Centre for
Data Analytics

Insight

Novel2Vec: Characterising 19th Century Literature via Word Embeddings

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene

20 September 2016

DCU NUI Galway OÉ Gaillimh UCC UCD sfi

Nation & Genre Gender

A unique collaboration between the Insight Centre and Humanities Institute in UCD.

Our Objective

Study a collection of Irish and British novels from the 19th Century using both quantitative and qualitative methodologies to examine works in new ways.

Talk Overview

This presentation will focus on our most recent work: The exploration of our textual corpus through the lens of word embeddings.

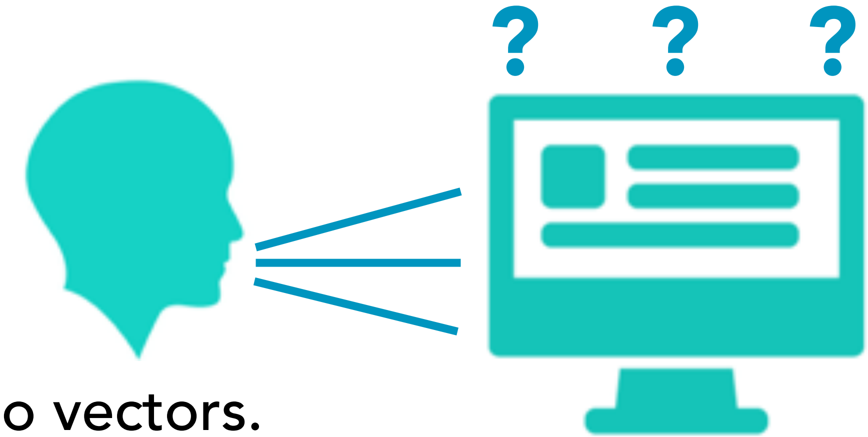
1. What are word embeddings?

2. How we've applied them to our corpus.

3. What insights have we gained as a result?

What are Word Embeddings?

Computers aren't adept at understanding natural language like humans.



Therefore, we convert words into vectors.

These vector representations of words are called

Word Embeddings

vec("king") - vec("man") + vec("woman") = vec("queen")

They can capture the underlying similarity between words and their semantic properties.

19th Century Word Embeddings

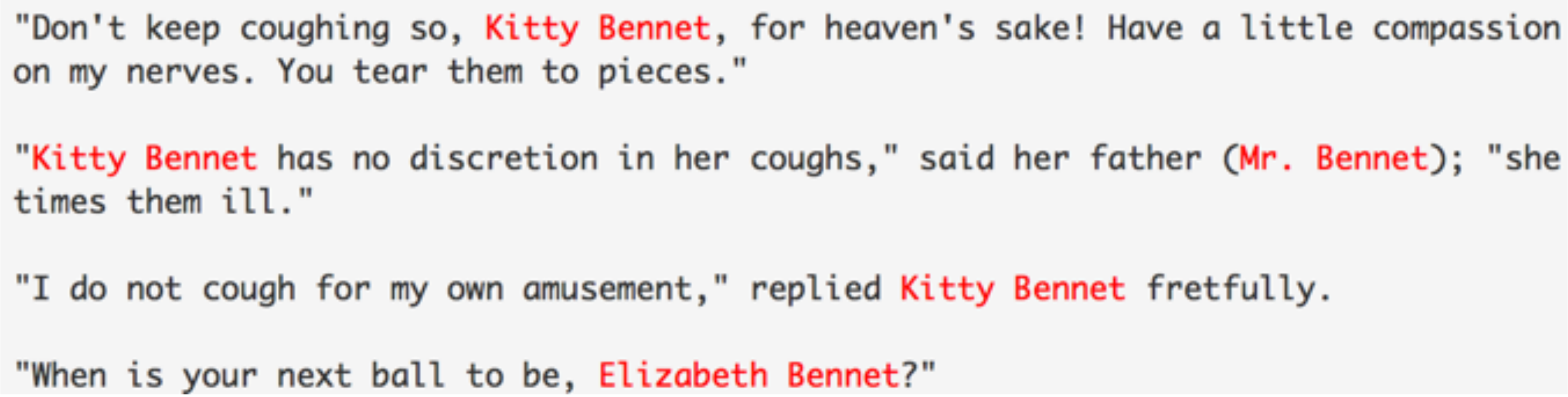
Novels Selected

- 6 novels by Jane Austen
- 3 by Charles Dickens - Bleak House, Oliver Twist, Great Expectations
- 3 by Arthur Conan Doyle - First 3 novels in the Sherlock Holmes series.

19th Century Word Embeddings

Methods: Data Preprocessing

- The text of each chapter of a novel is annotated by a literary scholar to identify all characters and their aliases.



"Don't keep coughing so, Kitty Bennet, for heaven's sake! Have a little compassion on my nerves. You tear them to pieces."

"Kitty Bennet has no discretion in her coughs," said her father (Mr. Bennet); "she times them ill."

"I do not cough for my own amusement," replied Kitty Bennet fretfully.

