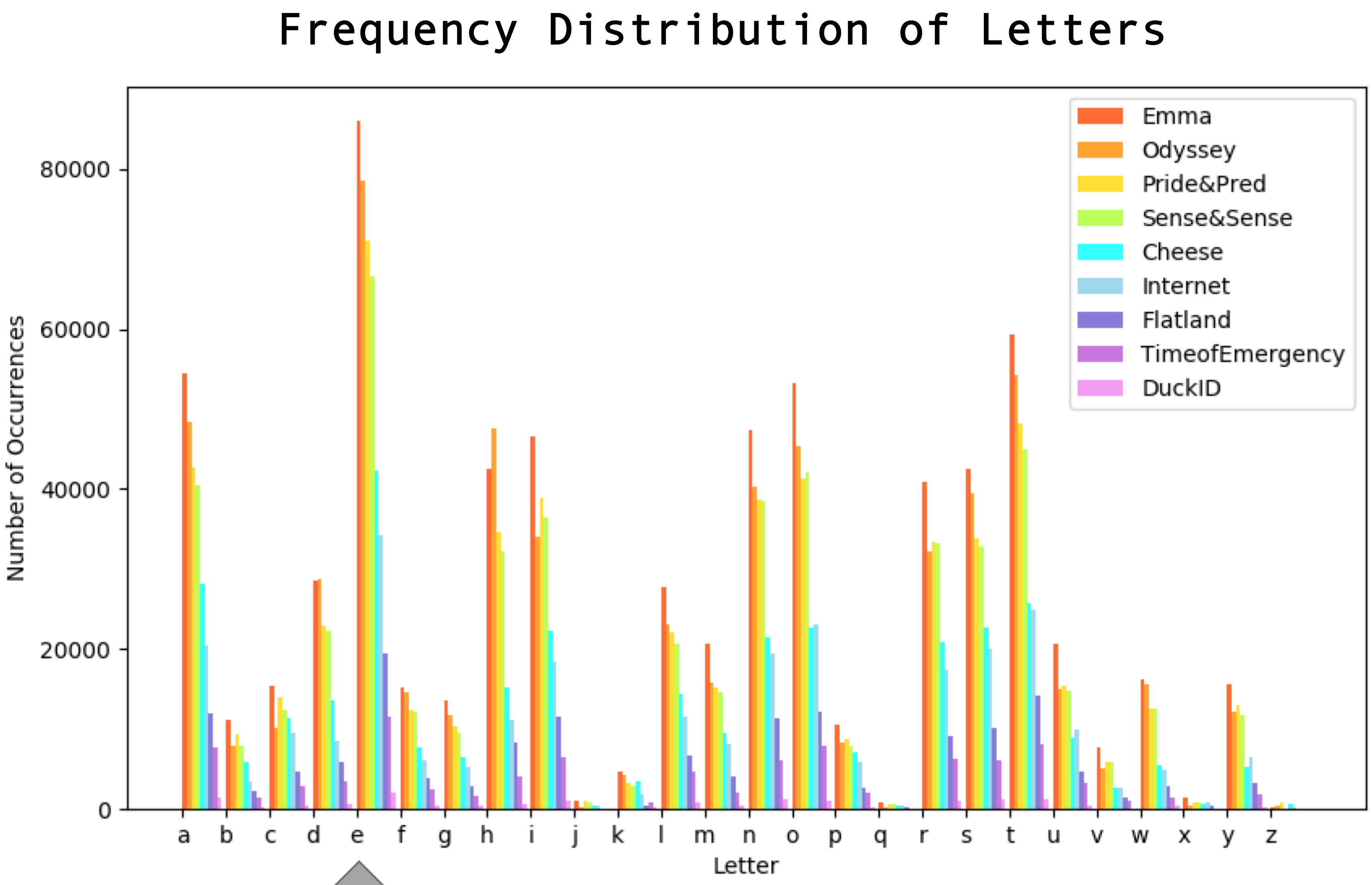
# Homework: Semantic Analyses

Elizabeth Johnson

Spring 2019

# Books

1. Jane Austen: Pride and Prejudice (1813)
2. Jane Austen: Emma (1815)
3. Jane Austen: Sense and Sensibility (1811)
4. Bob Hines: Ducks at a Distance: Waterfowl Identification Guide
5. Bob Brown: The Complete Book of Cheese (1955)
6. Edwin Abbott: Flatland: A Romance of Many Dimensions (1884)
7. Department of Defense: In Time of Emergency: A Citizen’s Handbook on Nuclear Attack and Natural Disasters (1968)
8. Homer (transl. by S.H. Butcher and A. Lang): The Odyssey (~8th century)
9. Electronic Frontier Foundation: Big Dummy’s Guide to the Internet (1993)



