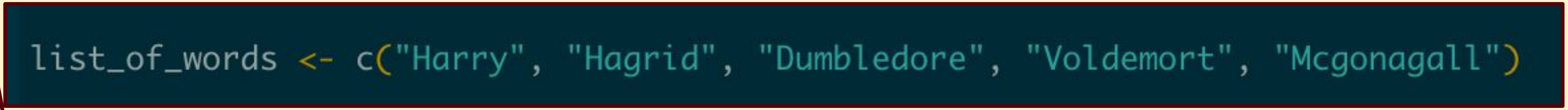
Character Network by Quotes



```
[40] "ly stopped him?\"    It seemed that Professor McGonagall had reached the point she was most anxious to dis
cuss, the real reason she had been waiting on a cold, hard wall all day, for neither as a cat nor as a woman ha
d she fixed Dumbledore with such a piercing stare as she did now. It was plain that whatever \"everyone\" was s
aying"
```

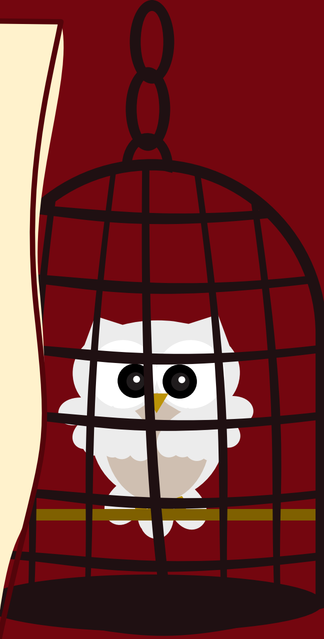


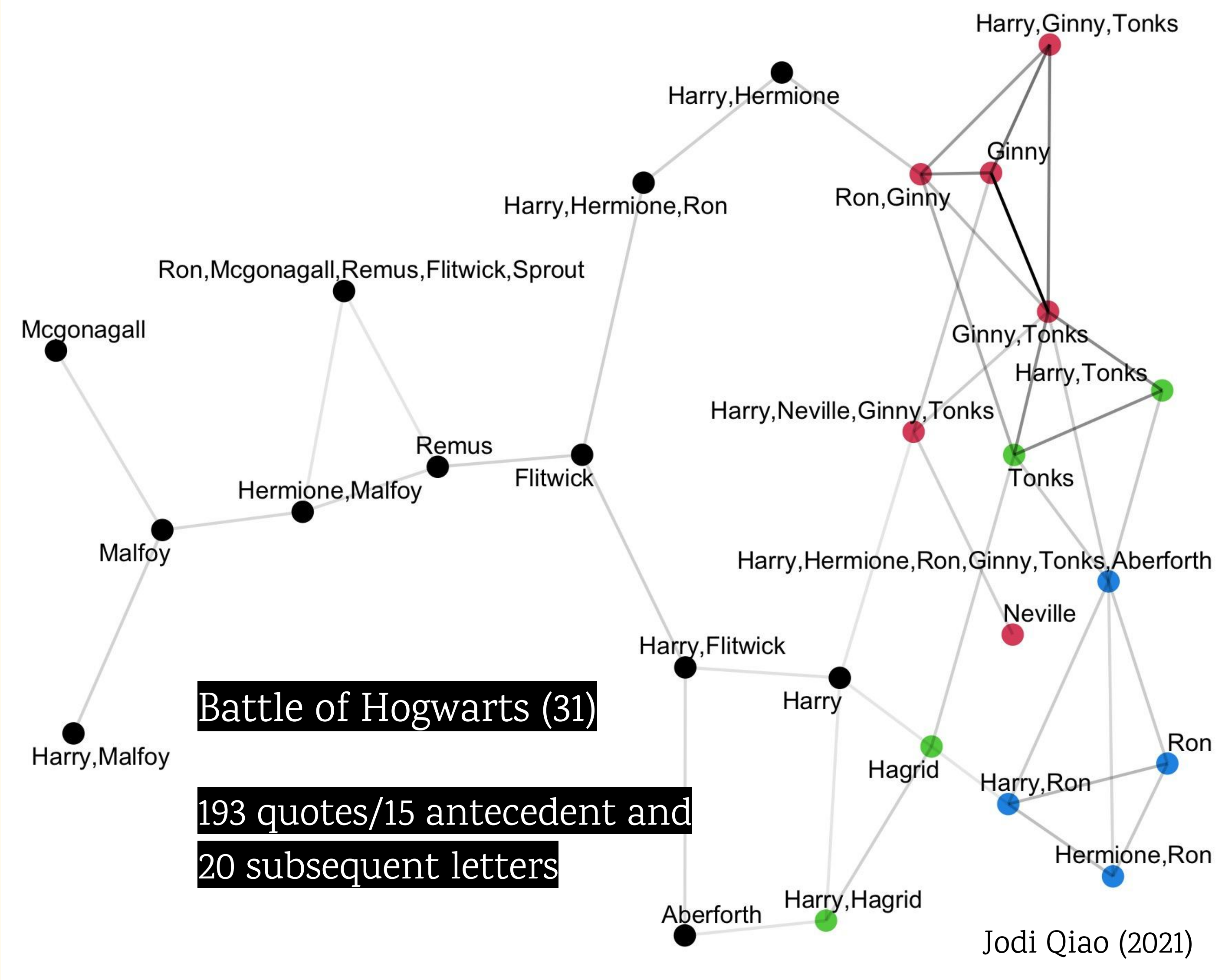
```
[44] "s she went on. \"That's not all. They're saying he tried to kill the Potter's son, Harry. But -- he could
n't. He couldn't kill that little boy. No one knows why, or how, but they're saying that when he couldn't kill
 Harry Potter, Voldemort's power somehow broke -- and that's why he's gone.    Dumbledore nodded glumly.    \"I
t's -- it's true?\" "
```



```
list_of_words <- c("Harry", "Hagrid", "Dumbledore", "Voldemort", "Mcgonagall")
```

Character Network by Quotes





Battle of Hogwarts (31)

193 quotes/15 antecedent and
20 subsequent letters

Jodi Qiao (2021)

Thank you!

References

C. Bail, “Text networks,” *Text Networks*. [Online]. Available: https://sicss.io/2018/materials/day3-text-analysis/text-networks/rmarkdown/SICSS_Text_Networks.html. [Accessed: 06-Dec-2021].

C. Bail, “Word Embeddings,” *Word embeddings*. [Online]. Available: https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html. [Accessed: 06-Dec-2021].

E. Fast, W. McGrath, P. Rajpurkar, and M. S. Bernstein, “Augur,” *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

J. Seo, G.-M. Park, S.-H. Kim, and H.-G. Cho, “Characteristic analysis of social network constructed from literary fiction,” *2013 International Conference on Cyberworlds*, 2013.

V. Kantorovich, “Quantitative methods and the analysis of literature,” *Soviet Studies in Literature*, vol. 13, no. 3, pp. 86–96, 1977.