Text Summarization

1- Text Summarization with Word Frequencies:

In [31]:
```
from bs4 import BeautifulSoup as bs
'https://www.scientificamerican.com/article/nasas-james-webb-space-telescope-will-face-
import urllib.request as ur
import re,certifi
from pprint import pprint
```

1.1 - Scrapping & Processing:

In [32]:
```
def scrape_site(url):
    response = ur.urlopen(url,cafile=certifi.where())
    text = response.read()
    parsed = bs(text,'lxml')
    paras = parsed.findAll('p')
    paragraphs =''
    for p in paras:
        paragraphs+=p.text
    return paragraphs
src='https://en.wikipedia.org/wiki/Natural_language_processing'
text = scrape_site(src)
pprint(text)
```

```
<ipython-input-32-1fa1d72a8b93>:2: DeprecationWarning: cafile, capath and cadefault are
deprecated, use a custom context instead.
  response = ur.urlopen(url,cafile=certifi.where())
('Natural language processing (NLP) is a subfield of linguistics, computer '
 'science, and artificial intelligence concerned with the interactions between '
 'computers and human language, in particular how to program computers to '
 'process and analyze large amounts of natural language data.  The goal is a '
 'computer capable of "understanding" the contents of documents, including the '
 'contextual nuances of the language within them. The technology can then '
 'accurately extract information and insights contained in the documents as '
 'well as categorize and organize the documents themselves.\n'
 'Challenges in natural language processing frequently involve speech '
 'recognition, natural language understanding, and natural language '
 'generation.\n'
 'Natural language processing has its roots in the 1950s. Already in 1950, '
 'Alan Turing published an article titled "Computing Machinery and '
 'Intelligence" which proposed what is now called the Turing test as a '
 'criterion of intelligence, a task that involves the automated interpretation '
 'and generation of natural language, but at the time not articulated as a '
 'problem separate from artificial intelligence.\n'
 "The premise of symbolic NLP is well-summarized by John Searle's Chinese room "
 'experiment: Given a collection of rules (e.g., a Chinese phrasebook, with '
 'questions and matching answers), the computer emulates natural language '
 'understanding (or other NLP tasks) by applying those rules to the data it is '
 'confronted with.\n'
 'Up to the 1980s, most natural language processing systems were based on '
 'complex sets of hand-written rules.  Starting in the late 1980s, however, '
 'there was a revolution in natural language processing with the introduction '
 'of machine learning algorithms for language processing.  This was due to '
 "both the steady increase in computational power (see Moore's law) and the "
 'gradual lessening of the dominance of Chomskyan theories of linguistics '
 '(e.g. transformational grammar), whose theoretical underpinnings discouraged '
```

'the sort of corpus linguistics that underlies the machine-learning approach '
'to language processing.[6]\n'
'In the 2010s, representation learning and deep neural network-style machine '
'learning methods became widespread in natural language processing, due in '
'part to a flurry of results showing that such techniques[7][8] can achieve '
'state-of-the-art results in many natural language tasks, for example in '
'language modeling,[9] parsing,[10][11] and many others. This is increasingly '
'important in medicine and healthcare, where NLP is being used to analyze '
'notes and text in electronic health records that would otherwise be '
'inaccessible for study when seeking to improve care.[12]\n'
'In the early days, many language-processing systems were designed by '
'symbolic methods, i.e., the hand-coding of a set of rules, coupled with a '
'dictionary lookup:[13][14] such as by writing grammars or devising heuristic '
'rules for stemming.\n'
'More recent systems based on machine-learning algorithms have many '
'advantages over hand-produced rules: \n'
'Despite the popularity of machine learning in NLP research, symbolic methods '
'are still (2020) commonly used:\n'
'Since the so-called "statistical revolution"[15][16] in the late 1980s and '
'mid-1990s, much natural language processing research has relied heavily on '
'machine learning. The machine-learning paradigm calls instead for using '
'statistical inference to automatically learn such rules through the analysis '
'of large corpora (the plural form of corpus, is a set of documents, possibly '
'with human or computer annotations) of typical real-world examples.\n'
'Many different classes of machine-learning algorithms have been applied to '
'natural-language-processing tasks. These algorithms take as input a large '
'set of "features" that are generated from the input data. Increasingly, '
'however, research has focused on statistical models, which make soft, '
'probabilistic decisions based on attaching real-valued weights to each input '
'feature (complex-valued embeddings,[17] and neural networks in general have '
'also been proposed, for e.g. speech[18]). Such models have the advantage '
'that they can express the relative certainty of many different possible '
'answers rather than only one, producing more reliable results when such a '
'model is included as a component of a larger system.\n'
'Some of the earliest-used machine learning algorithms, such as decision '
'trees, produced systems of hard if-then rules similar to existing '
'hand-written rules. However, part-of-speech tagging introduced the use of '
'hidden Markov models to natural language processing, and increasingly, '
'research has focused on statistical models, which make soft, probabilistic '
'decisions based on attaching real-valued weights to the features making up '
'the input data. The cache language models upon which many speech recognition '
'systems now rely are examples of such statistical models. Such models are '
'generally more robust when given unfamiliar input, especially input that '
'contains errors (as is very common for real-world data), and produce more '
'reliable results when integrated into a larger system comprising multiple '
'subtasks.\n'
'Since the neural turn, statistical methods in NLP research have been largely '
'replaced by neural networks. However, they continue to be relevant for '
'contexts in which statistical interpretability and transparency is '
'required.\n'
'A major drawback of statistical methods is that they require elaborate '
'feature engineering. Since 2015,[19] the field has thus largely abandoned '
'statistical methods and shifted to neural networks for machine learning. '
'Popular techniques include the use of word embeddings to capture semantic '
'properties of words, and an increase in end-to-end learning of a '
'higher-level task (e.g., question answering) instead of relying on a '
'pipeline of separate intermediate tasks (e.g., part-of-speech tagging and '
'dependency parsing). In some areas, this shift has entailed substantial '
'changes in how NLP systems are designed, such that deep neural network-based '
'approaches may be viewed as a new paradigm distinct from statistical natural '
'language processing. For instance, the term neural machine translation (NMT) '
'emphasizes the fact that deep learning-based approaches to machine '
'translation directly learn sequence-to-sequence transformations, obviating '
'the need for intermediate steps such as word alignment and language modeling '
'that was used in statistical machine translation (SMT). Latest works tend to '

```
'use non-technical structure of a given task to build proper neural '
'network.[20]\n'
'The following is a list of some of the most commonly researched tasks in '
'natural language processing. Some of these tasks have direct real-world '
'applications, while others more commonly serve as subtasks that are used to '
'aid in solving larger tasks.\n'
'Though natural language processing tasks are closely intertwined, they can '
'be subdivided into categories for convenience. A coarse division is given '
'below.\n'
'Based on long-standing trends in the field, it is possible to extrapolate '
'future directions of NLP. As of 2020, three trends among the topics of the '
'long-standing series of CoNLL Shared Tasks can be observed:[36]\n'
'Most higher-level NLP applications involve aspects that emulate intelligent '
'behaviour and apparent comprehension of natural language. More broadly '
'speaking, the technical operationalization of increasingly advanced aspects '
'of cognitive behaviour represents one of the developmental trajectories of '
'NLP (see trends among CoNLL shared tasks above).\n'
'Cognition refers to "the mental action or process of acquiring knowledge and '
'understanding through thought, experience, and the senses."[37] Cognitive '
'science is the interdisciplinary, scientific study of the mind and its '
'processes.[38] Cognitive linguistics is an interdisciplinary branch of '
'linguistics, combining knowledge and research from both psychology and '
'linguistics.[39] Especially during the age of symbolic NLP, the area of '
'computational linguistics maintained strong ties with cognitive studies.\n'
'As an example, George Lakoff offers a methodology to build natural language '
'processing (NLP) algorithms through the perspective of cognitive science, '
'along with the findings of cognitive linguistics,[40] with two defining '
'aspects:\n'
'Ties with cognitive linguistics are part of the historical heritage of NLP, '
'but they have been less frequently addressed since the statistical turn '
'during the 1990s. Nevertheless, approaches to develop cognitive models '
'towards technically operationalizable frameworks have been pursued in the '
'context of various frameworks, e.g., of cognitive grammar,[42] functional '
'grammar,[43] construction grammar,[44] computational psycholinguistics and '
'cognitive neuroscience (e.g., ACT-R), however, with limited uptake in '
'mainstream NLP (as measured by presence on major conferences[45] of the '
'ACL). More recently, ideas of cognitive NLP have been revived as an approach '
'to achieve explainability, e.g., under the notion of "cognitive AI".[46] '
'Likewise, ideas of cognitive NLP are inherent to neural models multimodal '
'NLP (although rarely made explicit).[47]\n'
' Media related to Natural language processing at Wikimedia Commons\n')
```

1.2 - Tokenization:

In [33]:

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.tokenize import sent_tokenize
from nltk.tag import pos_tag
# Remove Punctuation
from nltk.tokenize import RegexpTokenizer
```

In [34]:

```
tokenizer = RegexpTokenizer(r'\w+')
tokens= tokenizer.tokenize(text)
print('Number of tokens without stop words = {}'.format(len(tokens)))
print(tokens)
```

```
Number of tokens without stop words = 1365
['Natural', 'language', 'processing', 'NLP', 'is', 'a', 'subfield', 'of', 'linguistics',
'computer', 'science', 'and', 'artificial', 'intelligence', 'concerned', 'with', 'the',
'interactions', 'between', 'computers', 'and', 'human', 'language', 'in', 'particular',
'how', 'to', 'program', 'computers', 'to', 'process', 'and', 'analyze', 'large', 'amount
s', 'of', 'natural', 'language', 'data', 'The', 'goal', 'is', 'a', 'computer', 'capabl
```