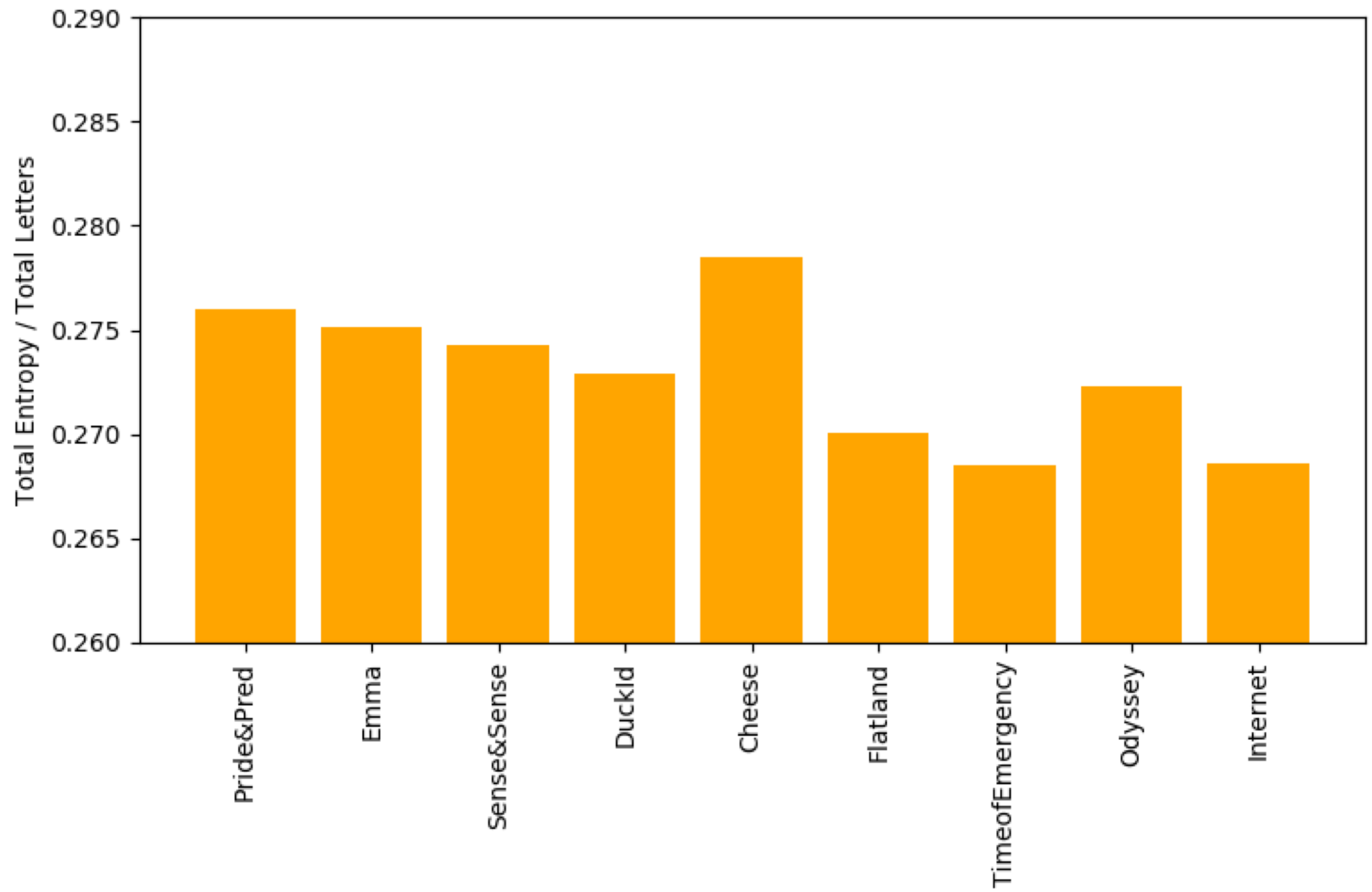
Here we are clearly seeing certain letters are used far more frequently than others. “e” is used most often while “j”, ”q”, “x”, and “z” are hardly used at all. This means that the probability distribution of letters is not constant.

I was expecting vowels like “a”, “e”, “i”, etc. to be used most often- and they are!

# Computed letter-based entropy for each book

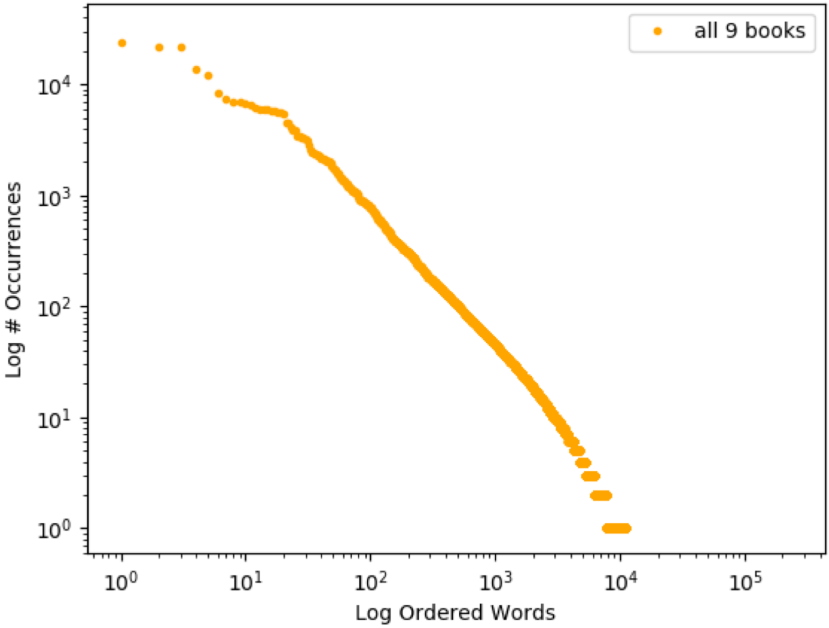
(using each book’s individual letter probabilities of occurrence): $H = -\sum P_i \log P_i$

