Ragheb AlGhezi
LING581 – Assignment#4

Question1

i. How many words in the corpus?

The number of words without NON-words is 1037490

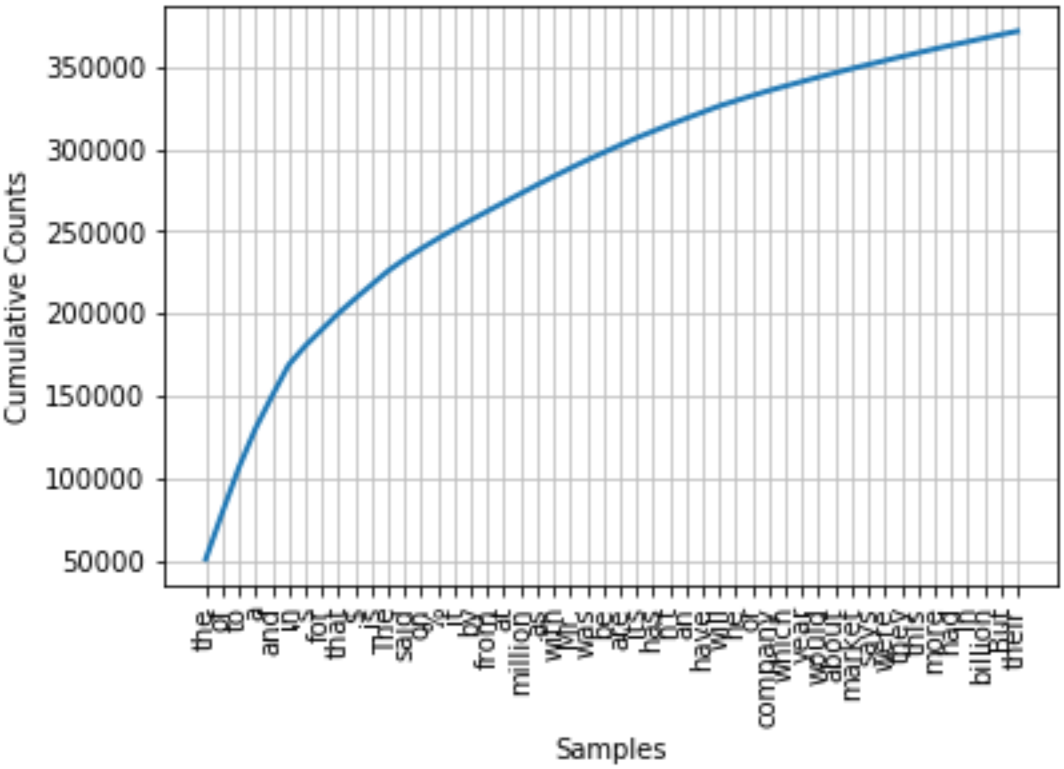
ii. How many distinct words?

The number of distinct words without NON-words is 49184

iii. Compute the lexical diversity (to 3 significant figures)

Lexical diversity is 0.047406721992501136 (4.7%)

iv. Plot the cumulative frequency distribution graph



v. How many top frequency words do you need to account for at least 50% of the words in the corpus?