```
e', 'of', 'understanding', 'the', 'contents', 'of', 'documents', 'including', 'the', 'co
ntextual', 'nuances', 'of', 'the', 'language', 'within', 'them', 'The', 'technology', 'c
an', 'then', 'accurately', 'extract', 'information', 'and', 'insights', 'contained', 'i
n', 'the', 'documents', 'as', 'well', 'as', 'categorize', 'and', 'organize', 'the', 'doc
uments', 'themselves', 'Challenges', 'in', 'natural', 'language', 'processing', 'frequen
tly', 'involve', 'speech', 'recognition', 'natural', 'language', 'understanding', 'and',
'natural', 'language', 'generation', 'Natural', 'language', 'processing', 'has', 'its',
'roots', 'in', 'the', '1950s', 'Already', 'in', '1950', 'Alan', 'Turing', 'published',
'an', 'article', 'titled', 'Computing', 'Machinery', 'and', 'Intelligence', 'which', 'pr
oposed', 'what', 'is', 'now', 'called', 'the', 'Turing', 'test', 'as', 'a', 'criterion',
'of', 'intelligence', 'a', 'task', 'that', 'involves', 'the', 'automated', 'interpretati
on', 'and', 'generation', 'of', 'natural', 'language', 'but', 'at', 'the', 'time', 'no
t', 'articulated', 'as', 'a', 'problem', 'separate', 'from', 'artificial', 'intelligenc
e', 'The', 'premise', 'of', 'symbolic', 'NLP', 'is', 'well', 'summarized', 'by', 'John',
'Searle', 's', 'Chinese', 'room', 'experiment', 'Given', 'a', 'collection', 'of', 'rule
s', 'e', 'g', 'a', 'Chinese', 'phrasebook', 'with', 'questions', 'and', 'matching', 'ans
wers', 'the', 'computer', 'emulates', 'natural', 'language', 'understanding', 'or', 'oth
er', 'NLP', 'tasks', 'by', 'applying', 'those', 'rules', 'to', 'the', 'data', 'it', 'i
s', 'confronted', 'with', 'Up', 'to', 'the', '1980s', 'most', 'natural', 'language', 'pr
ocessing', 'systems', 'were', 'based', 'on', 'complex', 'sets', 'of', 'hand', 'written',
'rules', 'Starting', 'in', 'the', 'late', '1980s', 'however', 'there', 'was', 'a', 'revo
lution', 'in', 'natural', 'language', 'processing', 'with', 'the', 'introduction', 'of',
'machine', 'learning', 'algorithms', 'for', 'language', 'processing', 'This', 'was', 'du
e', 'to', 'both', 'the', 'steady', 'increase', 'in', 'computational', 'power', 'see', 'M
oore', 's', 'law', 'and', 'the', 'gradual', 'lessening', 'of', 'the', 'dominance', 'of',
'Chomskyan', 'theories', 'of', 'linguistics', 'e', 'g', 'transformational', 'grammar',
'whose', 'theoretical', 'underpinnings', 'discouraged', 'the', 'sort', 'of', 'corpus',
'linguistics', 'that', 'underlies', 'the', 'machine', 'learning', 'approach', 'to', 'lan
guage', 'processing', '6', 'In', 'the', '2010s', 'representation', 'learning', 'and', 'd
eep', 'neural', 'network', 'style', 'machine', 'learning', 'methods', 'became', 'widespr
ead', 'in', 'natural', 'language', 'processing', 'due', 'in', 'part', 'to', 'a', 'flurr
y', 'of', 'results', 'showing', 'that', 'such', 'techniques', '7', '8', 'can', 'achiev
e', 'state', 'of', 'the', 'art', 'results', 'in', 'many', 'natural', 'language', 'task
s', 'for', 'example', 'in', 'language', 'modeling', '9', 'parsing', '10', '11', 'and',
'many', 'others', 'This', 'is', 'increasingly', 'important', 'in', 'medicine', 'and', 'h
ealthcare', 'where', 'NLP', 'is', 'being', 'used', 'to', 'analyze', 'notes', 'and', 'tex
t', 'in', 'electronic', 'health', 'records', 'that', 'would', 'otherwise', 'be', 'inacce
ssible', 'for', 'study', 'when', 'seeking', 'to', 'improve', 'care', '12', 'In', 'the',
'early', 'days', 'many', 'language', 'processing', 'systems', 'were', 'designed', 'by',
'symbolic', 'methods', 'i', 'e', 'the', 'hand', 'coding', 'of', 'a', 'set', 'of', 'rule
s', 'coupled', 'with', 'a', 'dictionary', 'lookup', '13', '14', 'such', 'as', 'by', 'wri
ting', 'grammars', 'or', 'devising', 'heuristic', 'rules', 'for', 'stemming', 'More', 'r
ecent', 'systems', 'based', 'on', 'machine', 'learning', 'algorithms', 'have', 'many',
'advantages', 'over', 'hand', 'produced', 'rules', 'Despite', 'the', 'popularity', 'of',
'machine', 'learning', 'in', 'NLP', 'research', 'symbolic', 'methods', 'are', 'still',
'2020', 'commonly', 'used', 'Since', 'the', 'so', 'called', 'statistical', 'revolution',
'15', '16', 'in', 'the', 'late', '1980s', 'and', 'mid', '1990s', 'much', 'natural', 'lan
guage', 'processing', 'research', 'has', 'relied', 'heavily', 'on', 'machine', 'learnin
g', 'The', 'machine', 'learning', 'paradigm', 'calls', 'instead', 'for', 'using', 'stati
stical', 'inference', 'to', 'automatically', 'learn', 'such', 'rules', 'through', 'the',
'analysis', 'of', 'large', 'corpora', 'the', 'plural', 'form', 'of', 'corpus', 'is',
'a', 'set', 'of', 'documents', 'possibly', 'with', 'human', 'or', 'computer', 'annotatio
ns', 'of', 'typical', 'real', 'world', 'examples', 'Many', 'different', 'classes', 'of',
'machine', 'learning', 'algorithms', 'have', 'been', 'applied', 'to', 'natural', 'langua
ge', 'processing', 'tasks', 'These', 'algorithms', 'take', 'as', 'input', 'a', 'large',
'set', 'of', 'features', 'that', 'are', 'generated', 'from', 'the', 'input', 'data', 'In
creasingly', 'however', 'research', 'has', 'focused', 'on', 'statistical', 'models', 'wh
ich', 'make', 'soft', 'probabilistic', 'decisions', 'based', 'on', 'attaching', 'real',
'valued', 'weights', 'to', 'each', 'input', 'feature', 'complex', 'valued', 'embedding
s', '17', 'and', 'neural', 'networks', 'in', 'general', 'have', 'also', 'been', 'propose
d', 'for', 'e', 'g', 'speech', '18', 'Such', 'models', 'have', 'the', 'advantage', 'tha
t', 'they', 'can', 'express', 'the', 'relative', 'certainty', 'of', 'many', 'different',
'possible', 'answers', 'rather', 'than', 'only', 'one', 'producing', 'more', 'reliable',
'results', 'when', 'such', 'a', 'model', 'is', 'included', 'as', 'a', 'component', 'of',
'a', 'larger', 'system', 'Some', 'of', 'the', 'earliest', 'used', 'machine', 'learning',
```

'algorithms', 'such', 'as', 'decision', 'trees', 'produced', 'systems', 'of', 'hard', 'i
f', 'then', 'rules', 'similar', 'to', 'existing', 'hand', 'written', 'rules', 'However',
'part', 'of', 'speech', 'tagging', 'introduced', 'the', 'use', 'of', 'hidden', 'Markov',
'models', 'to', 'natural', 'language', 'processing', 'and', 'increasingly', 'research',
'has', 'focused', 'on', 'statistical', 'models', 'which', 'make', 'soft', 'probabilisti
c', 'decisions', 'based', 'on', 'attaching', 'real', 'valued', 'weights', 'to', 'the',
'features', 'making', 'up', 'the', 'input', 'data', 'The', 'cache', 'language', 'model
s', 'upon', 'which', 'many', 'speech', 'recognition', 'systems', 'now', 'rely', 'are',
'examples', 'of', 'such', 'statistical', 'models', 'Such', 'models', 'are', 'generally',
'more', 'robust', 'when', 'given', 'unfamiliar', 'input', 'especially', 'input', 'that',
'contains', 'errors', 'as', 'is', 'very', 'common', 'for', 'real', 'world', 'data', 'an
d', 'produce', 'more', 'reliable', 'results', 'when', 'integrated', 'into', 'a', 'large
r', 'system', 'comprising', 'multiple', 'subtasks', 'Since', 'the', 'neural', 'turn', 's
tatistical', 'methods', 'in', 'NLP', 'research', 'have', 'been', 'largely', 'replaced',
'by', 'neural', 'networks', 'However', 'they', 'continue', 'to', 'be', 'relevant', 'fo
r', 'contexts', 'in', 'which', 'statistical', 'interpretability', 'and', 'transparency',
'is', 'required', 'A', 'major', 'drawback', 'of', 'statistical', 'methods', 'is', 'tha
t', 'they', 'require', 'elaborate', 'feature', 'engineering', 'Since', '2015', '19', 'th
e', 'field', 'has', 'thus', 'largely', 'abandoned', 'statistical', 'methods', 'and', 'sh
ifted', 'to', 'neural', 'networks', 'for', 'machine', 'learning', 'Popular', 'technique
s', 'include', 'the', 'use', 'of', 'word', 'embeddings', 'to', 'capture', 'semantic', 'p
roperties', 'of', 'words', 'and', 'an', 'increase', 'in', 'end', 'to', 'end', 'learnin
g', 'of', 'a', 'higher', 'level', 'task', 'e', 'g', 'question', 'answering', 'instead',
'of', 'relying', 'on', 'a', 'pipeline', 'of', 'separate', 'intermediate', 'tasks', 'e',
'g', 'part', 'of', 'speech', 'tagging', 'and', 'dependency', 'parsing', 'In', 'some', 'a
reas', 'this', 'shift', 'has', 'entailed', 'substantial', 'changes', 'in', 'how', 'NLP',
'systems', 'are', 'designed', 'such', 'that', 'deep', 'neural', 'network', 'based', 'app
roaches', 'may', 'be', 'viewed', 'as', 'a', 'new', 'paradigm', 'distinct', 'from', 'stat
istical', 'natural', 'language', 'processing', 'For', 'instance', 'the', 'term', 'neura
l', 'machine', 'translation', 'NMT', 'emphasizes', 'the', 'fact', 'that', 'deep', 'learn
ing', 'based', 'approaches', 'to', 'machine', 'translation', 'directly', 'learn', 'seque
nce', 'to', 'sequence', 'transformations', 'obviating', 'the', 'need', 'for', 'intermedi
ate', 'steps', 'such', 'as', 'word', 'alignment', 'and', 'language', 'modeling', 'that',
'was', 'used', 'in', 'statistical', 'machine', 'translation', 'SMT', 'Latest', 'works',
'tend', 'to', 'use', 'non', 'technical', 'structure', 'of', 'a', 'given', 'task', 'to',
'build', 'proper', 'neural', 'network', '20', 'The', 'following', 'is', 'a', 'list', 'o
f', 'some', 'of', 'the', 'most', 'commonly', 'researched', 'tasks', 'in', 'natural', 'la
nguage', 'processing', 'Some', 'of', 'these', 'tasks', 'have', 'direct', 'real', 'worl
d', 'applications', 'while', 'others', 'more', 'commonly', 'serve', 'as', 'subtasks', 't
hat', 'are', 'used', 'to', 'aid', 'in', 'solving', 'larger', 'tasks', 'Though', 'natura
l', 'language', 'processing', 'tasks', 'are', 'closely', 'intertwined', 'they', 'can',
'be', 'subdivided', 'into', 'categories', 'for', 'convenience', 'A', 'coarse', 'divisio
n', 'is', 'given', 'below', 'Based', 'on', 'long', 'standing', 'trends', 'in', 'the', 'f
ield', 'it', 'is', 'possible', 'to', 'extrapolate', 'future', 'directions', 'of', 'NLP',
'As', 'of', '2020', 'three', 'trends', 'among', 'the', 'topics', 'of', 'the', 'long', 's
tanding', 'series', 'of', 'CoNLL', 'Shared', 'Tasks', 'can', 'be', 'observed', '36', 'Mo
st', 'higher', 'level', 'NLP', 'applications', 'involve', 'aspects', 'that', 'emulate',
'intelligent', 'behaviour', 'and', 'apparent', 'comprehension', 'of', 'natural', 'langua
ge', 'More', 'broadly', 'speaking', 'the', 'technical', 'operationalization', 'of', 'inc
reasingly', 'advanced', 'aspects', 'of', 'cognitive', 'behaviour', 'represents', 'one',
'of', 'the', 'developmental', 'trajectories', 'of', 'NLP', 'see', 'trends', 'among', 'Co
NLL', 'shared', 'tasks', 'above', 'Cognition', 'refers', 'to', 'the', 'mental', 'actio
n', 'or', 'process', 'of', 'acquiring', 'knowledge', 'and', 'understanding', 'through',
'thought', 'experience', 'and', 'the', 'senses', '37', 'Cognitive', 'science', 'is', 'th
e', 'interdisciplinary', 'scientific', 'study', 'of', 'the', 'mind', 'and', 'its', 'proc
esses', '38', 'Cognitive', 'linguistics', 'is', 'an', 'interdisciplinary', 'branch', 'o
f', 'linguistics', 'combining', 'knowledge', 'and', 'research', 'from', 'both', 'psychol
ogy', 'and', 'linguistics', '39', 'Especially', 'during', 'the', 'age', 'of', 'symboli
c', 'NLP', 'the', 'area', 'of', 'computational', 'linguistics', 'maintained', 'strong',
'ties', 'with', 'cognitive', 'studies', 'As', 'an', 'example', 'George', 'Lakoff', 'offe
rs', 'a', 'methodology', 'to', 'build', 'natural', 'language', 'processing', 'NLP', 'alg
orithms', 'through', 'the', 'perspective', 'of', 'cognitive', 'science', 'along', 'wit
h', 'the', 'findings', 'of', 'cognitive', 'linguistics', '40', 'with', 'two', 'definin
g', 'aspects', 'Ties', 'with', 'cognitive', 'linguistics', 'are', 'part', 'of', 'the',
'historical', 'heritage', 'of', 'NLP', 'but', 'they', 'have', 'been', 'less', 'frequentl