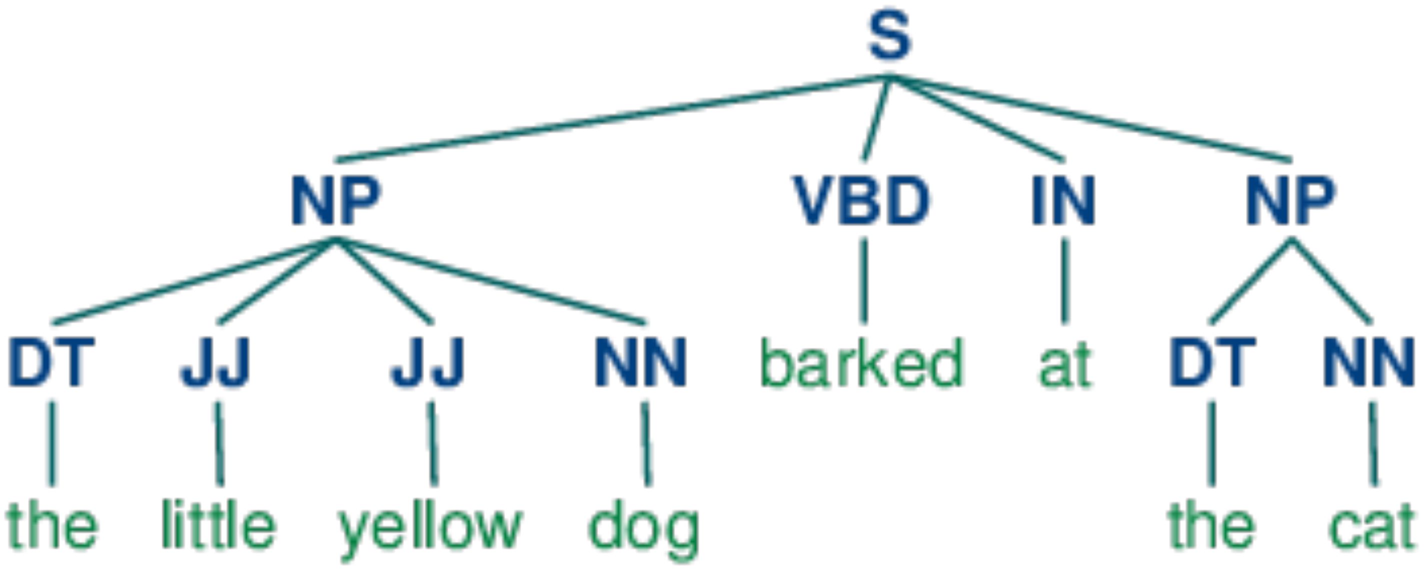
"When is your next ball to be, Elizabeth Bennet?"

- A character dictionary is then created, mapping all aliases for a character to their definitive name.

19th Century Word Embeddings

Methods: Data Preprocessing

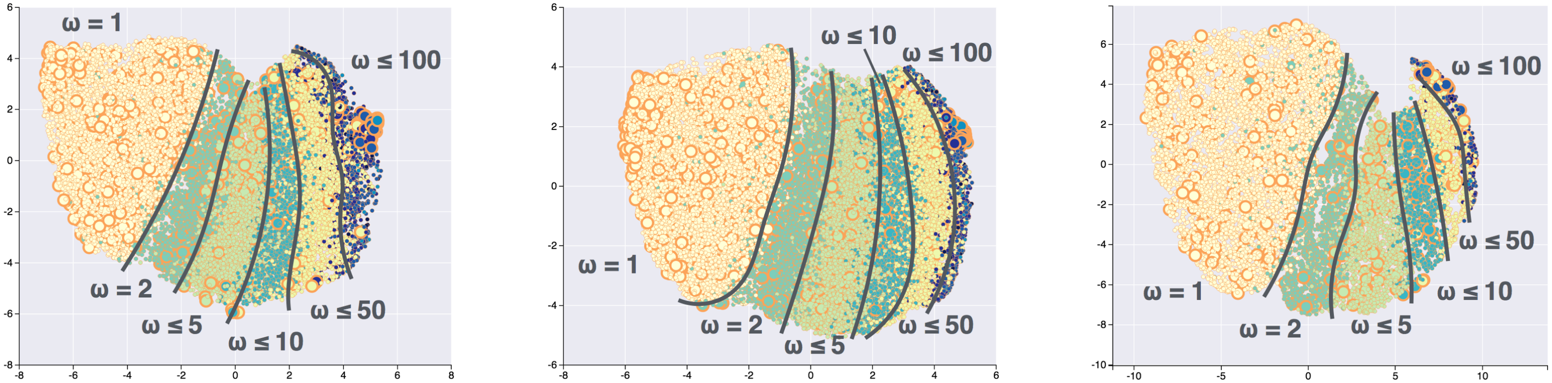
- Part-of-speech tagging (POS tagging) was applied to each text using the Natural Language Toolkit (NLTK) PerceptronTagger Implementation.



