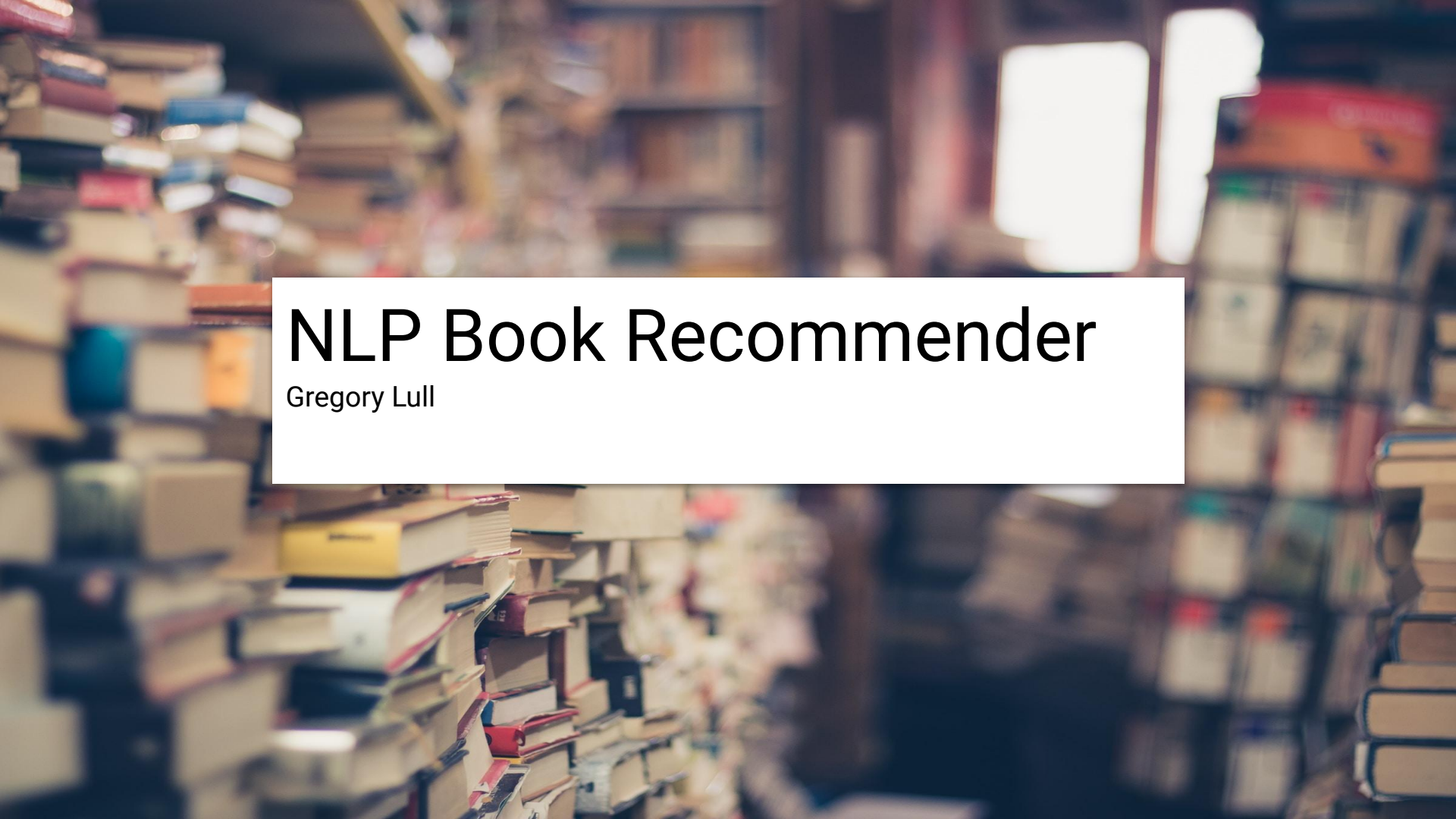


NLP Book Recommender

Gregory Lull

Reading Stats

- Range of books read per year: 4 - 13
- Lifetime: 100 - 1000 books

Data collection

- 500 Kindle books
- 500 out of 50,000 Project Gutenberg books
- Cross referenced with Goodreads top 10k books

Methodology

- Clean text
- → tokenize
- → reduce dimension
- → cluster
- → cosine similarity

Analysis

Similar books will cluster together, e.g. Harry Potter 1 should be in the same cluster as Harry Potter 6...and they are!