```
y', 'addressed', 'since', 'the', 'statistical', 'turn', 'during', 'the', '1990s', 'Never
theless', 'approaches', 'to', 'develop', 'cognitive', 'models', 'towards', 'technicall
y', 'operationalizable', 'frameworks', 'have', 'been', 'pursued', 'in', 'the', 'contex
t', 'of', 'various', 'frameworks', 'e', 'g', 'of', 'cognitive', 'grammar', '42', 'functi
onal', 'grammar', '43', 'construction', 'grammar', '44', 'computational', 'psycholinguis
tics', 'and', 'cognitive', 'neuroscience', 'e', 'g', 'ACT', 'R', 'however', 'with', 'lim
ited', 'uptake', 'in', 'mainstream', 'NLP', 'as', 'measured', 'by', 'presence', 'on', 'm
ajor', 'conferences', '45', 'of', 'the', 'ACL', 'More', 'recently', 'ideas', 'of', 'cogn
itive', 'NLP', 'have', 'been', 'revived', 'as', 'an', 'approach', 'to', 'achieve', 'expl
ainability', 'e', 'g', 'under', 'the', 'notion', 'of', 'cognitive', 'AI', '46', 'Likewis
e', 'ideas', 'of', 'cognitive', 'NLP', 'are', 'inherent', 'to', 'neural', 'models', 'mul
timodal', 'NLP', 'although', 'rarely', 'made', 'explicit', '47', 'Media', 'related', 't
o', 'Natural', 'language', 'processing', 'at', 'Wikimedia', 'Commons']
```

Omission of Stopwords:

In [35]:

```
from nltk.corpus import stopwords

stop_words=stopwords.words('english')
def nostop(txt):
    clean= [word for word in txt if word.lower() not in stop_words]
    return clean
c
print('Number of tokens without stop words = {}'.format(len(txt_nostop)))
print(txt_nostop)
```

