2.2 Feature Selection

Using the point-wise mutual information (PMI) measure, we explored the most discriminating n-grams for each author: Charles Dickens, Arthur Doyle, and Mark Twain.

c) A table of the top 10 unigram features for each author (highest positive PMI value):

	Dickens	Doyle	Twain
1	('her',)	('is',)	('t',)
2	('with',)	('you',)	('and',)
3	('his',)	('had',)	('but',)
4	('in',)	('i',)	('it',)
5	('to',)	('that',)	('a',)
6	('of',)	('his',)	('was',)
7	('and',)	('the',)	('he',)
8	('had',)	('of',)	('to',)
9	('the',)	('he')	('the',)
10	('that',)	('it',)	('i',)

Unigram features for all three authors very similar; there are many common features. Therefore, they cannot give any information for author identification or for characterization of author style. Though there are miserable differences: each author has at least one unique unigram: Dickens – (‘her’), Doyle – (‘you’), Twain - ('t'), ('but'), ('was',), however it cannot provide strong assurance.

d) A table of the top 10 bigram features for each author (highest positive PMI value):

	Dickens	Doyle	Twain
1	('of', 'his')	('it', 'is')	('<s>', 'and')
2	('to', 'be')	('<s>', 'you')	('<s>', 'the')
3	('that', 'i')	('i', 'have')	(('it', '</s>')
4	('i', 'have')	('that', 'i')	('<s>', 'he')
5	('<s>', 'and')	('<s>', 'it')	('in', 'a')
6	('in', 'a')	('on', 'the')	('he', 'was')
7	('<s>', 'but')	('at', 'the')	('<s>', 'but')
8	('he', 'was')	('and', 'the')	('was', 'a')
9	('don', 't')	('it', 'was')	('in', 'the')
10	('with', 'a')	('of', 'the')	('of', 'the')

Bigram features provide more information for characterization of author style or author identification. Each author has specific features that could point out at characteristic structures in their works:

- Dickens: ('<s>', 'but'), ('don', 't') – frequently used negation and imperative mood.

- Doyle: ('it', 'is'), ('it', 'was') – explanations of what happened; ('<s>', 'you') – a lot of dialogs.
- Twain: ('he', 'was'), ('was', 'a') – narration in the past tense; ('<s>', 'he'), ('he', 'was') – male heroes.