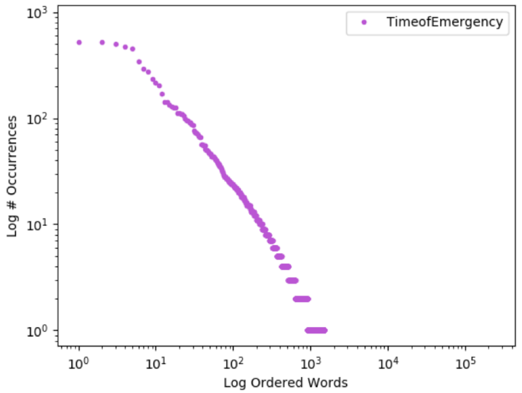
Then divided by total letters in the book to compare (obviously the longest book had the highest total entropy… so dividing by total letters normalizes them)

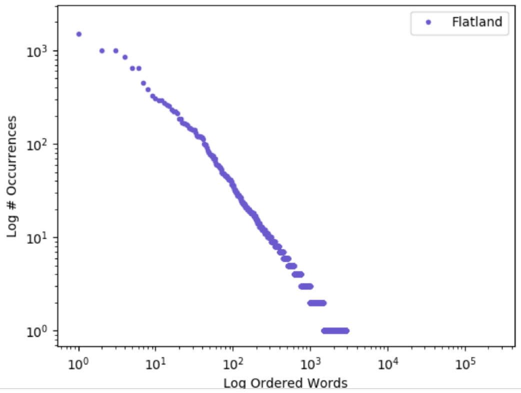


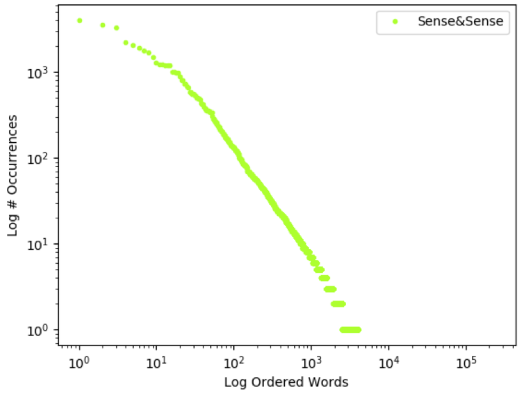
It looks like the book about cheese contains the most information per letter (i.e. the most unexpected combinations of letters) and the Time of Emergency book contains the least. This probably speaks to the content of the books. The cheese book has weird words (cheese names) and the Time of Emergency is plain English.

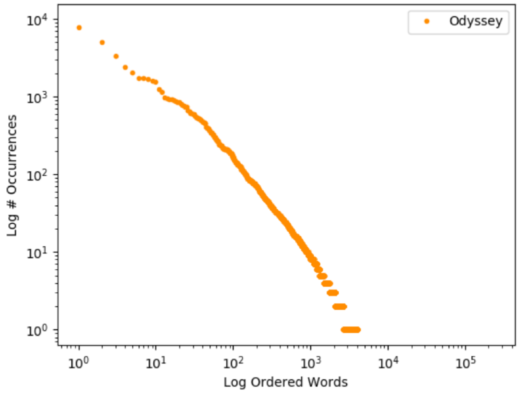
# Zipf’s Law

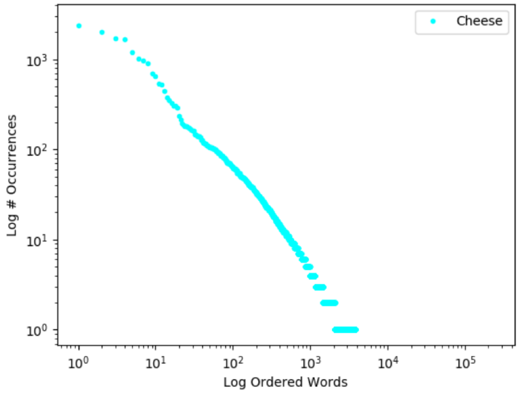

all 9 books
Log # Occurrences
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words