19th Century Word Embeddings

Methods: Word Embedding Generation

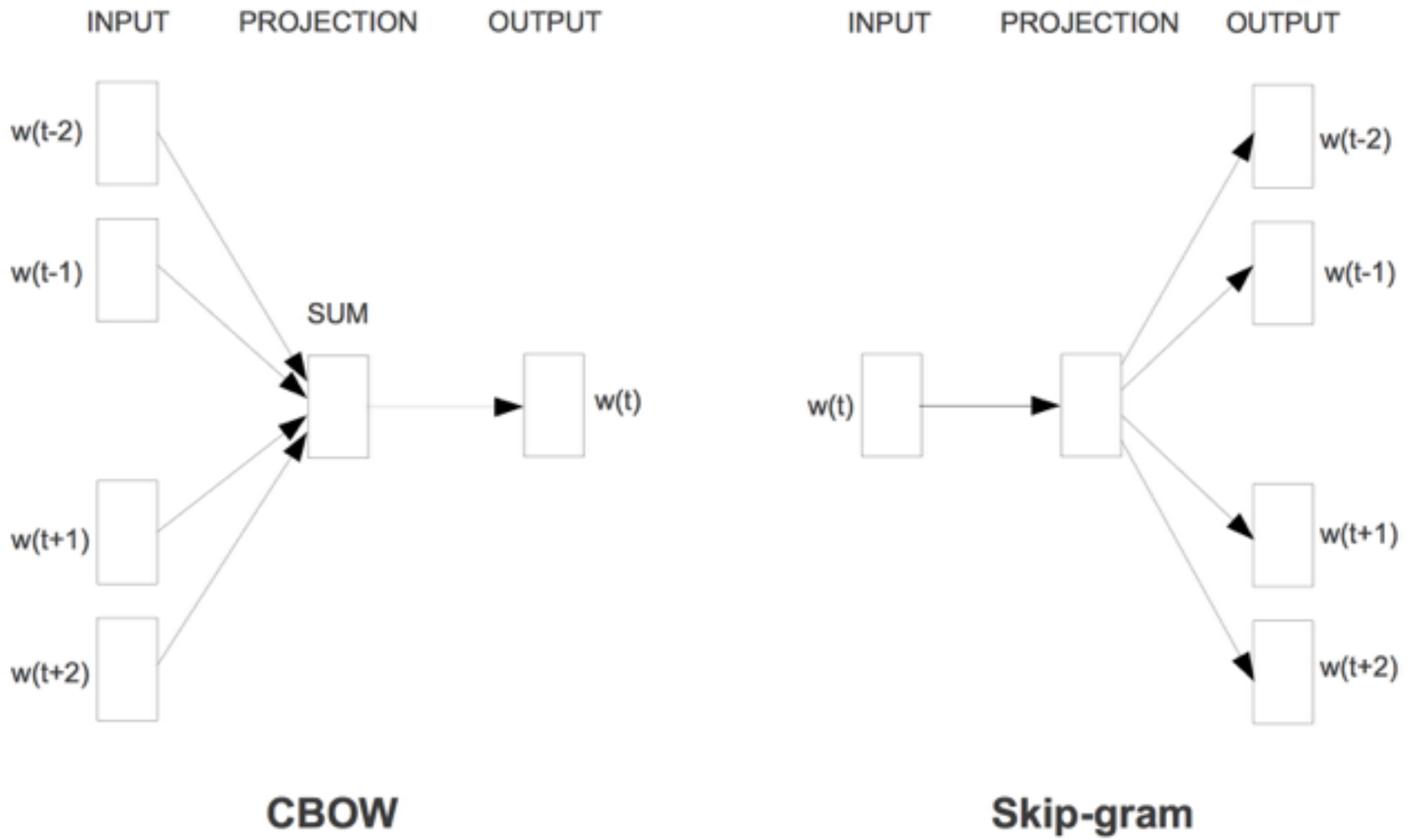
- Word2vec is a two-layer neural network that processes text into a set of feature vectors distributed within a dense dimensional space.
- For our purposes, we used the Gensim implementation.



- We visualised the generated word embeddings in 2 dimensional space using the dimensionality reduction technique known as t-SNE initialised with PCA.