Topic 0: pron, say, know, like, look, come, think, time, man, good

Topic 1: mr, holmes, customer, bitcoin, amazon, bezos, mr pickwick, pickwick

Conclusion

- Topics modeling words are related to each other
- Clustering speeds up the cosine similarity check
- This can return a list of “similar books” but needs

Future work

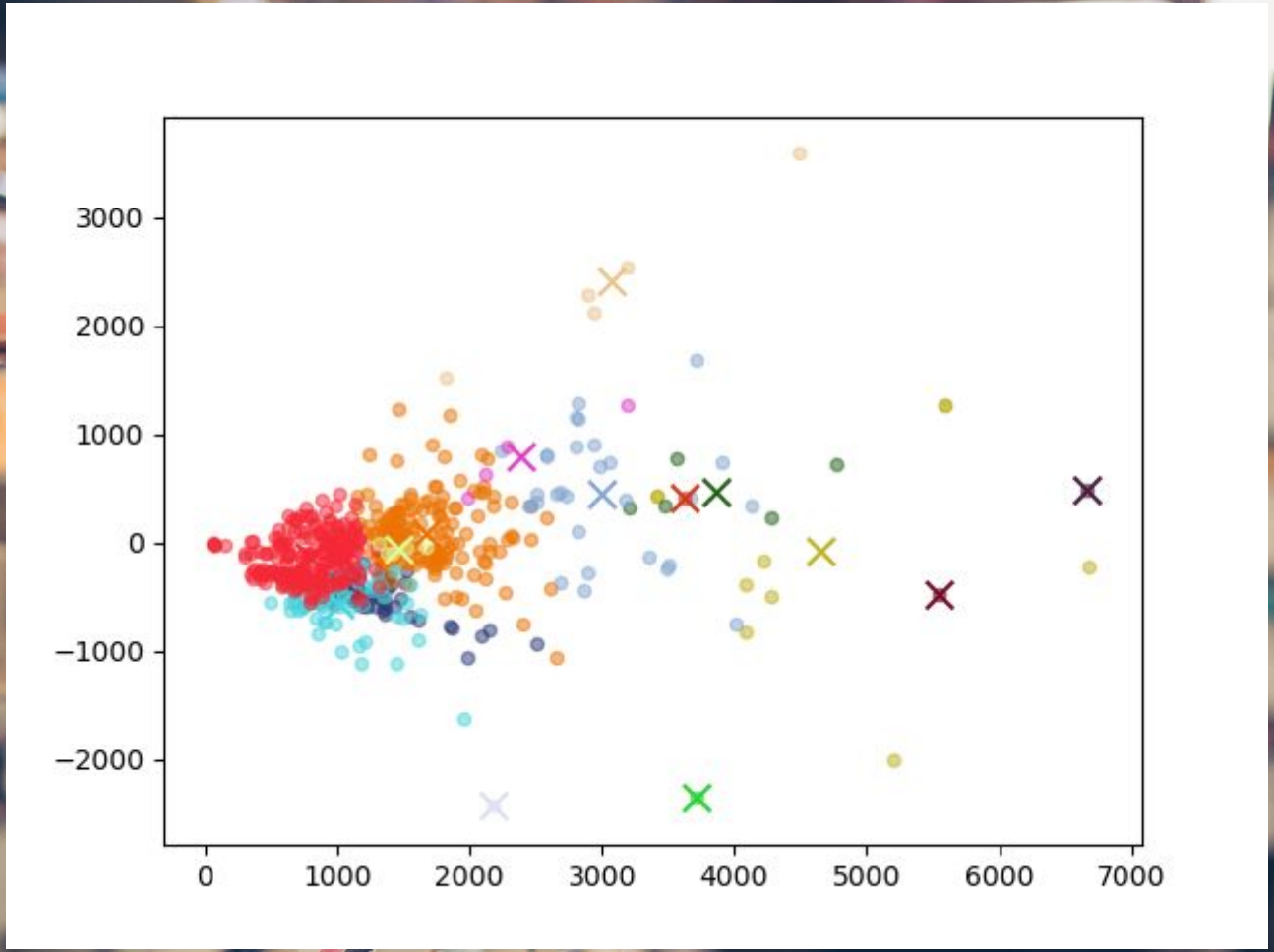
- Use Goodreads user ratings to sort the popularity
- Get Kindle preview samples

Appendix

Topic 0 pron, say, know, like, look, come, think, time, man, good, mr, tell, day, want, little, pron know, way, thing, hand, ask

Topic 1 garfield, man, lincoln, seward, new, time, day, year, war, sleep, state, president, washington, russell, union, bitcoin, hamilton, sumter, come, old

Topic 2 mr, say, oliver, sir, mr pickwick, say mr, mrs, pickwick, gentleman, squeers, man, boffin, pecksniff, holmes, little, mr pecksniff, come, gradgrind, dear, lucy


3000
2000
1000
0
−1000
−2000
0
1000
2000
3000
4000
5000
6000
7000