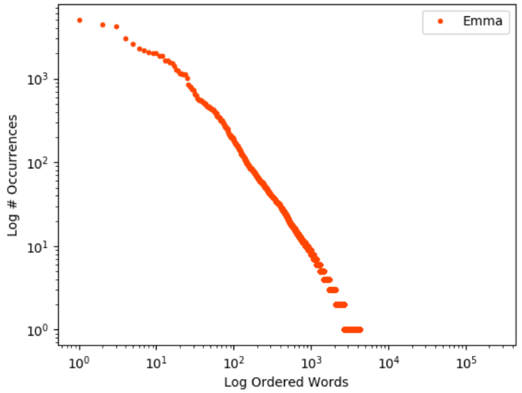

TimeofEmergency
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words

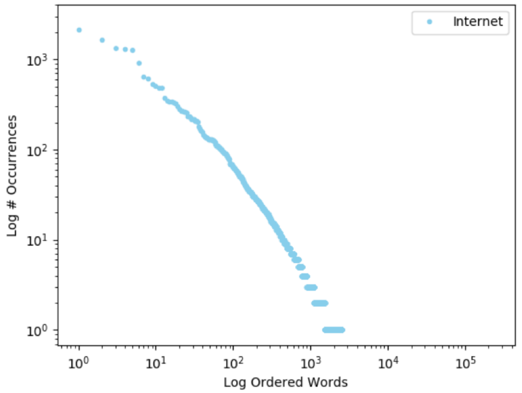

Flatland
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words

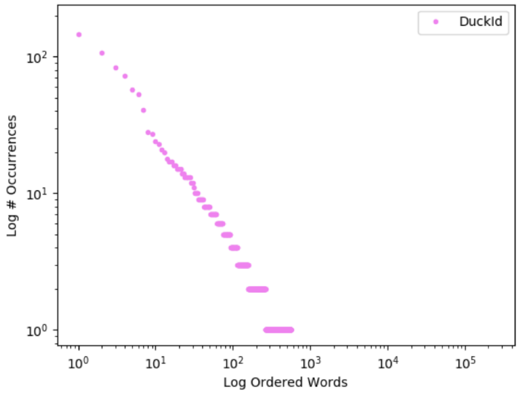

Sense&Sense
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words

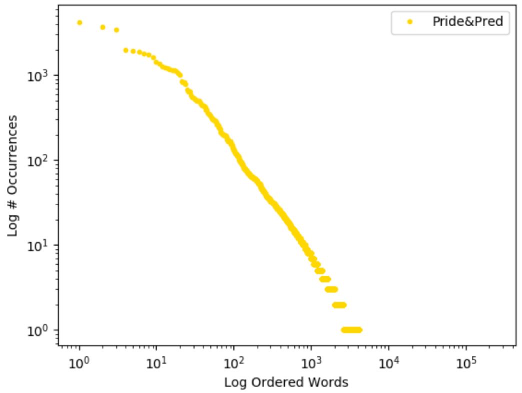

Odyssey
Log # Occurrences
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words