```
Number of tokens without stop words = 878
['Natural', 'language', 'processing', 'NLP', 'subfield', 'linguistics', 'computer', 'sci
ence', 'artificial', 'intelligence', 'concerned', 'interactions', 'computers', 'human',
'language', 'particular', 'program', 'computers', 'process', 'analyze', 'large', 'amount
s', 'natural', 'language', 'data', 'goal', 'computer', 'capable', 'understanding', 'cont
ents', 'documents', 'including', 'contextual', 'nuances', 'language', 'within', 'technol
ogy', 'accurately', 'extract', 'information', 'insights', 'contained', 'documents', 'wel
l', 'categorize', 'organize', 'documents', 'Challenges', 'natural', 'language', 'process
ing', 'frequently', 'involve', 'speech', 'recognition', 'natural', 'language', 'understa
nding', 'natural', 'language', 'generation', 'Natural', 'language', 'processing', 'root
s', '1950s', 'Already', '1950', 'Alan', 'Turing', 'published', 'article', 'titled', 'Com
puting', 'Machinery', 'Intelligence', 'proposed', 'called', 'Turing', 'test', 'criterio
n', 'intelligence', 'task', 'involves', 'automated', 'interpretation', 'generation', 'na
tural', 'language', 'time', 'articulated', 'problem', 'separate', 'artificial', 'intelli
gence', 'premise', 'symbolic', 'NLP', 'well', 'summarized', 'John', 'Searle', 'Chinese',
'room', 'experiment', 'Given', 'collection', 'rules', 'e', 'g', 'Chinese', 'phrasebook',
'questions', 'matching', 'answers', 'computer', 'emulates', 'natural', 'language', 'unde
rstanding', 'NLP', 'tasks', 'applying', 'rules', 'data', 'confronted', '1980s', 'natura
l', 'language', 'processing', 'systems', 'based', 'complex', 'sets', 'hand', 'written',
'rules', 'Starting', 'late', '1980s', 'however', 'revolution', 'natural', 'language', 'p
rocessing', 'introduction', 'machine', 'learning', 'algorithms', 'language', 'processin
g', 'due', 'steady', 'increase', 'computational', 'power', 'see', 'Moore', 'law', 'gradu
al', 'lessening', 'dominance', 'Chomskyan', 'theories', 'linguistics', 'e', 'g', 'transf
ormational', 'grammar', 'whose', 'theoretical', 'underpinnings', 'discouraged', 'sort',
'corpus', 'linguistics', 'underlies', 'machine', 'learning', 'approach', 'language', 'pr
ocessing', '6', '2010s', 'representation', 'learning', 'deep', 'neural', 'network', 'sty
le', 'machine', 'learning', 'methods', 'became', 'widespread', 'natural', 'language', 'p
rocessing', 'due', 'part', 'flurry', 'results', 'showing', 'techniques', '7', '8', 'achi
eve', 'state', 'art', 'results', 'many', 'natural', 'language', 'tasks', 'example', 'lan
guage', 'modeling', '9', 'parsing', '10', '11', 'many', 'others', 'increasingly', 'impor
tant', 'medicine', 'healthcare', 'NLP', 'used', 'analyze', 'notes', 'text', 'electroni
c', 'health', 'records', 'would', 'otherwise', 'inaccessible', 'study', 'seeking', 'impr
ove', 'care', '12', 'early', 'days', 'many', 'language', 'processing', 'systems', 'desig
ned', 'symbolic', 'methods', 'e', 'hand', 'coding', 'set', 'rules', 'coupled', 'dictiona
ry', 'lookup', '13', '14', 'writing', 'grammars', 'devising', 'heuristic', 'rules', 'ste
mming', 'recent', 'systems', 'based', 'machine', 'learning', 'algorithms', 'many', 'adva
ntages', 'hand', 'produced', 'rules', 'Despite', 'popularity', 'machine', 'learning', 'N
```

LP', 'research', 'symbolic', 'methods', 'still', '2020', 'commonly', 'used', 'Since', 'c
alled', 'statistical', 'revolution', '15', '16', 'late', '1980s', 'mid', '1990s', 'muc
h', 'natural', 'language', 'processing', 'research', 'relied', 'heavily', 'machine', 'le
arning', 'machine', 'learning', 'paradigm', 'calls', 'instead', 'using', 'statistical',
'inference', 'automatically', 'learn', 'rules', 'analysis', 'large', 'corpora', 'plura
l', 'form', 'corpus', 'set', 'documents', 'possibly', 'human', 'computer', 'annotation
s', 'typical', 'real', 'world', 'examples', 'Many', 'different', 'classes', 'machine',
'learning', 'algorithms', 'applied', 'natural', 'language', 'processing', 'tasks', 'algo
rithms', 'take', 'input', 'large', 'set', 'features', 'generated', 'input', 'data', 'Inc
reasingly', 'however', 'research', 'focused', 'statistical', 'models', 'make', 'soft',
'probabilistic', 'decisions', 'based', 'attaching', 'real', 'valued', 'weights', 'inpu
t', 'feature', 'complex', 'valued', 'embeddings', '17', 'neural', 'networks', 'general',
'also', 'proposed', 'e', 'g', 'speech', '18', 'models', 'advantage', 'express', 'relativ
e', 'certainty', 'many', 'different', 'possible', 'answers', 'rather', 'one', 'producin
g', 'reliable', 'results', 'model', 'included', 'component', 'larger', 'system', 'earlie
st', 'used', 'machine', 'learning', 'algorithms', 'decision', 'trees', 'produced', 'syst
ems', 'hard', 'rules', 'similar', 'existing', 'hand', 'written', 'rules', 'However', 'pa
rt', 'speech', 'tagging', 'introduced', 'use', 'hidden', 'Markov', 'models', 'natural',
'language', 'processing', 'increasingly', 'research', 'focused', 'statistical', 'model
s', 'make', 'soft', 'probabilistic', 'decisions', 'based', 'attaching', 'real', 'value
d', 'weights', 'features', 'making', 'input', 'data', 'cache', 'language', 'models', 'up
on', 'many', 'speech', 'recognition', 'systems', 'rely', 'examples', 'statistical', 'mod
els', 'models', 'generally', 'robust', 'given', 'unfamiliar', 'input', 'especially', 'in
put', 'contains', 'errors', 'common', 'real', 'world', 'data', 'produce', 'reliable', 'r
esults', 'integrated', 'larger', 'system', 'comprising', 'multiple', 'subtasks', 'Sinc
e', 'neural', 'turn', 'statistical', 'methods', 'NLP', 'research', 'largely', 'replace
d', 'neural', 'networks', 'However', 'continue', 'relevant', 'contexts', 'statistical',
'interpretability', 'transparency', 'required', 'major', 'drawback', 'statistical', 'met
hods', 'require', 'elaborate', 'feature', 'engineering', 'Since', '2015', '19', 'field',
'thus', 'largely', 'abandoned', 'statistical', 'methods', 'shifted', 'neural', 'network
s', 'machine', 'learning', 'Popular', 'techniques', 'include', 'use', 'word', 'embedding
s', 'capture', 'semantic', 'properties', 'words', 'increase', 'end', 'end', 'learning',
'higher', 'level', 'task', 'e', 'g', 'question', 'answering', 'instead', 'relying', 'pip
eline', 'separate', 'intermediate', 'tasks', 'e', 'g', 'part', 'speech', 'tagging', 'dep
endency', 'parsing', 'areas', 'shift', 'entailed', 'substantial', 'changes', 'NLP', 'sys
tems', 'designed', 'deep', 'neural', 'network', 'based', 'approaches', 'may', 'viewed',
'new', 'paradigm', 'distinct', 'statistical', 'natural', 'language', 'processing', 'inst
ance', 'term', 'neural', 'machine', 'translation', 'NMT', 'emphasizes', 'fact', 'deep',
'learning', 'based', 'approaches', 'machine', 'translation', 'directly', 'learn', 'seque
nce', 'sequence', 'transformations', 'obviating', 'need', 'intermediate', 'steps', 'wor
d', 'alignment', 'language', 'modeling', 'used', 'statistical', 'machine', 'translatio
n', 'SMT', 'Latest', 'works', 'tend', 'use', 'non', 'technical', 'structure', 'given',
'task', 'build', 'proper', 'neural', 'network', '20', 'following', 'list', 'commonly',
'researched', 'tasks', 'natural', 'language', 'processing', 'tasks', 'direct', 'real',
'world', 'applications', 'others', 'commonly', 'serve', 'subtasks', 'used', 'aid', 'solv
ing', 'larger', 'tasks', 'Though', 'natural', 'language', 'processing', 'tasks', 'closel
y', 'intertwined', 'subdivided', 'categories', 'convenience', 'coarse', 'division', 'giv
en', 'Based', 'long', 'standing', 'trends', 'field', 'possible', 'extrapolate', 'futur
e', 'directions', 'NLP', '2020', 'three', 'trends', 'among', 'topics', 'long', 'standin
g', 'series', 'CoNLL', 'Shared', 'Tasks', 'observed', '36', 'higher', 'level', 'NLP', 'a
pplications', 'involve', 'aspects', 'emulate', 'intelligent', 'behaviour', 'apparent',
'comprehension', 'natural', 'language', 'broadly', 'speaking', 'technical', 'operational
ization', 'increasingly', 'advanced', 'aspects', 'cognitive', 'behaviour', 'represents',
'one', 'developmental', 'trajectories', 'NLP', 'see', 'trends', 'among', 'CoNLL', 'share
d', 'tasks', 'Cognition', 'refers', 'mental', 'action', 'process', 'acquiring', 'knowled
ge', 'understanding', 'thought', 'experience', 'senses', '37', 'Cognitive', 'science',
'interdisciplinary', 'scientific', 'study', 'mind', 'processes', '38', 'Cognitive', 'lin
guistics', 'interdisciplinary', 'branch', 'linguistics', 'combining', 'knowledge', 'rese
arch', 'psychology', 'linguistics', '39', 'Especially', 'age', 'symbolic', 'NLP', 'are
a', 'computational', 'linguistics', 'maintained', 'strong', 'ties', 'cognitive', 'studie
s', 'example', 'George', 'Lakoff', 'offers', 'methodology', 'build', 'natural', 'languag
e', 'processing', 'NLP', 'algorithms', 'perspective', 'cognitive', 'science', 'along',
'findings', 'cognitive', 'linguistics', '40', 'two', 'defining', 'aspects', 'Ties', 'cog
nitive', 'linguistics', 'part', 'historical', 'heritage', 'NLP', 'less', 'frequently',
'addressed', 'since', 'statistical', 'turn', '1990s', 'Nevertheless', 'approaches', 'dev

```
elop', 'cognitive', 'models', 'towards', 'technically', 'operationalizable', 'framework
s', 'pursued', 'context', 'various', 'frameworks', 'e', 'g', 'cognitive', 'grammar', '4
2', 'functional', 'grammar', '43', 'construction', 'grammar', '44', 'computational', 'ps
ycholinguistics', 'cognitive', 'neuroscience', 'e', 'g', 'ACT', 'R', 'however', 'limite
d', 'uptake', 'mainstream', 'NLP', 'measured', 'presence', 'major', 'conferences', '45',
'ACL', 'recently', 'ideas', 'cognitive', 'NLP', 'revived', 'approach', 'achieve', 'expla
inability', 'e', 'g', 'notion', 'cognitive', 'AI', '46', 'Likewise', 'ideas', 'cognitiv
e', 'NLP', 'inherent', 'neural', 'models', 'multimodal', 'NLP', 'although', 'rarely', 'm
ade', 'explicit', '47', 'Media', 'related', 'Natural', 'language', 'processing', 'Wikime
dia', 'Commons']
```

1.4 - Word Frequencies:

In [36]:

```
from nltk.stem.wordnet import WordNetLemmatizer

lmtz = nltk.WordNetLemmatizer()

lemmatized_text=[lmtz.lemmatize(w) for w in txt_nostop]

#print(lemmatized_text)

freqdist = nltk.FreqDist(lemmatized_text)
freqdist.most_common(15)
#print(max(freqdist.values()))
```

Out[36]:

```
[('language', 29),
 ('natural', 18),
 ('processing', 17),
 ('NLP', 17),
 ('machine', 13),
 ('learning', 13),
 ('task', 12),
 ('statistical', 12),
 ('cognitive', 11),
 ('model', 10),
 ('linguistics', 9),
 ('rule', 9),
 ('e', 9),
 ('neural', 9),
 ('g', 8)]
```

In [37]:

```
#weighted_freq=freqdist.values()/max(freqdist.values())
#print(weighted_freq)
fd_weighted={}
for i in freqdist:
    fd_weighted[i]=round(freqdist[i]/max(freqdist.values()),2)
print(fd_weighted)
```

```
{'language': 1.0, 'natural': 0.62, 'processing': 0.59, 'NLP': 0.59, 'machine': 0.45, 'le
arning': 0.45, 'task': 0.41, 'statistical': 0.41, 'cognitive': 0.38, 'model': 0.34, 'lin
guistics': 0.31, 'rule': 0.31, 'e': 0.31, 'neural': 0.31, 'g': 0.28, 'system': 0.28, 'co
mputer': 0.21, 'based': 0.21, 'algorithm': 0.21, 'network': 0.21, 'method': 0.21, 'man
y': 0.21, 'research': 0.21, 'input': 0.21, 'data': 0.17, 'speech': 0.17, 'grammar': 0.1
7, 'approach': 0.17, 'used': 0.17, 'real': 0.17, 'understanding': 0.14, 'document': 0.1
4, 'symbolic': 0.14, 'set': 0.14, 'hand': 0.14, 'part': 0.14, 'result': 0.14, 'example':
0.14, 'feature': 0.14, 'Natural': 0.1, 'science': 0.1, 'intelligence': 0.1, 'process':
0.1, 'large': 0.1, '1980s': 0.1, 'however': 0.1, 'computational': 0.1, 'corpus': 0.1, 'd
eep': 0.1, 'increasingly': 0.1, 'study': 0.1, 'commonly': 0.1, 'Since': 0.1, 'world': 0.
1, 'decision': 0.1, 'valued': 0.1, 'larger': 0.1, 'use': 0.1, 'given': 0.1, 'word': 0.1,
'translation': 0.1, 'trend': 0.1, 'aspect': 0.1, 'artificial': 0.07, 'human': 0.07, 'ana
```

lyze': 0.07, 'well': 0.07, 'frequently': 0.07, 'involve': 0.07, 'recognition': 0.07, 'ge
neration': 0.07, 'Turing': 0.07, 'proposed': 0.07, 'called': 0.07, 'separate': 0.07, 'Ch
inese': 0.07, 'question': 0.07, 'answer': 0.07, 'complex': 0.07, 'written': 0.07, 'lat
e': 0.07, 'revolution': 0.07, 'due': 0.07, 'increase': 0.07, 'see': 0.07, 'technique':
0.07, 'achieve': 0.07, 'modeling': 0.07, 'parsing': 0.07, 'others': 0.07, 'designed': 0.
07, 'advantage': 0.07, 'produced': 0.07, '2020': 0.07, '1990s': 0.07, 'paradigm': 0.07,
'instead': 0.07, 'learn': 0.07, 'different': 0.07, 'focused': 0.07, 'make': 0.07, 'sof
t': 0.07, 'probabilistic': 0.07, 'attaching': 0.07, 'weight': 0.07, 'embeddings': 0.07,
'possible': 0.07, 'one': 0.07, 'reliable': 0.07, 'However': 0.07, 'tagging': 0.07, 'subt
asks': 0.07, 'turn': 0.07, 'largely': 0.07, 'context': 0.07, 'major': 0.07, 'field': 0.0
7, 'end': 0.07, 'higher': 0.07, 'level': 0.07, 'intermediate': 0.07, 'area': 0.07, 'sequ
ence': 0.07, 'technical': 0.07, 'build': 0.07, 'application': 0.07, 'long': 0.07, 'stand
ing': 0.07, 'among': 0.07, 'CoNLL': 0.07, 'behaviour': 0.07, 'knowledge': 0.07, 'Cogniti
ve': 0.07, 'interdisciplinary': 0.07, 'framework': 0.07, 'idea': 0.07, 'subfield': 0.03,
'concerned': 0.03, 'interaction': 0.03, 'particular': 0.03, 'program': 0.03, 'amount':
0.03, 'goal': 0.03, 'capable': 0.03, 'content': 0.03, 'including': 0.03, 'contextual':
0.03, 'nuance': 0.03, 'within': 0.03, 'technology': 0.03, 'accurately': 0.03, 'extract':
0.03, 'information': 0.03, 'insight': 0.03, 'contained': 0.03, 'categorize': 0.03, 'orga
nize': 0.03, 'Challenges': 0.03, 'root': 0.03, '1950s': 0.03, 'Already': 0.03, '1950':
0.03, 'Alan': 0.03, 'published': 0.03, 'article': 0.03, 'titled': 0.03, 'Computing': 0.0
3, 'Machinery': 0.03, 'Intelligence': 0.03, 'test': 0.03, 'criterion': 0.03, 'involves':
0.03, 'automated': 0.03, 'interpretation': 0.03, 'time': 0.03, 'articulated': 0.03, 'pro
blem': 0.03, 'premise': 0.03, 'summarized': 0.03, 'John': 0.03, 'Searle': 0.03, 'room':
0.03, 'experiment': 0.03, 'Given': 0.03, 'collection': 0.03, 'phrasebook': 0.03, 'matchi
ng': 0.03, 'emulates': 0.03, 'applying': 0.03, 'confronted': 0.03, 'Starting': 0.03, 'in
troduction': 0.03, 'steady': 0.03, 'power': 0.03, 'Moore': 0.03, 'law': 0.03, 'gradual':
0.03, 'lessening': 0.03, 'dominance': 0.03, 'Chomskyan': 0.03, 'theory': 0.03, 'transfor
mational': 0.03, 'whose': 0.03, 'theoretical': 0.03, 'underpinnings': 0.03, 'discourage
d': 0.03, 'sort': 0.03, 'underlies': 0.03, '6': 0.03, '2010s': 0.03, 'representation':
0.03, 'style': 0.03, 'became': 0.03, 'widespread': 0.03, 'flurry': 0.03, 'showing': 0.0
3, '7': 0.03, '8': 0.03, 'state': 0.03, 'art': 0.03, '9': 0.03, '10': 0.03, '11': 0.03,
'important': 0.03, 'medicine': 0.03, 'healthcare': 0.03, 'note': 0.03, 'text': 0.03, 'el
ectronic': 0.03, 'health': 0.03, 'record': 0.03, 'would': 0.03, 'otherwise': 0.03, 'inac
cessible': 0.03, 'seeking': 0.03, 'improve': 0.03, 'care': 0.03, '12': 0.03, 'early': 0.
03, 'day': 0.03, 'coding': 0.03, 'coupled': 0.03, 'dictionary': 0.03, 'lookup': 0.03, '1
3': 0.03, '14': 0.03, 'writing': 0.03, 'devising': 0.03, 'heuristic': 0.03, 'stemming':
0.03, 'recent': 0.03, 'Despite': 0.03, 'popularity': 0.03, 'still': 0.03, '15': 0.03, '1
6': 0.03, 'mid': 0.03, 'much': 0.03, 'relied': 0.03, 'heavily': 0.03, 'call': 0.03, 'usi
ng': 0.03, 'inference': 0.03, 'automatically': 0.03, 'analysis': 0.03, 'plural': 0.03,
'form': 0.03, 'possibly': 0.03, 'annotation': 0.03, 'typical': 0.03, 'Many': 0.03, 'clas
s': 0.03, 'applied': 0.03, 'take': 0.03, 'generated': 0.03, 'Increasingly': 0.03, '17':
0.03, 'general': 0.03, 'also': 0.03, '18': 0.03, 'express': 0.03, 'relative': 0.03, 'cer
tainty': 0.03, 'rather': 0.03, 'producing': 0.03, 'included': 0.03, 'component': 0.03,
'earliest': 0.03, 'tree': 0.03, 'hard': 0.03, 'similar': 0.03, 'existing': 0.03, 'introd
uced': 0.03, 'hidden': 0.03, 'Markov': 0.03, 'making': 0.03, 'cache': 0.03, 'upon': 0.0
3, 'rely': 0.03, 'generally': 0.03, 'robust': 0.03, 'unfamiliar': 0.03, 'especially': 0.
03, 'contains': 0.03, 'error': 0.03, 'common': 0.03, 'produce': 0.03, 'integrated': 0.0
3, 'comprising': 0.03, 'multiple': 0.03, 'replaced': 0.03, 'continue': 0.03, 'relevant':
0.03, 'interpretability': 0.03, 'transparency': 0.03, 'required': 0.03, 'drawback': 0.0
3, 'require': 0.03, 'elaborate': 0.03, 'engineering': 0.03, '2015': 0.03, '19': 0.03, 't
hus': 0.03, 'abandoned': 0.03, 'shifted': 0.03, 'Popular': 0.03, 'include': 0.03, 'captu
re': 0.03, 'semantic': 0.03, 'property': 0.03, 'answering': 0.03, 'relying': 0.03, 'pipe
line': 0.03, 'dependency': 0.03, 'shift': 0.03, 'entailed': 0.03, 'substantial': 0.03,
'change': 0.03, 'may': 0.03, 'viewed': 0.03, 'new': 0.03, 'distinct': 0.03, 'instance':
0.03, 'term': 0.03, 'NMT': 0.03, 'emphasizes': 0.03, 'fact': 0.03, 'directly': 0.03, 'tr
ansformation': 0.03, 'obviating': 0.03, 'need': 0.03, 'step': 0.03, 'alignment': 0.03,
'SMT': 0.03, 'Latest': 0.03, 'work': 0.03, 'tend': 0.03, 'non': 0.03, 'structure': 0.03,
'proper': 0.03, '20': 0.03, 'following': 0.03, 'list': 0.03, 'researched': 0.03, 'direc
t': 0.03, 'serve': 0.03, 'aid': 0.03, 'solving': 0.03, 'Though': 0.03, 'closely': 0.03,
'intertwined': 0.03, 'subdivided': 0.03, 'category': 0.03, 'convenience': 0.03, 'coars
e': 0.03, 'division': 0.03, 'Based': 0.03, 'extrapolate': 0.03, 'future': 0.03, 'directi
on': 0.03, 'three': 0.03, 'topic': 0.03, 'series': 0.03, 'Shared': 0.03, 'Tasks': 0.03,
'observed': 0.03, '36': 0.03, 'emulate': 0.03, 'intelligent': 0.03, 'apparent': 0.03, 'c
omprehension': 0.03, 'broadly': 0.03, 'speaking': 0.03, 'operationalization': 0.03, 'adv
anced': 0.03, 'represents': 0.03, 'developmental': 0.03, 'trajectory': 0.03, 'shared':

