



National University
Of Computer and Emerging Sciences

Name:

Dawood Sarfraz

Roll no:

20P-0153

Section:

BSCS-7A

Course:

Natural Language Processing

Lab:

01

Submitted to:

Dr. Omer Usman Khan

FAST National University of Computer and Emerging Sciences

Using the above information, fill the following table:

Corpus	Text1	Text2	Text3	Text4	Text5	Text6	Text7	Text8	Text9
Corpus Name	Moby Dick by Herman Melville 1851	Sense and Sensibility by Jane Austen 1811	The Book of Genesis	Inaugural Address Corpus	Chat Corpus	Monty Python and the Holy Grail	Wall Street Journal	Personals Corpus	The Man Who Was Thursday by G . K . Chesterton 1908
Corpus Length	260819	141576	44764	152901	45010	16967	100676	4867	69213
Unique Words	19317	6833	2789	10025	6066	2166	12408	1108	6807
Lexical Richness	0.074062855 85022564	0.048263 83002768 831	0.06230 453042 623537	0.065565 3004231 4962	0.1347 700510 997556 2	0.1276 59574 46808 51	0.1232 468512 853112 9	0.227655 64002465 585	0.09834857613 45412

===== Names of Texts Used =====

Name of Text 1 is ----->> <Text: Moby Dick by Herman Melville 1851>
 Name of Text 2 is ----->> <Text: Sense and Sensibility by Jane Austen 1811>
 Name of Text 3 is ----->> <Text: The Book of Genesis>
 Name of Text 4 is ----->> <Text: Inaugural Address Corpus>
 Name of Text 5 is ----->> <Text: Chat Corpus>
 Name of Text 6 is ----->> <Text: Monty Python and the Holy Grail>
 Name of Text 7 is ----->> <Text: Wall Street Journal>
 Name of Text 8 is ----->> <Text: Personals Corpus>
 Name of Text 9 is ----->> <Text: The Man Who Was Thursday by G . K . Chesterton 1908>

===== Length of Texts Used =====

Length of Text 1 is ----->> 260819
 Length of Text 2 is ----->> 141576
 Length of Text 3 is ----->> 44764
 Length of Text 4 is ----->> 152901
 Length of Text 5 is ----->> 45010
 Length of Text 6 is ----->> 16967
 Length of Text 7 is ----->> 100676
 Length of Text 8 is ----->> 4867
 Length of Text 9 is ----->> 69213

===== Unique Words of Texts Used =====

Length of Unique words in Text 1	----->>	19317
Length of Unique words in Text 2	----->>	6833
Length of Unique words in Text 3	----->>	2789
Length of Unique words in Text 4	----->>	10025
Length of Unique words in Text 5	----->>	6066
Length of Unique words in Text 6	----->>	2166
Length of Unique words in Text 7	----->>	12408
Length of Unique words in Text 8	----->>	1108
Length of Unique words in Text 9	----->>	6807

===== Lexical Richness of Texts Used =====

Lexical Richness of Text 1	----->>	0.07406285585022564
Lexical Richness of Text 2	----->>	0.04826383002768831
Lexical Richness of Text 3	----->>	0.06230453042623537
Lexical Richness of Text 4	----->>	0.06556530042314962
Lexical Richness of Text 5	----->>	0.13477005109975562
Lexical Richness of Text 6	----->>	0.1276595744680851
Lexical Richness of Text 7	----->>	0.12324685128531129
Lexical Richness of Text 8	----->>	0.22765564002465585
Lexical Richness of Text 9	----->>	0.0983485761345412

Now, fill the table below:

	Text 1		
Tokens	TF()	LOGTF()	IDF()
monster	0.018786974875296663	1.6989700043360185	-0.7359535705891886
evil	0.004217484155678842	1.0791812460476247	0.08715017571890013
devil	0.01955379017632918	1.716003343634799	0.7533276666586114
the	5.260736372733581	4.137417414990392	-3.1831432548946452
Common word ,	7.174707364110744	4.272166625140787	-3.3179009081517252
Common word the	5.260736372733581	4.137417414990392	-3.1831432548946452
Common word .	2.630943297842565	3.836513998890671	-2.8822082042808295

Start for word ----->> monster

```
Term Frequency of "monster" ----->> 0.018786974875296663
Term Frequency Log "monster" ----->> 1.6989700043360185
Inverse Document Frequency "monster" ----->> -0.7359535705891886
=====
```

Start for word ----->> evil

```
Term Frequency of "evil" ----->> 0.004217484155678842
Term Frequency Log "evil" ----->> 1.0791812460476247
Inverse Document Frequency "evil" ----->> -0.08715017571890013
=====
```

Start for word ----->> devil

```
Term Frequency of "devil" ----->> 0.01955379017632918
Term Frequency Log "devil" ----->> 1.716003343634799
Inverse Document Frequency "devil" ----->> -0.7533276666586114
=====
```

Start for word ----->> the

```
Term Frequency of "the" ----->> 5.260736372733581
Term Frequency Log "the" ----->> 4.137417414990392
Inverse Document Frequency "the" ----->> -3.1831432548946452
=====
=====
```

Start for Most Common word ----->> ,

Term Frequency of "," ----->> 7.174707364110744

Term Frequency Log of"," ----->> 4.272166625140787

Inverse Document Frequency of "," ----->> -3.3179009081517252

=====

Start for Most Common word ----->> the

Term Frequency of "the" ----->> 5.260736372733581

Term Frequency Log of"the" ----->> 4.137417414990392

Inverse Document Frequency of "the" ----->> -3.1831432548946452

=====

Start for Most Common word ----->> .

Term Frequency of "." ----->> 2.630943297842565

Term Frequency Log of"." ----->> 3.836513998890671

Inverse Document Frequency of "." ----->> -2.8822082042808295

=====