Plan:

1. Introduce TF-IDF
2. Work through example of TF-IDF

Text Analysis: TF-IDF

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

TF-IDF:
Term Frequency - Inverse Document Frequency

Term Frequency (TF) : how frequently a word occurs in a document

Inverse document frequency (IDF) : intended to measure how important a word is to a document

decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

TF-IDF:

Term Frequency - Inverse Document Frequency

the frequency of a term adjusted for how rarely it is used

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

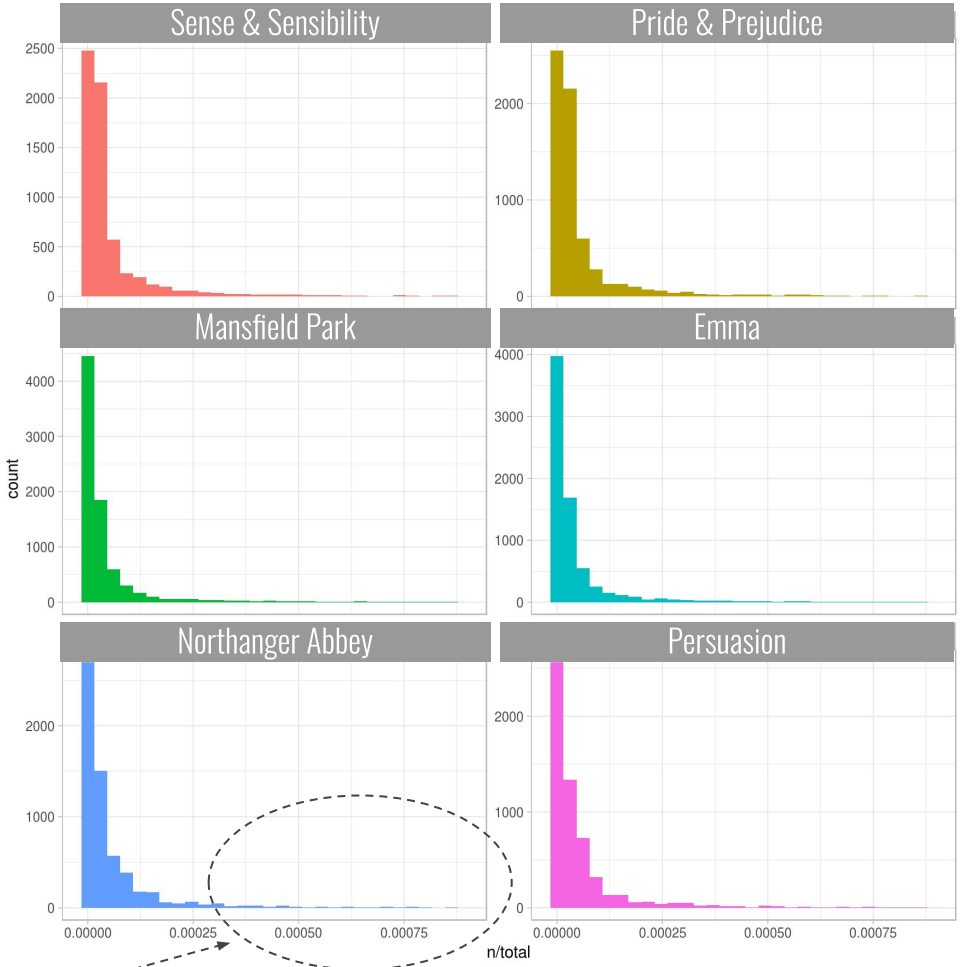
N = total number of documents

What are the most commonly used words in Jane Austen's novels?

Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

Calculating TF-IDF attempts to find the words that are important (i.e., common) in a text, but not *too* common

Frequency Distribution in Jane Austen's Novels



The long tails in each plot are those very frequent words

	book	word	n	total	tf	idf	tf_idf
	<fct>	<chr>	<int>	<int>	<dbl>	<dbl>	<dbl>
1	Mansfield Park	the	6206	160460	0.0387	0	0
2	Mansfield Park	to	5475	160460	0.0341	0	0
3	Mansfield Park	and	5438	160460	0.0339	0	0
4	Emma	to	5239	160996	0.0325	0	0
5	Emma	the	5201	160996	0.0323	0	0
6	Emma	and	4896	160996	0.0304	0	0
7	Mansfield Park	of	4778	160460	0.0298	0	0
8	Pride & Prejudice	the	4331	122204	0.0354	0	0
9	Emma	of	4291	160996	0.0267	0	0
10	Pride & Prejudice	to	4162	122204	0.0341	0	0

... with 40,369 more rows

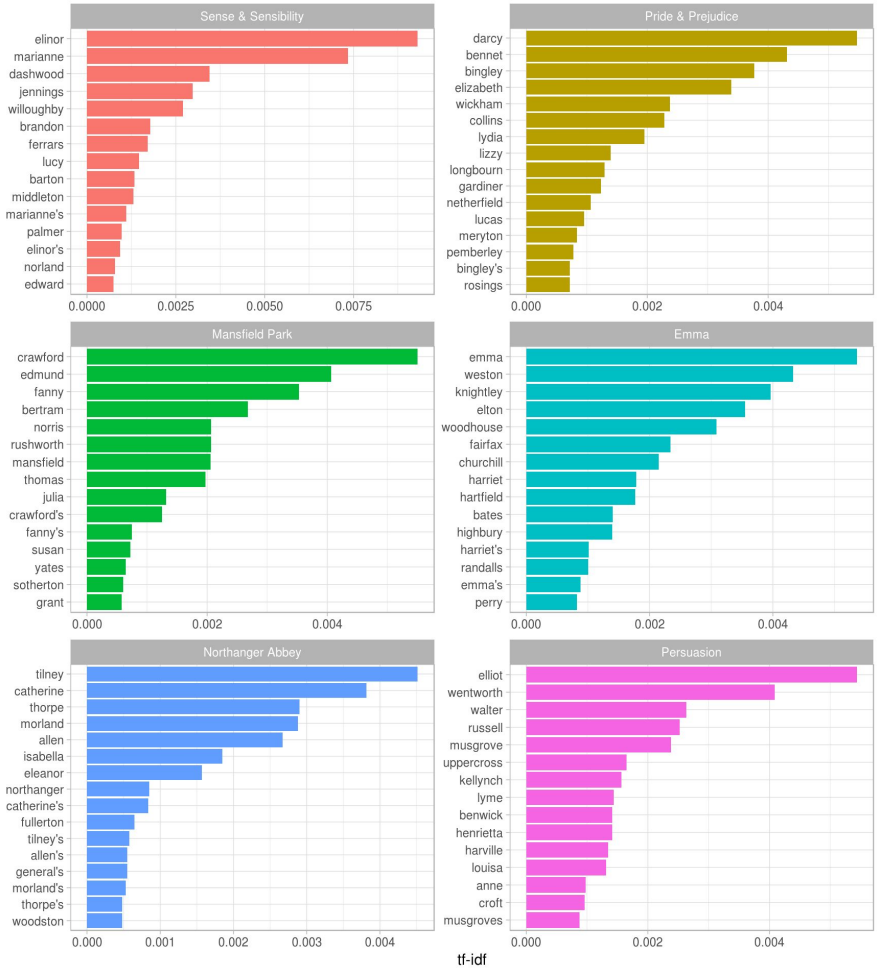
Super common words will have TF-IDF of zero....since they occur frequently across all documents