The number of top word that constitute 50% of the corpus is 217

vi. Print those words

Ragheb AlGhezi
LING581 – Assignment#4

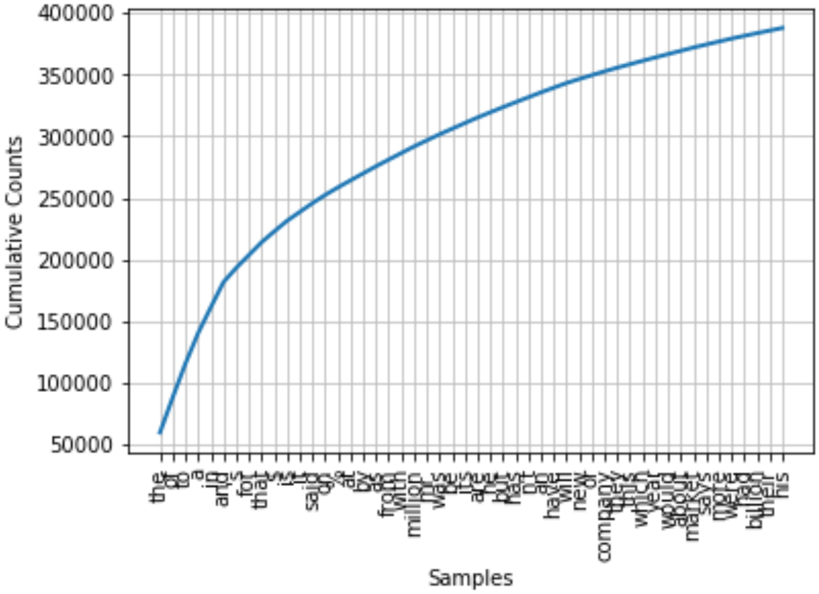
```
['the', 'of', 'to', 'a', 'and', 'in', "'s", 'for', 'that', '$', 'is', 'The', 'said', 'on', '%', 'it', 'by', 'from',
'at', 'million', 'as', 'with', 'Mr.', 'was', 'be', 'are', 'its', 'has', "n't", 'an', 'have', 'will', 'he',
'or', 'company', 'which', 'year', 'would', 'about', 'market', 'says', 'were', 'they', 'this',
'more', 'had', 'In', 'billion', 'But', 'their', 'up', 'U.S.', 'but', 'than', 'his', 'been', 'who',
'share', 'new', 'also', 'one', 'other', 'not', 'stock', 'some', 'Corp.', 'New', 'I', 'shares', 'years',
'It', 'trading', 'Inc.', 'all', 'could', 'last', 'two', '&', 'out', 'because', 'when', 'after', 'do',
'sales', 'can', 'York', 'only', 'into', 'Co.', 'president', 'A', 'such', 'over', 'first', 'He', 'business',
'if', 'most', 'may', 'companies', 'cents', 'government', 'prices', 'any', 'quarter', 'we',
'down', 'time', 'many', 'no', 'say', 'price', 'there', 'months', 'now', 'much', 'week',
'investors', 'rose', 'them', 'people', 'so', 'group', 'yesterday', 'bonds', "'", 'stocks', 'even',
'you', 'did', 'interest', '1', 'earnings', 'major', 'make', 'next', 'American', 'We', 'three',
'through', 'what', 'net', 'chief', 'executive', 'just', 'And', 'under', 'before', '10', 'Friday',
'expected', 'off', 'earlier', 'industry', 'made', 'money', 'program', 'unit', 'those', 'month',
'investment', 'rate', 'federal', 'days', 'while', 'officials', 'still', '30', 'like', 'does', 'sell',
'since', 'buy', 'between', 'against', 'Exchange', 'financial', 'profit', 'firm', 'plan', 'ago',
'Japanese', 'get', 'recent', 'That', 'They', 'big', 'For', 'income', 'chairman', 'back', 'rates',
'state', 'own', 'fell', 'offer', 'should', 'issue', 'well', 'markets', 'debt', 'bank', 'these',
'analysts', 'products', 'securities', 'higher', 'funds', 'including', 'take', '15', 'part', 'This', '8',
'she', '1988', 'Japan']
```

Question 2

I. **The number of words without NON-words is 1037490**

II. **The number of distinct words without NON-words is 43746**

III. **Lexical diversity is 0.0421652256889223 (4.2%)**



IV. **the number of top word that constitute 50% of the corpus is 176**

V. Those words are: ['the', 'of', 'to', 'a', 'in', 'and', "'s", 'for', 'that', '$', 'is', 'it', 'said', 'on', '%', 'at', 'by', 'as', 'from', 'with', 'million', 'mr.', 'was', 'be', 'its', 'are', 'he', 'but', 'has', "n't", 'an', 'have', 'will', 'new', 'or', 'company', 'they', 'this', 'which', 'year', 'would', 'about', 'market', 'says', 'more', 'were', 'had', 'billion', 'their', 'his', 'up', 'u.s.', 'one', 'than', 'stock', 'been', 'some', 'who', 'also', 'other', 'share', 'not', 'we', 'corp.', 'when', 'last', 'if', 'i', 'all', 'shares', 'president', 'years', 'trading', 'first', 'two', 'after', 'inc.', 'because', 'could', 'sales', '&', 'out', 'there', 'do', 'only', 'business', 'such', 'most', 'can', 'co.', 'york', 'into', 'may', 'over', 'group', 'many', 'time', 'now', 'federal', 'companies', 'prices', 'no', 'government', 'so', 'any', 'cents', 'quarter', 'bank', 'investors', 'down', 'you', 'price', 'exchange', 'what', 'people', 'even', 'say', 'yesterday', 'much', 'big', 'while', 'months', 'securities', 'under', 'week', 'rose', 'them', 'bonds', 'stocks', 'major', 'next', 'three', 'net', 'interest', "''", 'earnings', 'did', 'financial', 'still', '1', 'make', 'chairman', 'american', 'just', 'earlier', 'board', 'through', 'investment', 'before', 'those', 'since', 'chief', 'industry', 'executive', 'these', 'state', 'money', 'national', 'program', 'off', 'officials', '10', 'friday', 'expected', 'made', 'analysts', 'like', 'rate', 'she', 'unit', 'month', 'markets', 'days', 'does', 'house', '30']