```
0.03, 'Cognition': 0.03, 'refers': 0.03, 'mental': 0.03, 'action': 0.03, 'acquiring': 0.
03, 'thought': 0.03, 'experience': 0.03, 'sens': 0.03, '37': 0.03, 'scientific': 0.03,
'mind': 0.03, '38': 0.03, 'branch': 0.03, 'combining': 0.03, 'psychology': 0.03, '39':
0.03, 'Especially': 0.03, 'age': 0.03, 'maintained': 0.03, 'strong': 0.03, 'tie': 0.03,
'George': 0.03, 'Lakoff': 0.03, 'offer': 0.03, 'methodology': 0.03, 'perspective': 0.03,
'along': 0.03, 'finding': 0.03, '40': 0.03, 'two': 0.03, 'defining': 0.03, 'Ties': 0.03,
'historical': 0.03, 'heritage': 0.03, 'le': 0.03, 'addressed': 0.03, 'since': 0.03, 'Nev
ertheless': 0.03, 'develop': 0.03, 'towards': 0.03, 'technically': 0.03, 'operationaliza
ble': 0.03, 'pursued': 0.03, 'various': 0.03, '42': 0.03, 'functional': 0.03, '43': 0.0
3, 'construction': 0.03, '44': 0.03, 'psycholinguistics': 0.03, 'neuroscience': 0.03, 'A
CT': 0.03, 'R': 0.03, 'limited': 0.03, 'uptake': 0.03, 'mainstream': 0.03, 'measured':
0.03, 'presence': 0.03, 'conference': 0.03, '45': 0.03, 'ACL': 0.03, 'recently': 0.03,
'revived': 0.03, 'explainability': 0.03, 'notion': 0.03, 'AI': 0.03, '46': 0.03, 'Likewi
se': 0.03, 'inherent': 0.03, 'multimodal': 0.03, 'although': 0.03, 'rarely': 0.03, 'mad
e': 0.03, 'explicit': 0.03, '47': 0.03, 'Media': 0.03, 'related': 0.03, 'Wikimedia': 0.0
3, 'Commons': 0.03}
```

1.5 - Sentence Scroing:

In [38]:

```
# Sentence tokenization and cleaning
sentences=sent_tokenize(text)

sentences_clean=[]

for i in sentences:
    s=tokenizer.tokenize(i)
    s=nostop(s)
    sentences_clean.append(' '.join(s))


sentence_scores={}

for sentence in sentences_clean:
    for word in word_tokenize(sentence.lower()):
        # Avoid lengthy/wordy sentences
        if word in fd_weighted and len(sentence.split(' '))<20:
            if sentence not in sentence_scores:
                sentence_scores[sentence] = fd_weighted[word]
            else:
                sentence_scores[sentence] += fd_weighted[word]

print(sentence_scores)
```

```
{'goal computer capable understanding contents documents including contextual nuances la
nguage within': 1.5000000000000002, 'technology accurately extract information insights
contained documents well categorize organize documents': 0.28, 'Challenges natural langu
age processing frequently involve speech recognition natural language understanding natu
ral language generation': 6.039999999999999, 'Natural language processing roots 1950s':
2.2399999999999998, '1980s natural language processing systems based complex sets hand w
ritten rules': 2.8, 'Starting late 1980s however revolution natural language processing
introduction machine learning algorithms language processing': 5.07, 'due steady increas
e computational power see Moore law gradual lessening dominance Chomskyan theories lingu
istics e g': 1.3900000000000001, 'transformational grammar whose theoretical underpinnin
gs discouraged sort corpus linguistics underlies machine learning approach language proc
essing': 3.4499999999999997, 'increasingly important medicine healthcare NLP used analyz
e notes text electronic health records would otherwise inaccessible study seeking improv
e care': 0.8000000000000002, 'Many different classes machine learning algorithms applied
natural language processing tasks': 3.42, 'algorithms take input large set features gene
rated input data': 0.89, 'speech 18': 0.2, 'models advantage express relative certainty
many different possible answers rather one producing reliable results model included com
ponent larger system': 1.4900000000000004, 'earliest used machine learning algorithms de
```

```
cision trees produced systems hard rules similar existing hand written rules': 1.5700000
000000005, 'cache language models upon many speech recognition systems rely examples sta
tistical models': 1.95, 'Since neural turn statistical methods NLP research largely repl
aced neural networks': 1.4400000000000002, 'However continue relevant contexts statistic
al interpretability transparency required': 0.66, 'major drawback statistical methods re
quire elaborate feature engineering': 0.7400000000000001, 'Since 2015 19 field thus larg
ely abandoned statistical methods shifted neural networks machine learning': 1.94, 'Late
st works tend use non technical structure given task build proper neural network': 1.390
0000000000001, '20 following list commonly researched tasks natural language processin
g': 2.4299999999999997, 'tasks direct real world applications others commonly serve subt
asks used aid solving larger tasks': 0.9000000000000001, 'Though natural language proces
sing tasks closely intertwined subdivided categories convenience': 2.329999999999999, 'c
oarse division given': 0.16, 'Based long standing trends field possible extrapolate futu
re directions NLP': 0.55, 'Cognition refers mental action process acquiring knowledge un
derstanding thought experience senses': 0.4900000000000001, '37 Cognitive science interd
isciplinary scientific study mind processes': 0.7400000000000001, '38 Cognitive linguist
ics interdisciplinary branch linguistics combining knowledge research psychology linguis
tics': 1.7800000000000002, '39 Especially age symbolic NLP area computational linguistic
s maintained strong ties cognitive studies': 1.15, 'recently ideas cognitive NLP revived
approach achieve explainability e g notion cognitive AI': 1.7100000000000004, '46 Likewi
se ideas cognitive NLP inherent neural models multimodal NLP although rarely made explic
it': 0.9000000000000001, '47 Media related Natural language processing Wikimedia Common
s': 2.27}
```

1.6 - Summarize:

In [41]:

```
import heapq

# Re-introducing punctuation and stop words for proper and readable summaries
#This is based of the top n sentences with highest scores
# Summary based on # sentences:
sentences=sent_tokenize(text)
sentence_scores={}

for sentence in sentences:
    for word in word_tokenize(sentence.lower()):
        # Avoid lengthy/wordy sentences
        if word in fd_weighted and len(sentence.split(' '))<20:
            if sentence not in sentence_scores:
                sentence_scores[sentence] = fd_weighted[word]
            else:
                sentence_scores[sentence] += fd_weighted[word]
summary_sentences = heapq.nlargest(5, sentence_scores, key=sentence_scores.get)

summary = ' '.join(summary_sentences)

print(summary)
```

```
Challenges in natural language processing frequently involve speech recognition, natural
language understanding, and natural language generation. Up to the 1980s, most natural l
anguage processing systems were based on complex sets of hand-written rules. transformat
ional grammar), whose theoretical underpinnings discouraged the sort of corpus linguisti
cs that underlies the machine-learning approach to language processing. [20]
The following is a list of some of the most commonly researched tasks in natural languag
e processing. Though natural language processing tasks are closely intertwined, they can
be subdivided into categories for convenience.
```

In [10]:

```
# Summary with # of words restriction:
count = 0
max_words = 50
summary = ""
```

```
for i in sentence_scores:
    if count < max_words:
        if len(i.split(' ')) < max_words:
            count+=len(i.split(' '))
        if count < max_words:
            summary+=i+' '

print(summary+'\n','Length of summary: ',len(summary.split(' ')))
```

```
Challenges in natural language processing frequently involve speech recognition, natural
language understanding, and natural language generation. Natural language processing has
its roots in the 1950s. Up to the 1980s, most natural language processing systems were b
ased on complex sets of hand-written rules.
 Length of summary:  43
```

In [11]:

```
# Summary with ratio restriction:
ratio=0.15 # Summary size/Original size
count = 0
summary = ""
for i in sentence_scores:
    if count/len(tokens) <= ratio:
        if len(i.split(' '))/len(tokens) < ratio:
            count+=len(i.split(' '))
        if count/len(tokens) < max_words:
            summary+=i+' '

print(summary+'\n','Length of summary: ',len(summary.split(' ')),'\nRatio=',len(summary
```