	book	word	n	tf	idf	tf_idf
	<fct>	<chr>	<int>	<dbl>	<dbl>	<dbl>
1	Sense & Sensibility	elinor	623	0.00519	1.79	0.00931
2	Sense & Sensibility	marianne	492	0.00410	1.79	0.00735
3	Mansfield Park	crawford	493	0.00307	1.79	0.00551
4	Pride & Prejudice	darcy	373	0.00305	1.79	0.00547
5	Persuasion	elliot	254	0.00304	1.79	0.00544
6	Emma	emma	786	0.00488	1.10	0.00536
7	Northanger Abbey	tilney	196	0.00252	1.79	0.00452
8	Emma	weston	389	0.00242	1.79	0.00433
9	Pride & Prejudice	bennet	294	0.00241	1.79	0.00431
10	Persuasion	wentworth	191	0.00228	1.79	0.00409

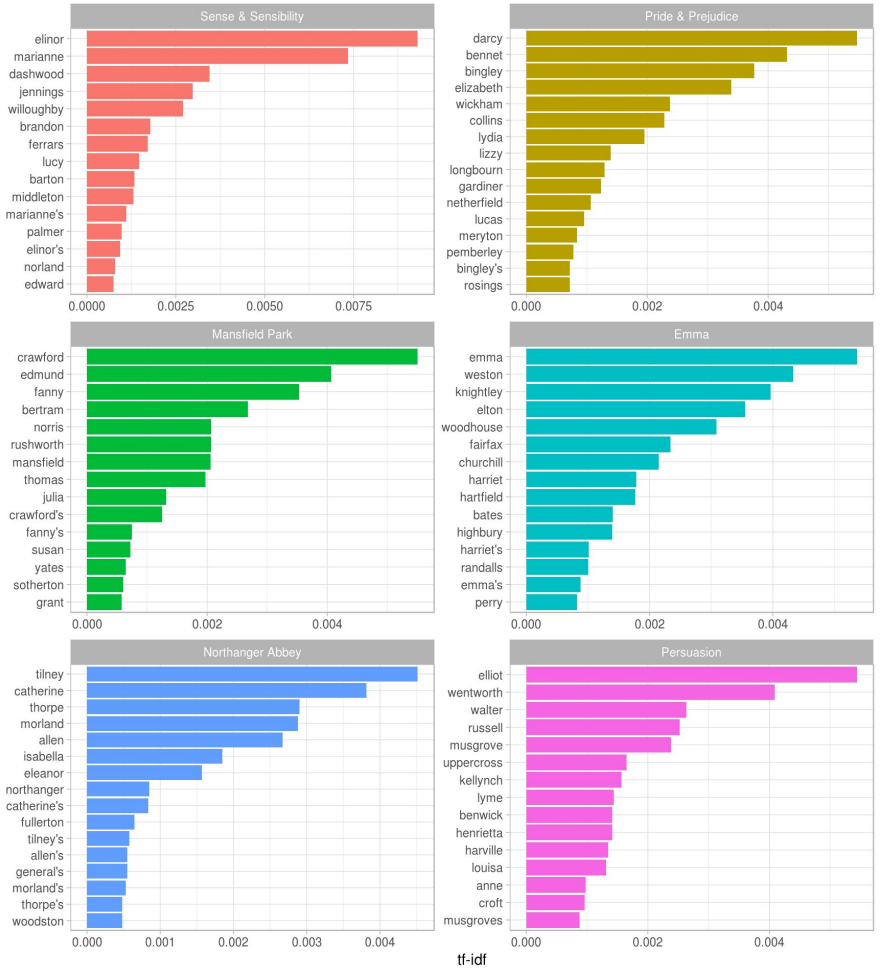
... with 40,369 more rows

Proper nouns, like character names, are important to a specific novel, and have a higher TF-IDF

High TF-IDF words broken down by Austen novel



Can conclude that "Jane Austen used similar language across her six novels, and *what distinguishes one novel from the rest within the collection of her works are the proper nouns*, the names of people and places"



A quick look at TF-IDF in another corpus: classic physics texts from Project Gutenberg