19th Century Word Embeddings

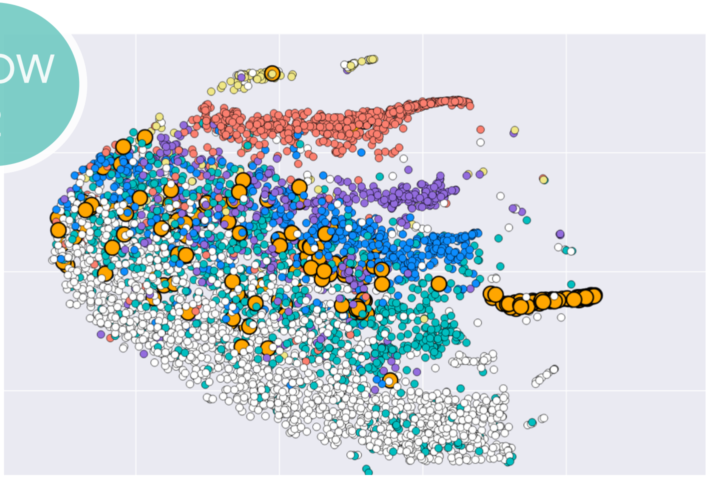
Word2Vec: CBOW Versus Skip-Gram NS



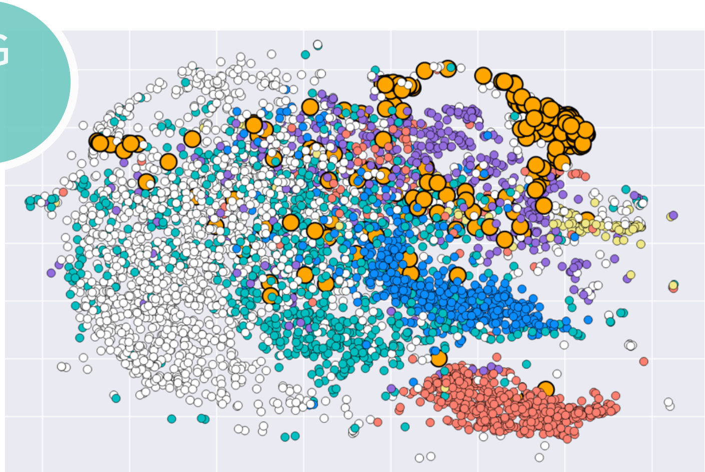
19th Century Word Embeddings

Resulting Word Embeddings - Austen