Cheese
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words

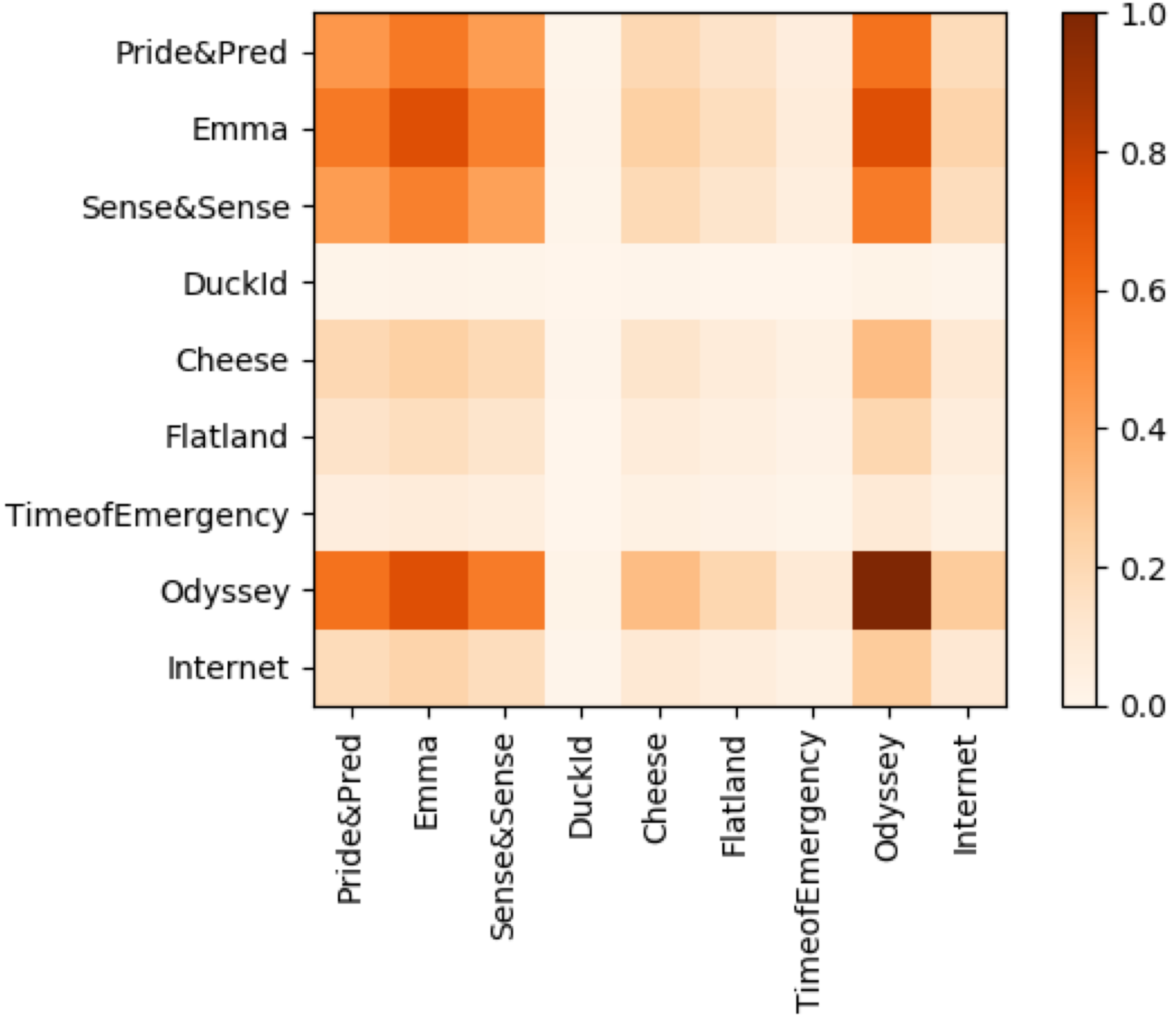

Emma
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words


Internet
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words


DuckId
Log # Occurrences
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words


Pride&Pred
Log # Occurrences
$10^3$
$10^2$
$10^1$
$10^0$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Log Ordered Words

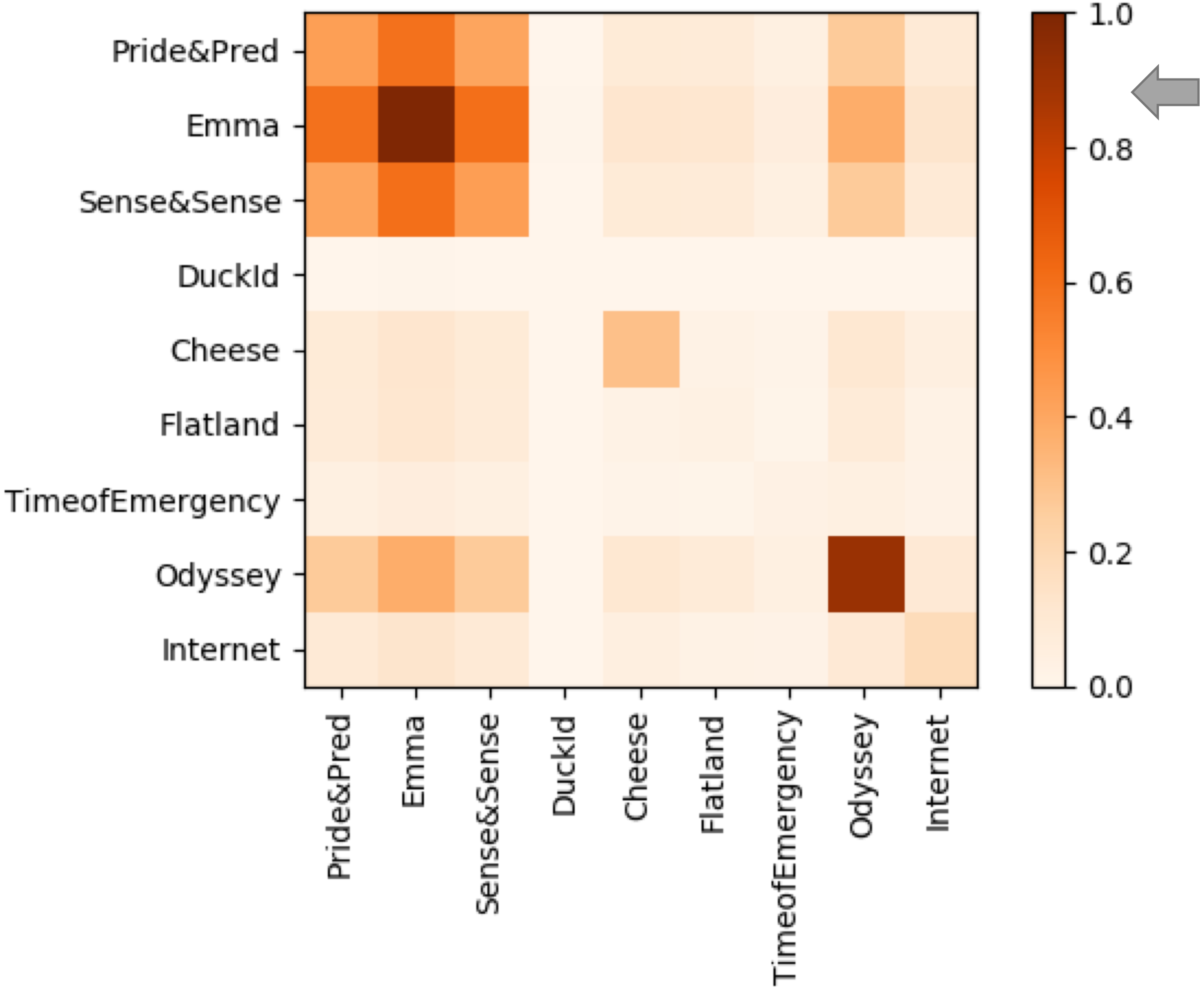
## **With Stopwords Still Included**: Could Confuse Homer with Jane Austen



I normalized the dot product values (instead of having really large numbers here)

At this point, the most similar works appear to be the Jane Austen books and Homer’s Odyssey. I expected this would be the case since they’re all lengthy stories.