```
Challenges in natural language processing frequently involve speech recognition, natural
language understanding, and natural language generation. Natural language processing has
its roots in the 1950s. Up to the 1980s, most natural language processing systems were b
ased on complex sets of hand-written rules. transformational grammar), whose theoretical
underpinnings discouraged the sort of corpus linguistics that underlies the machine-lear
ning approach to language processing. Many different classes of machine-learning algorit
hms have been applied to natural-language-processing tasks. These algorithms take as inp
ut a large set of "features" that are generated from the input data. speech[18]). The ca
che language models upon which many speech recognition systems now rely are examples of
such statistical models. Since the neural turn, statistical methods in NLP research have
been largely replaced by neural networks. However, they continue to be relevant for cont
exts in which statistical interpretability and transparency is required. A major drawbac
k of statistical methods is that they require elaborate feature engineering. Since 2015,
[19] the field has thus largely abandoned statistical methods and shifted to neural netw
orks for machine learning. Latest works tend to use non-technical structure of a given t
ask to build proper neural network. [20]
The following is a list of some of the most commonly researched tasks in natural languag
e processing.
 Length of summary:  206
Ratio= 0.1509157509157509
```

2-Text Summarization with N-grams

2.1-2:

In [42]:

```
from nltk.util import ngrams

def generate_ngrams(txt,n):
    n_grams = ngrams(word_tokenize(txt.lower()),n)
    return [' '.join(gram) for gram in n_grams]
text = scrape_site(src)
```

