

	Skipgram	fastText, h=10, bigram	MDLText	
Number	2,00	6,00	34,00	
average	57,50%	79,87%	86,62%	
min	43,00%	60,20%	67,20%	
max	72,00%	95,70%	98,50%	
difference	29,00%	35,50%	31,30%	
type	dataset	parent dataset	Skipgram	fastText, h=10, bigram
news articles	20Newsgroups	-		
news articles	Irish economic	-		
news articles	La1S	TREC-5		
news articles	La2S	TREC-5		
news articles	Latimes	LA Times		
news articles	RCV1 Top-4	-		
news articles	Re0	Reuters-21578		
news articles	Re1	Reuters-21578		
news articles	Re8	Reuters-21578		
news articles	Reuters	-		
news articles	Reviews	San Jose Mercury newspaper		
news articles	New3S	-		
news articles	AG News	-		92,50%
web pages	7Sectors	-		
web pages	Techtc300-1092-135724	-		
web pages	Techtc300-1092-789236	-		
scientific documents	ACM	-		
scientific documents	CSTR	Computer Science Technical Reports		
scientific documents	Oh0	HSUMED		
scientific documents	Oh10	HSUMED		
scientific documents	Oh15	HSUMED		
scientific documents	Oh5	HSUMED		
scientific documents	Ohscal	HSUMED		
scientific documents	Pubmed-Cancer	-		
scientific documents	Pubmed-Cancer-2000	-		
Mail	Enron	-		
Mail	Trec7-3000	-		
Q&A	Opinosis	-		
Q&A	Yelp P.	-		95,70%
Q&A	Yelp F	-		63,90%
Q&A	Yah A	-		72,30%
Q&A	Amz F	-		60,20%
Q&A	Amz P	-		94,60%

MDLText
90,20%
67,20%
87,70%
88,30%
84,70%
92,00%
81,50%
81,20%
91,40%
91,20%
90,30%
85,80%
91,30%
97,80%
98,50%
82,20%
89,20%
92,50%
81,90%
85,90%
90,70%
78,00%
92,00%
84,80%
71,60%
98,10%
67,40%

Text Retrival Conference	Tr11	TRE		
Text Retrival Conference	Tr12	TRE		
Text Retrival Conference	Tr21	TRE		
Text Retrival Conference	Tr23	TRE		
Text Retrival Conference	Tr31	TRE		
Text Retrival Conference	Tr41	TRE		
Text Retrival Conference	Tr45	TRE		
word pairs	RW	-	43,00%	
word pairs	WS353	-	72,00%	

82,70%
87,90%
83,60%
92,10%
82,60%
92,40%
90,30%