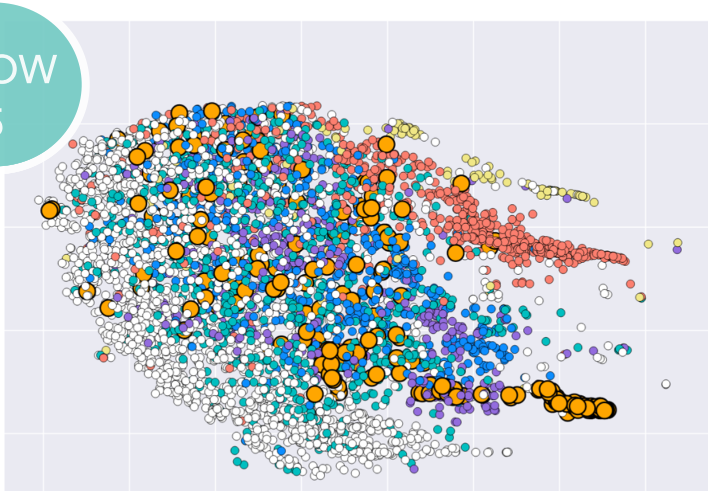
CBOW 2



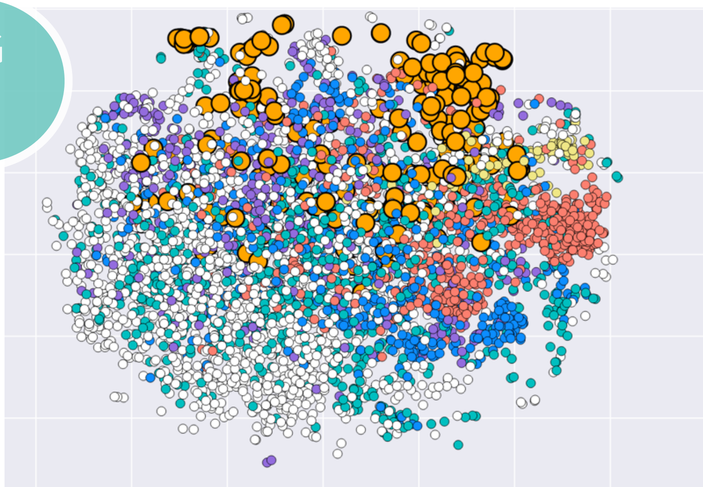
SG 2



CBOW 5

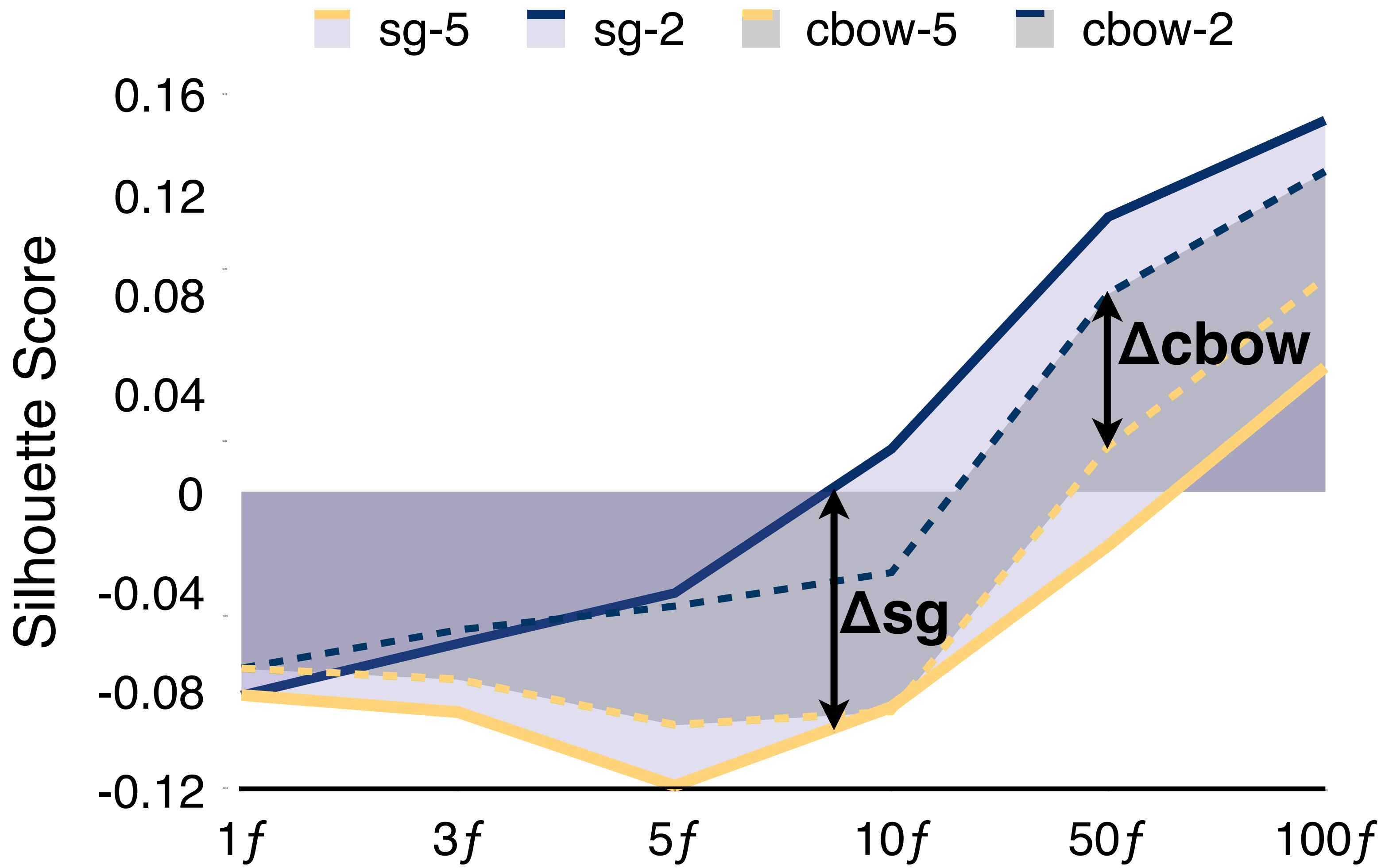


SG 5



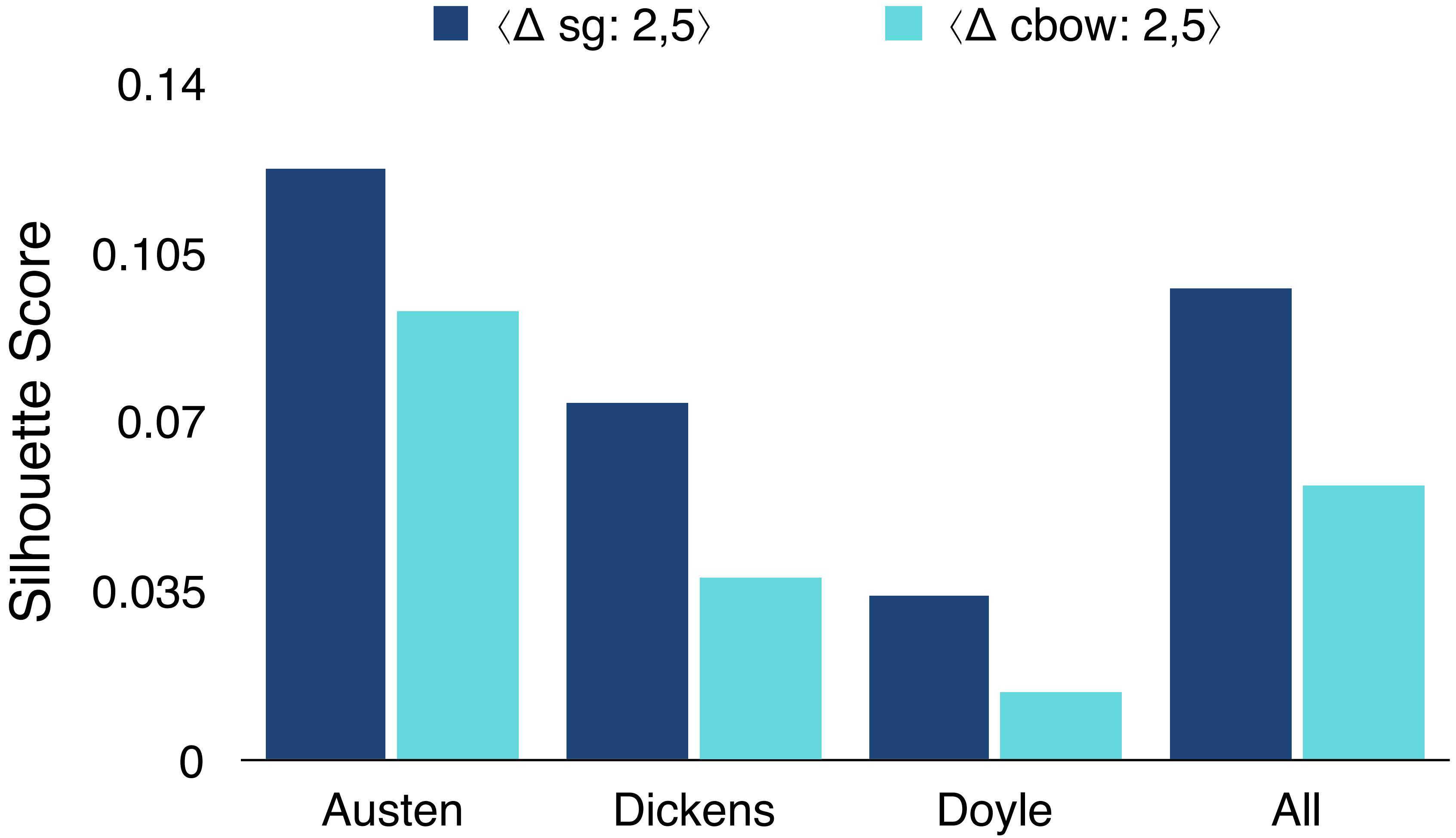
19th Century Word Embeddings

