1

2 Text Classification

2.1 Author indefication using LMs

Dickens				Doyle				Twain			
Unigrams		Bigrams		Unigrams		Bigrams		Unigrams		Bigrams	
'the'	65440	'< s >', 'i'	7392	'the'	72028	'< s >', 'i'	10863	'the'	67253	'< s >', 'i'	7105
'and'	50579	('of', 'the'	6559	'of'	34862	'of', 'the'	8869	'and'	59974	'of', 'the'	7029
'to'	36773	'in', 'the'	6135	'and'	34405	'in', 'the'	6003	'a'	34965	'in', 'the'	5607
'of'	36288	'< s >', 'the'	4212	'i'	31920	'< s >', 'it'	5271	'to'	33284	'< s >', 'the'	4997
'a'	28874	'to', 'the'	3495	'a'	31080	'< s >', 'the'	5181	'of'	30456	'< s >', 'he'	4062
'i'	26157	'< s >', 'he'	3188	'to'	30829	'< s >', 'he'	4855	'i'	26541	('it', 'was'	3687
'in'	24205	'to', 'be'	2981	'that'	21884	'it', 'was'	4068	'it'	25226	'it', $'$ < $/s$ >$'$	3259
'that'	18512	'it', 'was'	2615	'it'	21448	'to', 'the'	3589	'was'	20767	'and', 'the'	3214
'was'	17260)	on', 'the'	2550	'in'	21173	'it', 'is'	3290	'in'	19669	'< s >', 'it'	3178
'it'	16969	'< s >', 'it'	2528	'was'	18735	'< s >', 'you'	3117	'he'	18982	'to', 'the'	3142
'he'	16635	'and', 'the'	2441	'he'	18161	'i', 'have'	2767	'that'	17877	'he', 'was'	2459
'his'	15938	'in', 'a'	2255	'you'	17336	'at', 'the'	2714	'you'	13064	'don', 't'	2379
'her'	12926	'< s >', 'but'	2194	'his'	14661	'it', $'$ < $/s$ >$'$	2600	't'	11101	'< s >', 'but'	2349
'with'	12582	'with', 'a'	2136	'is'	12663	'< s >', 'but'	2483	'his'	10703	'was', 'a'	2227
'you'	12313	'of', 'his'	2071	'had'	11265	'that', 'i'	2348	'but'	10442	'< s >', 'and'	2170)

c)

f Unigrams:

	Dickens	Doyle	Twain
DickensLM	**963**	**965**	830
DoyleLM	1100	669	865
TwainLM	1033	1001	**513**

g Unigrams:

	Dickens	Doyle	Twain
DickensLM	**1039**	809	919
DoyleLM	1281	**320**	1017
TwainLM	1402	1003	**286**