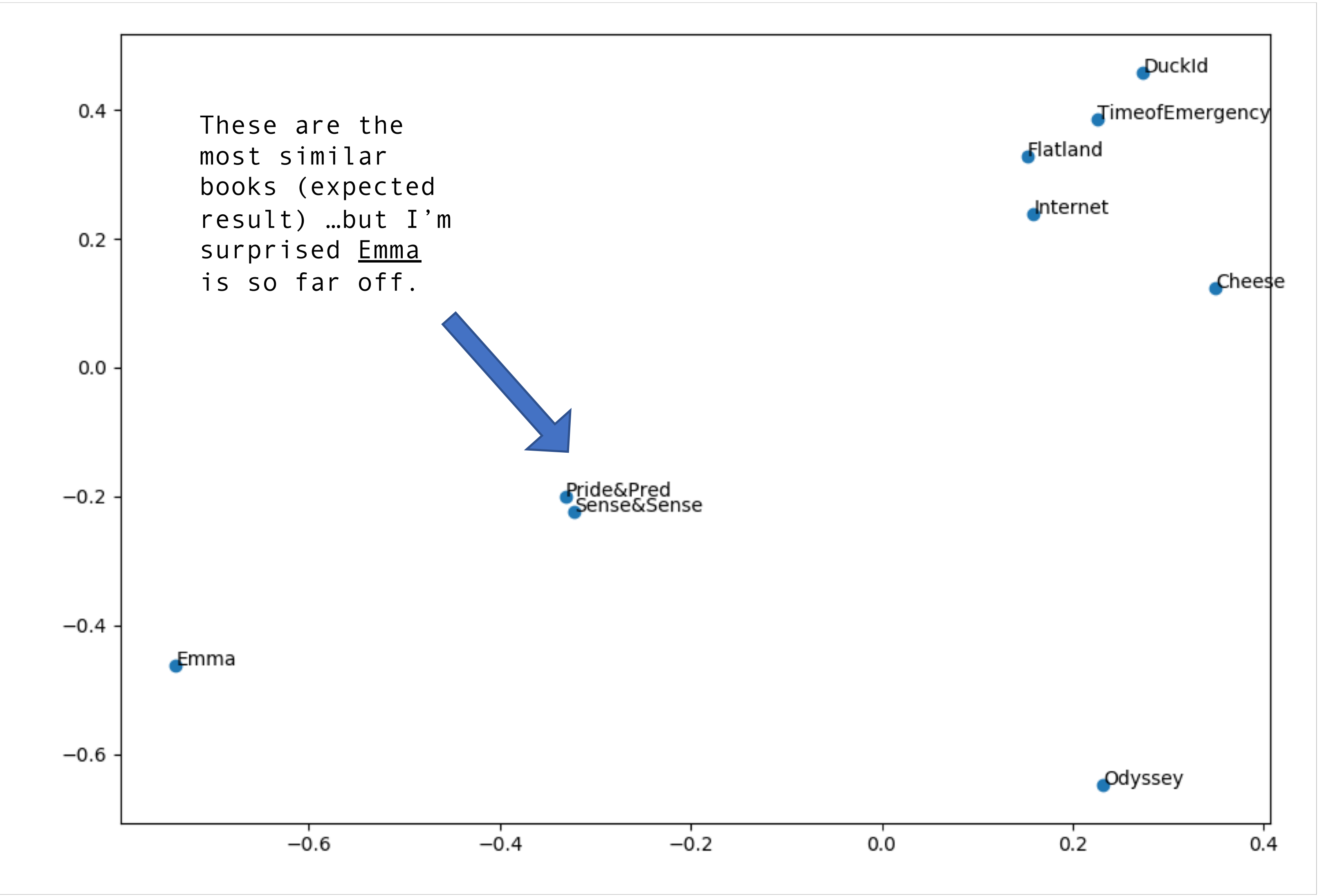
**With Stopwords NOT Included**: Jane Austen Books clearly similar in upper left corner



I normalized the dot product values (instead of having really large numbers here)

The results have definitely improved because common words like “a”, “the”, etc. were distorting the dot products. The longer books had more of those words so The Odyssey appeared like Jane Austen’s books. Now, it does not. Now, the dot products depend more on the content of the text. A word like “cheese” probably doesn’t occur in the Jane Austen books as much as the cheese guide book.

# **I’m not confident that I did this properly… but here goes:



These are the most similar books (expected result) …but I’m surprised Emma is so far off.

Is it like support vector machines? The mapping similarities concept feels familiar.

Except I haven’t done any machine learning to get my correlation matrix. …But it learns how to get the MDS? I’m using sklearn.manifold.MDS

# Lost Caravaggio Painting

I expect that Caravaggio painting experts can tell that he painted this newly discovered work of art because they know details of his signature style. The experts probably immediately noted that the subject matter of this new painting is similar to the grotesque, dark nature of his other works. Then, they probably inspected the colors and patterns in how he paints subjects. Maybe Caravaggio has a special way of painting ears that a trained expert would notice. Furthermore, maybe he has a unique brushstroke pattern resulting from the way he holds the paintbrush unlike all others. Knowing the location it was found could also hold clues to the painter- maybe it was found in the family home of one of his commissioners. I trust the experts know these details and can make a highly probable educated guess.