Context Window Sensitivity - Dickens Corpus



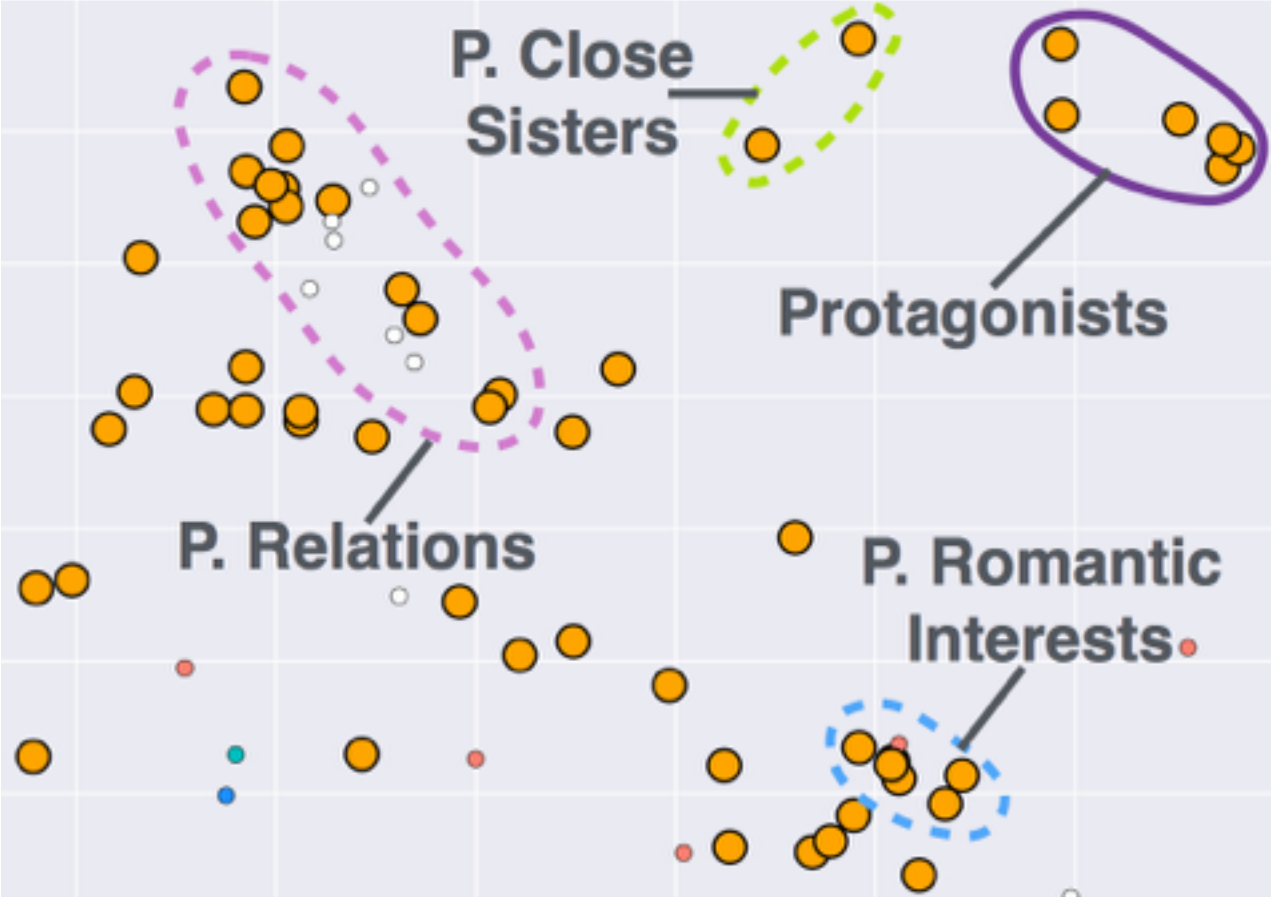
19th Century Word Embeddings

Context Window Sensitivity



Results

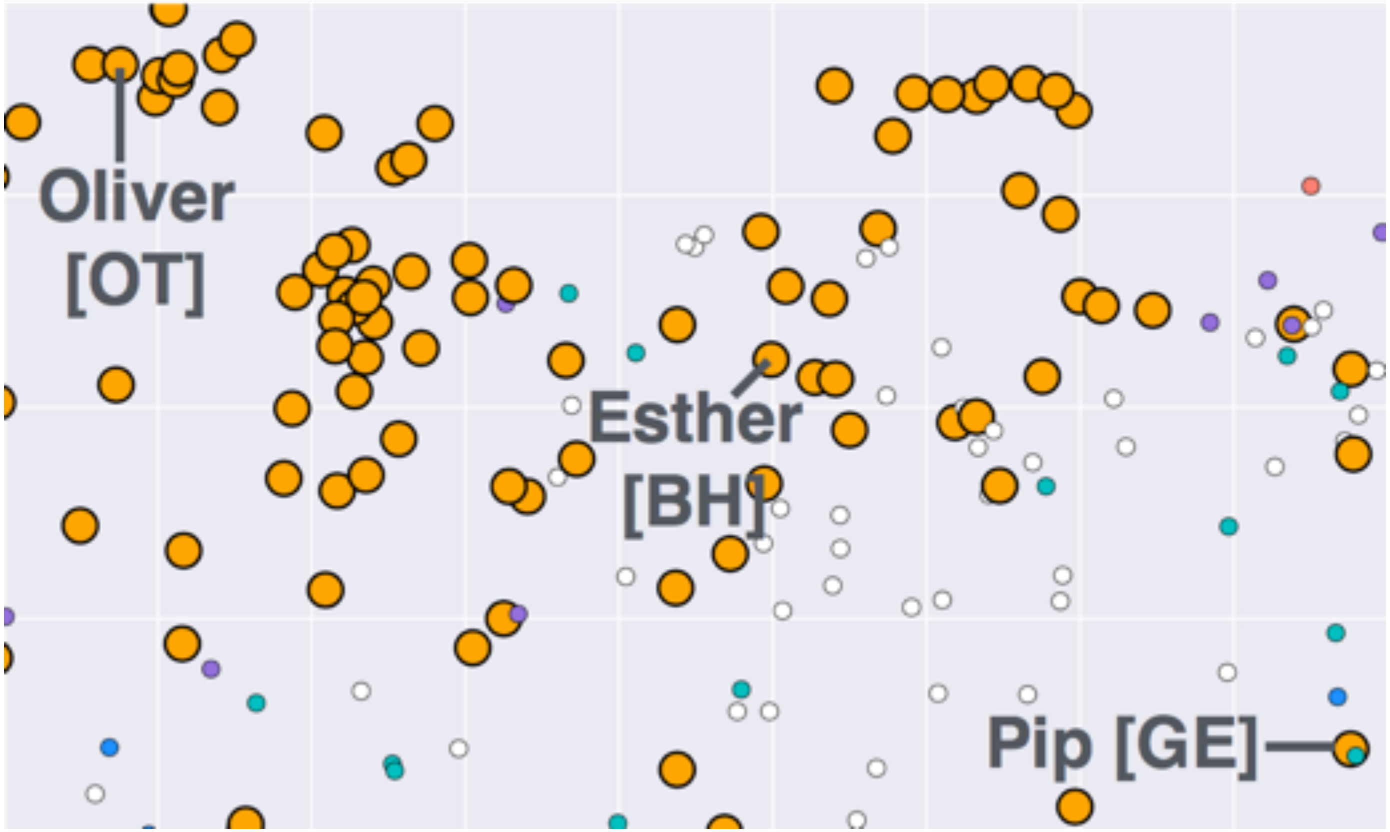
Jane Austen



The 6 protagonists of Austen can be found grouped together.

Results

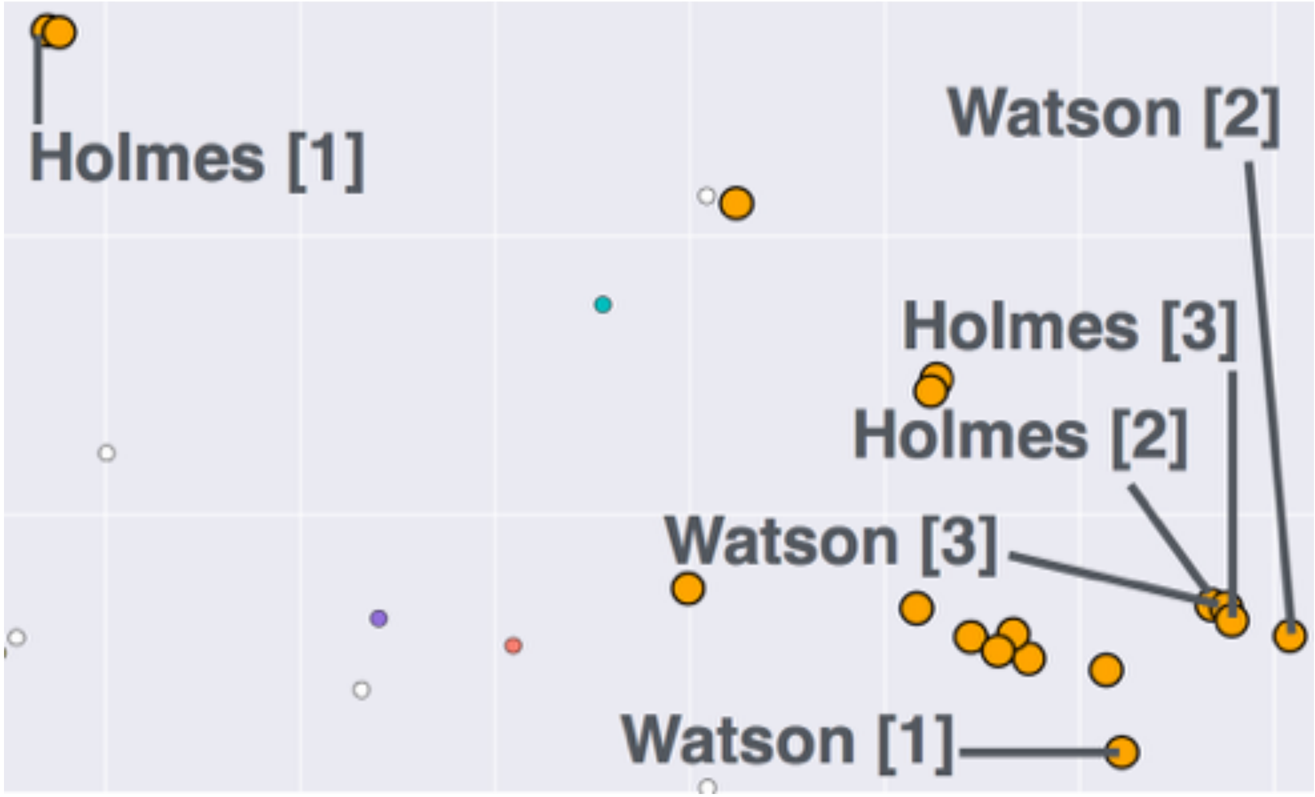
Charles Dickens



Unlike Austen, the protagonists within our Dickens dataset do not group together.