# Discovered an old, anonymous book!

To discover the book's authorship I'm thinking about using a method similar to problems 4 and 5 where we used dot products to find the most similar books. However, I can see this method would be difficult if the content of the mystery book were different from their other works. Maybe Jane Austen wrote a book about ducks so my method would confuse it with the waterfowl book in my list. But I know that Jane Austen has a signature style of penmanship that someone who reads and likes her books would know (my mom). There's more to it than words. Perhaps groups of words or phrases are common in her books. She might write that people say, "How do you do?" upon meeting, which is not a phrase I'd expect in my other books.

So I need to create a script that looks for groups of words. I could split my books into sentences and check if there are any identical sentences.

**Upon investigation I found that Jane Austen writes "Heaven forbid!" quite often. (My mom also picked up saying that phrase when I was growing up- haha.)

## Sentences In Common Between Pride & Prejudice and Emma

- {'Dear me!',

'Heaven forbid!',

'I never heard of such a thing.',

'Oh!',

'So much the better.',

'Very odd!',

'Well!',

'What are we to do?',

'no.'}

## Sentences In Common Between Pride & Prejudice and Sense & Sensibility

- {'Good gracious!’,

'Lord bless me!’,

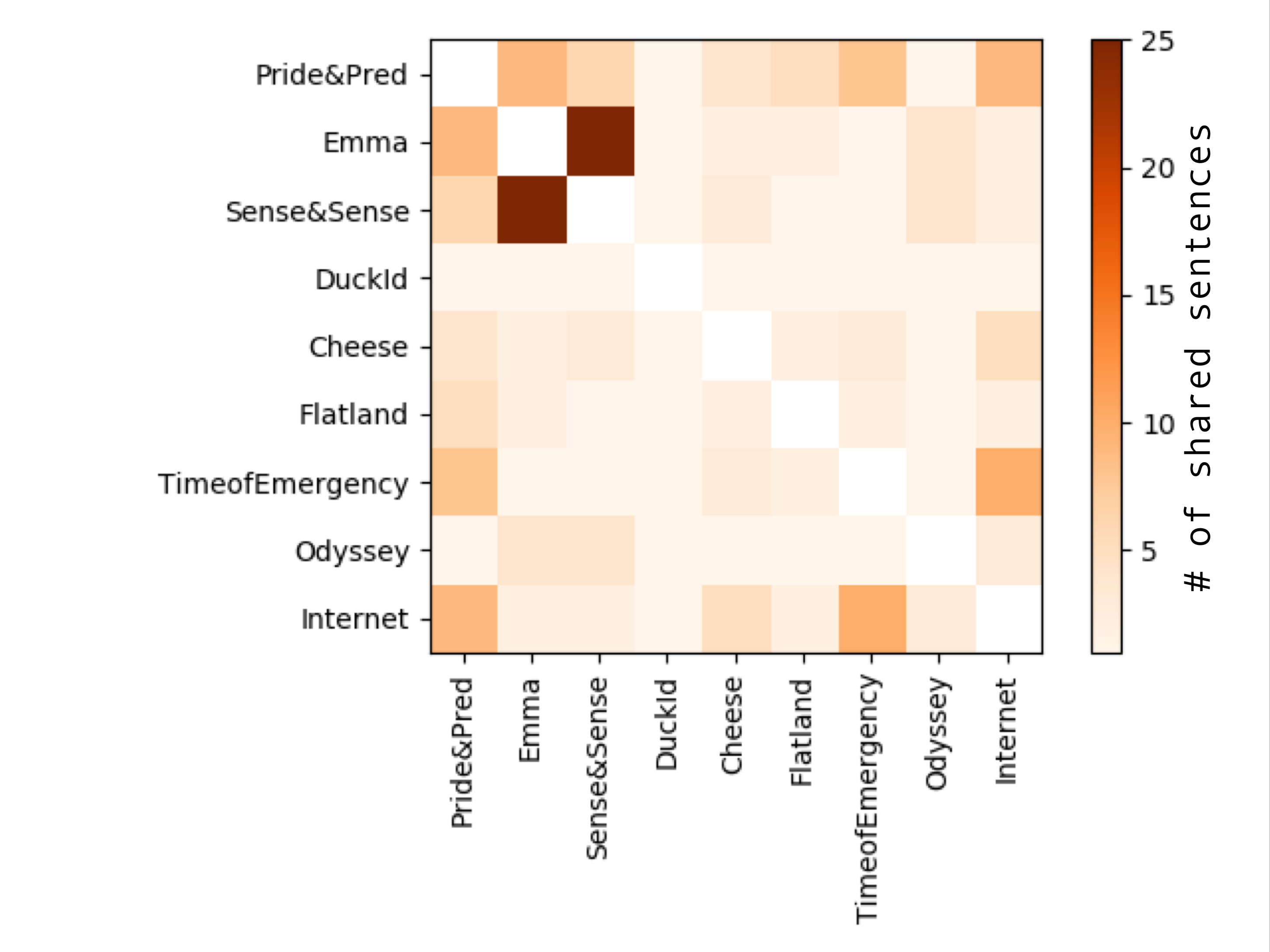
'Lord!’,

'Oh!’,

'Thank Heaven!’,

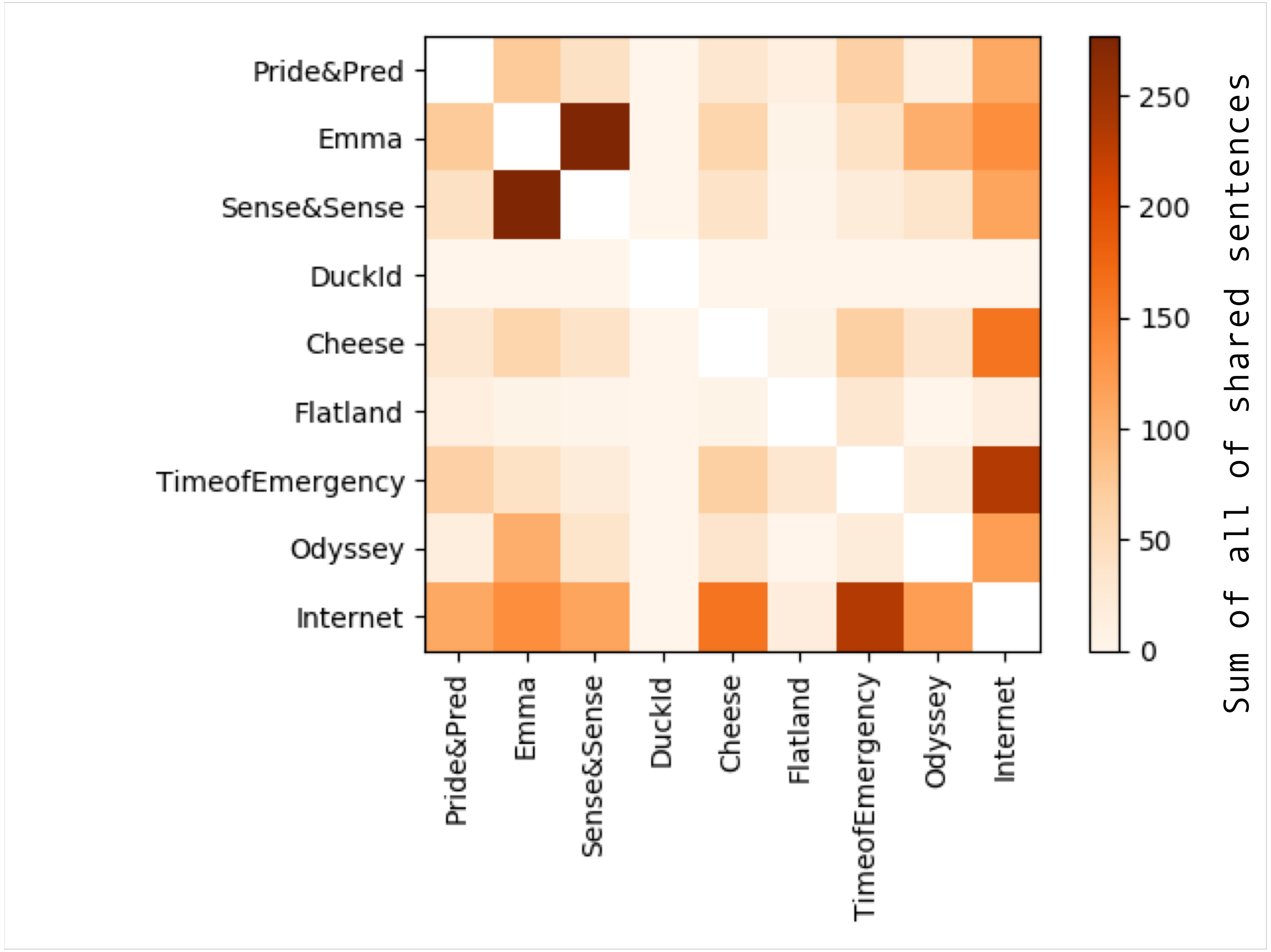
'Well!'}

# Shared Sentences



Clearly Emma and Sense & Sensibility are the most similar books. But it's also interesting here that the internet book and the handbook on emergencies are similar. This may be unsurprising because both were written with simple English words to a general American audience in the 90's.

# Sum of Total Shared Sentences



This method of adding up all the shared sentences (rather than just how many types of shared sentences) seems worse. I think what's happening is that the internet book and emergency guide are both guide books with directions so they share common directional phrases.

As is, I'd use my first method of just counting how many shared sentences there are rather than summing them.

# To Generalize:

Undoubtedly, my method needs refinement to use it in a larger application. It would be computationally expensive to ingest all books, separate out their sentences, and count up the similarities. (My script now takes a few seconds for 9 books.) But perhaps experts have already estimated the mystery book was from a specific time period and there is a list of candidate authors- this would make the task more manageable.