Results

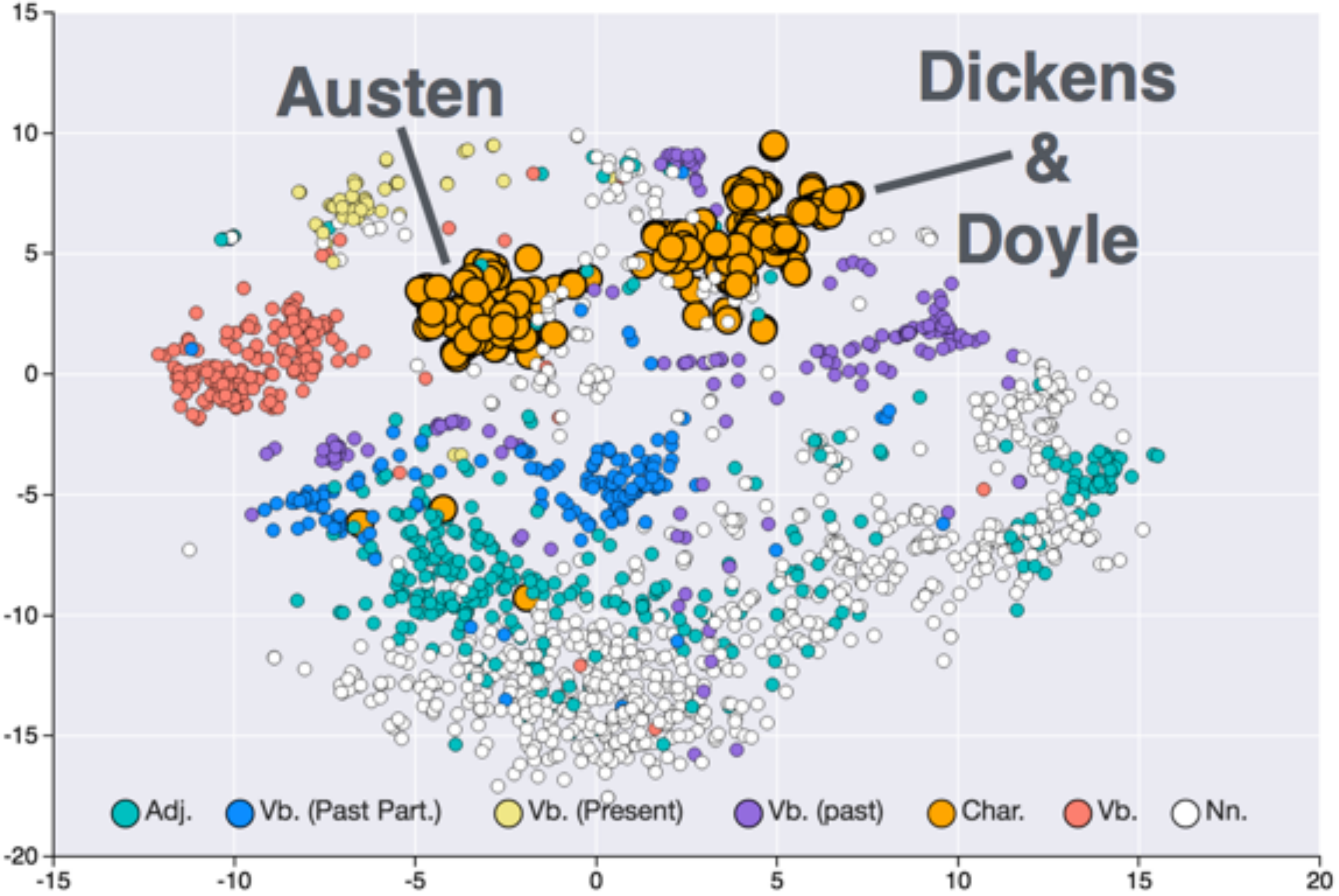
Arthur Conan Doyle



Sherlock Holmes from the first book in the series does not map into the same embedding space as later versions of himself.

Results

Aggregated: Austen, Dickens, Doyle



Conclusions and Future Work

- We have generated, visualised, and explored word embedding representations for four different datasets consisting of 12 popular 19th century novels.

- Our results suggest that word embeddings can potentially act as a useful tool in supporting quantitative literary analysis.

- Providing new ways of representing and visualising well-known literary texts that complement traditional "close reading" techniques.

Future Work

- In future work, we hope to extend our analysis to diachronic word embeddings to discover how word usage within our corpus changes over time.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality." *Advances in Neural Information Processing Systems*, 2013.

[2] L. v. d. Maatan and G. Hinton. "Visualizing data using t-sne." *Journal of Machine Learning Research*, 9, 2008.

[3] R. Rehurek and P. Sojka. "Software framework for topic modelling with large copora." *LREC Workshop on New Challenges for NLP Frameworks*, 2010.

[4] S. Bird, E. Klein, E. Loper. "Natural language processing with Python." *O'Reilly Media, Inc.*, 2009.

[5] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Diachronic word embeddings reveal historical laws of semantic change." *ACL*, 2016.

[6] S. Grayson, K. Wade, G. Meaney, and D. Greene. "The Sense and Sensibility of different sliding windows in constructing co-occurrence networks from literature." *2nd Int. Workshop on Computational History and Data-Driven Humanities*, 2016.