```
freqdist = nltk.FreqDist(generate_ngrams(text,3))
#freqdist.most_common(20)
freqdist.plot(20)
```

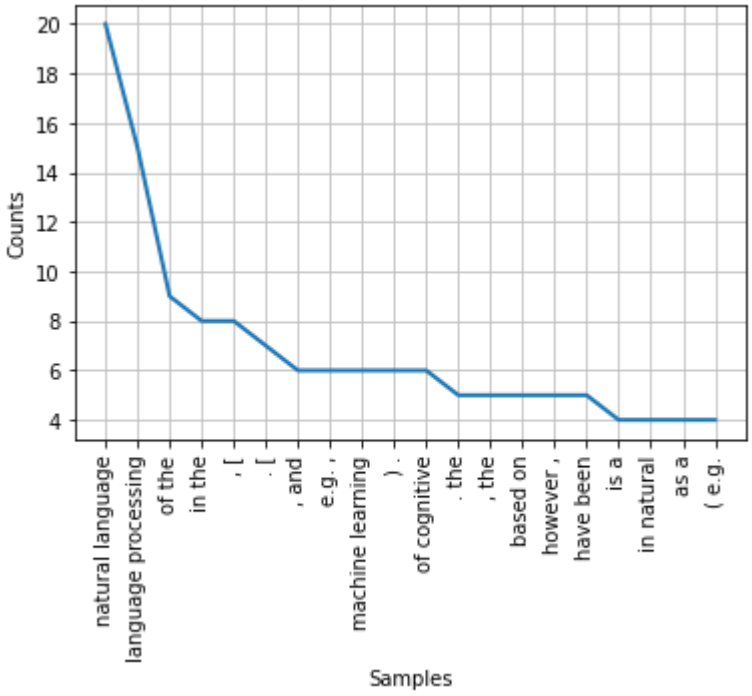
```
<ipython-input-32-1fa1d72a8b93>:2: DeprecationWarning: cafile, capath and cadefault are
deprecated, use a custom context instead.
  response = ur.urlopen(url,cafile=certifi.where())
['natural language', 'language processing', 'processing (', '( nlp', 'nlp )', ') is', 'i
s a', 'a subfield', 'subfield of', 'of linguistics', 'linguistics ,', ', computer', 'com
puter science', 'science ,', ', and', 'and artificial', 'artificial intelligence', 'inte
lligence concerned', 'concerned with', 'with the', 'the interactions', 'interactions bet
ween', 'between computers', 'computers and', 'and human', 'human language', 'language
,', ', in', 'in particular', 'particular how', 'how to', 'to program', 'program computer
s', 'computers to', 'to process', 'process and', 'and analyze', 'analyze large', 'large
amounts', 'amounts of', 'of natural', 'natural language', 'language data', 'data .', '.
the', 'the goal', 'goal is', 'is a', 'a computer', 'computer capable', 'capable of', 'of
``', ``` understanding', "understanding ''", "'' the", 'the contents', 'contents of', 'o
f documents', 'documents ,', ', including', 'including the', 'the contextual', 'contextu
al nuances', 'nuances of', 'of the', 'the language', 'language within', 'within them',
'them .', '. the', 'the technology', 'technology can', 'can then', 'then accurately', 'a
ccurately extract', 'extract information', 'information and', 'and insights', 'insights
contained', 'contained in', 'in the', 'the documents', 'documents as', 'as well', 'well
as', 'as categorize', 'categorize and', 'and organize', 'organize the', 'the documents',
'documents themselves', 'themselves .', '. challenges', 'challenges in', 'in natural',
'natural language', 'language processing', 'processing frequently', 'frequently involv
e', 'involve speech', 'speech recognition', 'recognition ,', ', natural', 'natural langu
age', 'language understanding', 'understanding ,', ', and', 'and natural', 'natural lang
uage', 'language generation', 'generation .', '. natural', 'natural language', 'language
processing', 'processing has', 'has its', 'its roots', 'roots in', 'in the', 'the 1950
s', '1950s .', '. already', 'already in', 'in 1950', '1950 ,', ', alan', 'alan turing',
'turing published', 'published an', 'an article', 'article titled', 'titled ``', ``` com
puting', 'computing machinery', 'machinery and', 'and intelligence', "intelligence ''",
"'' which", 'which proposed', 'proposed what', 'what is', 'is now', 'now called', 'calle
d the', 'the turing', 'turing test', 'test as', 'as a', 'a criterion', 'criterion of',
'of intelligence', 'intelligence ,', ', a', 'a task', 'task that', 'that involves', 'inv
olves the', 'the automated', 'automated interpretation', 'interpretation and', 'and gene
ration', 'generation of', 'of natural', 'natural language', 'language ,', ', but', 'but
at', 'at the', 'the time', 'time not', 'not articulated', 'articulated as', 'as a', 'a p
roblem', 'problem separate', 'separate from', 'from artificial', 'artificial intelligenc
e', 'intelligence .', '. the', 'the premise', 'premise of', 'of symbolic', 'symbolic nl
p', 'nlp is', 'is well-summarized', 'well-summarized by', 'by john', 'john searle', "sea
rle 's", "'s chinese", 'chinese room', 'room experiment', 'experiment :', ': given', 'gi
ven a', 'a collection', 'collection of', 'of rules', 'rules (', '( e.g.', 'e.g. ,', ',
a', 'a chinese', 'chinese phrasebook', 'phrasebook ,', ', with', 'with questions', 'ques
tions and', 'and matching', 'matching answers', 'answers )', ') ,', ', the', 'the comput
er', 'computer emulates', 'emulates natural', 'natural language', 'language understandin
g', 'understanding (', '( or', 'or other', 'other nlp', 'nlp tasks', 'tasks )', ') by',
'by applying', 'applying those', 'those rules', 'rules to', 'to the', 'the data', 'data
it', 'it is', 'is confronted', 'confronted with', 'with .', '. up', 'up to', 'to the',
'the 1980s', '1980s ,', ', most', 'most natural', 'natural language', 'language processi
ng', 'processing systems', 'systems were', 'were based', 'based on', 'on complex', 'comp
lex sets', 'sets of', 'of hand-written', 'hand-written rules', 'rules .', '. starting',
'starting in', 'in the', 'the late', 'late 1980s', '1980s ,', ', however', 'however ,',
', there', 'there was', 'was a', 'a revolution', 'revolution in', 'in natural', 'natural
language', 'language processing', 'processing with', 'with the', 'the introduction', 'in
troduction of', 'of machine', 'machine learning', 'learning algorithms', 'algorithms fo
r', 'for language', 'language processing', 'processing .', '. this', 'this was', 'was du
e', 'due to', 'to both', 'both the', 'the steady', 'steady increase', 'increase in', 'in
computational', 'computational power', 'power (', '( see', 'see moore', "moore 's", "'s
law", 'law )', ') and', 'and the', 'the gradual', 'gradual lessening', 'lessening of',
'of the', 'the dominance', 'dominance of', 'of chomskyan', 'chomskyan theories', 'theori
es of', 'of linguistics', 'linguistics (', '( e.g', 'e.g .', '. transformational', 'tran
sformational grammar', 'grammar )', ') ,', ', whose', 'whose theoretical', 'theoretical
underpinnings', 'underpinnings discouraged', 'discouraged the', 'the sort', 'sort of',
```

'of corpus', 'corpus linguistics', 'linguistics that', 'that underlies', 'underlies th
e', 'the machine-learning', 'machine-learning approach', 'approach to', 'to language',
'language processing', 'processing .', '. [', '[6', '6]', '] in', 'in the', 'the 2010
s', '2010s ,', ', representation', 'representation learning', 'learning and', 'and dee
p', 'deep neural', 'neural network-style', 'network-style machine', 'machine learning',
'learning methods', 'methods became', 'became widespread', 'widespread in', 'in natura
l', 'natural language', 'language processing', 'processing ,', ', due', 'due in', 'in pa
rt', 'part to', 'to a', 'a flurry', 'flurry of', 'of results', 'results showing', 'showi
ng that', 'that such', 'such techniques', 'techniques [', '[7', '7]', '] [', '[8', '8
]', '] can', 'can achieve', 'achieve state-of-the-art', 'state-of-the-art results', 'res
ults in', 'in many', 'many natural', 'natural language', 'language tasks', 'tasks ,', ',
for', 'for example', 'example in', 'in language', 'language modeling', 'modeling ,', ',
[', '[9', '9]', '] parsing', 'parsing ,', ', [', '[10', '10]', '] [', '[11', '11
]', '] and', 'and many', 'many others', 'others .', '. this', 'this is', 'is increasingl
y', 'increasingly important', 'important in', 'in medicine', 'medicine and', 'and health
care', 'healthcare ,', ', where', 'where nlp', 'nlp is', 'is being', 'being used', 'used
to', 'to analyze', 'analyze notes', 'notes and', 'and text', 'text in', 'in electronic',
'electronic health', 'health records', 'records that', 'that would', 'would otherwise',
'otherwise be', 'be inaccessible', 'inaccessible for', 'for study', 'study when', 'when
seeking', 'seeking to', 'to improve', 'improve care', 'care .', '. [', '[12', '12]',
'] in', 'in the', 'the early', 'early days', 'days ,', ', many', 'many language-processi
ng', 'language-processing systems', 'systems were', 'were designed', 'designed by', 'by
symbolic', 'symbolic methods', 'methods ,', ', i.e.', 'i.e. ,', ', the', 'the hand-codin
g', 'hand-coding of', 'of a', 'a set', 'set of', 'of rules', 'rules ,', ', coupled', 'co
upled with', 'with a', 'a dictionary', 'dictionary lookup', 'lookup :', ': [', '[13',
'13]', '] [', '[14', '14]', '] such', 'such as', 'as by', 'by writing', 'writing gram
mars', 'grammars or', 'or devising', 'devising heuristic', 'heuristic rules', 'rules fo
r', 'for stemming', 'stemming .', '. more', 'more recent', 'recent systems', 'systems ba
sed', 'based on', 'on machine-learning', 'machine-learning algorithms', 'algorithms hav
e', 'have many', 'many advantages', 'advantages over', 'over hand-produced', 'hand-produ
ced rules', 'rules :', ': despite', 'despite the', 'the popularity', 'popularity of', 'o
f machine', 'machine learning', 'learning in', 'in nlp', 'nlp research', 'research ,',
', symbolic', 'symbolic methods', 'methods are', 'are still', 'still (', '(2020', '2020
)', ') commonly', 'commonly used', 'used :', ': since', 'since the', 'the so-called', 's
o-called ``', '`` statistical', 'statistical revolution', "revolution ''", "'' [", '[1
5', '15]', '] [', '[16', '16]', '] in', 'in the', 'the late', 'late 1980s', '1980s an
d', 'and mid-1990s', 'mid-1990s ,', ', much', 'much natural', 'natural language', 'langu
age processing', 'processing research', 'research has', 'has relied', 'relied heavily',
'heavily on', 'on machine', 'machine learning', 'learning .', '. the', 'the machine-lear
ning', 'machine-learning paradigm', 'paradigm calls', 'calls instead', 'instead for', 'f
or using', 'using statistical', 'statistical inference', 'inference to', 'to automatical
ly', 'automatically learn', 'learn such', 'such rules', 'rules through', 'through the',
'the analysis', 'analysis of', 'of large', 'large corpora', 'corpora (', '(the', 'the p
lural', 'plural form', 'form of', 'of corpus', 'corpus ,', ', is', 'is a', 'a set', 'set
of', 'of documents', 'documents ,', ', possibly', 'possibly with', 'with human', 'human
or', 'or computer', 'computer annotations', 'annotations)', ') of', 'of typical', 'typi
cal real-world', 'real-world examples', 'examples .', '. many', 'many different', 'diffe
rent classes', 'classes of', 'of machine-learning', 'machine-learning algorithms', 'algo
rithms have', 'have been', 'been applied', 'applied to', 'to natural-language-processin
g', 'natural-language-processing tasks', 'tasks .', '. these', 'these algorithms', 'algo
rithms take', 'take as', 'as input', 'input a', 'a large', 'large set', 'set of', 'of `
`', '`` features', "features ''", "'' that", 'that are', 'are generated', 'generated fro
m', 'from the', 'the input', 'input data', 'data .', '. increasingly', 'increasingly ,',
', however', 'however ,', ', research', 'research has', 'has focused', 'focused on', 'on
statistical', 'statistical models', 'models ,', ', which', 'which make', 'make soft', 's
oft ,', ', probabilistic', 'probabilistic decisions', 'decisions based', 'based on', 'on
attaching', 'attaching real-valued', 'real-valued weights', 'weights to', 'to each', 'ea
ch input', 'input feature', 'feature (', '(complex-valued', 'complex-valued embedding
s', 'embeddings ,', ', [', '[17', '17]', '] and', 'and neural', 'neural networks', 'ne
tworks in', 'in general', 'general have', 'have also', 'also been', 'been proposed', 'pr
oposed ,', ', for', 'for e.g', 'e.g .', '. speech', 'speech [', '[18', '18]', '])',
') .', '. such', 'such models', 'models have', 'have the', 'the advantage', 'advantage t
hat', 'that they', 'they can', 'can express', 'express the', 'the relative', 'relative c
ertainty', 'certainty of', 'of many', 'many different', 'different possible', 'possible
answers', 'answers rather', 'rather than', 'than only', 'only one', 'one ,', ', producin

g', 'producing more', 'more reliable', 'reliable results', 'results when', 'when such',
'such a', 'a model', 'model is', 'is included', 'included as', 'as a', 'a component', 'c
omponent of', 'of a', 'a larger', 'larger system', 'system .', '. some', 'some of', 'of
the', 'the earliest-used', 'earliest-used machine', 'machine learning', 'learning algori
thms', 'algorithms ,', ', such', 'such as', 'as decision', 'decision trees', 'trees ,',
', produced', 'produced systems', 'systems of', 'of hard', 'hard if-then', 'if-then rule
s', 'rules similar', 'similar to', 'to existing', 'existing hand-written', 'hand-written
rules', 'rules .', '. however', 'however ,', ', part-of-speech', 'part-of-speech taggin
g', 'tagging introduced', 'introduced the', 'the use', 'use of', 'of hidden', 'hidden ma
rkov', 'markov models', 'models to', 'to natural', 'natural language', 'language process
ing', 'processing ,', ', and', 'and increasingly', 'increasingly ,', ', research', 'rese
arch has', 'has focused', 'focused on', 'on statistical', 'statistical models', 'models
,', ', which', 'which make', 'make soft', 'soft ,', ', probabilistic', 'probabilistic de
cisions', 'decisions based', 'based on', 'on attaching', 'attaching real-valued', 'real-
valued weights', 'weights to', 'to the', 'the features', 'features making', 'making up',
'up the', 'the input', 'input data', 'data .', '. the', 'the cache', 'cache language',
'language models', 'models upon', 'upon which', 'which many', 'many speech', 'speech rec
ognition', 'recognition systems', 'systems now', 'now rely', 'rely are', 'are examples',
'examples of', 'of such', 'such statistical', 'statistical models', 'models .', '. suc
h', 'such models', 'models are', 'are generally', 'generally more', 'more robust', 'robu
st when', 'when given', 'given unfamiliar', 'unfamiliar input', 'input ,', ', especiall
y', 'especially input', 'input that', 'that contains', 'contains errors', 'errors (', '(
as', 'as is', 'is very', 'very common', 'common for', 'for real-world', 'real-world dat
a', 'data)', ') ,', ', and', 'and produce', 'produce more', 'more reliable', 'reliable
results', 'results when', 'when integrated', 'integrated into', 'into a', 'a larger', 'l
arger system', 'system comprising', 'comprising multiple', 'multiple subtasks', 'subtask
s .', '. since', 'since the', 'the neural', 'neural turn', 'turn ,', ', statistical', 's
tatistical methods', 'methods in', 'in nlp', 'nlp research', 'research have', 'have bee
n', 'been largely', 'largely replaced', 'replaced by', 'by neural', 'neural networks',
'networks .', '. however', 'however ,', ', they', 'they continue', 'continue to', 'to b
e', 'be relevant', 'relevant for', 'for contexts', 'contexts in', 'in which', 'which sta
tistical', 'statistical interpretability', 'interpretability and', 'and transparency',
'transparency is', 'is required', 'required .', '. a', 'a major', 'major drawback', 'dra
wback of', 'of statistical', 'statistical methods', 'methods is', 'is that', 'that the
y', 'they require', 'require elaborate', 'elaborate feature', 'feature engineering', 'en
gineering .', '. since', 'since 2015', '2015 ,', ', [', '[19', '19]', '] the', 'the fi
eld', 'field has', 'has thus', 'thus largely', 'largely abandoned', 'abandoned statistic
al', 'statistical methods', 'methods and', 'and shifted', 'shifted to', 'to neural', 'ne
ural networks', 'networks for', 'for machine', 'machine learning', 'learning .', '. popu
lar', 'popular techniques', 'techniques include', 'include the', 'the use', 'use of', 'o
f word', 'word embeddings', 'embeddings to', 'to capture', 'capture semantic', 'semantic
properties', 'properties of', 'of words', 'words ,', ', and', 'and an', 'an increase',
'increase in', 'in end-to-end', 'end-to-end learning', 'learning of', 'of a', 'a higher-
level', 'higher-level task', 'task (', '(e.g.', 'e.g. ,', ', question', 'question answe
ring', 'answering)', ') instead', 'instead of', 'of relying', 'relying on', 'on a', 'a
pipeline', 'pipeline of', 'of separate', 'separate intermediate', 'intermediate tasks',
'tasks (', '(e.g.', 'e.g. ,', ', part-of-speech', 'part-of-speech tagging', 'tagging an
d', 'and dependency', 'dependency parsing', 'parsing)', ') .', '. in', 'in some', 'some
areas', 'areas ,', ', this', 'this shift', 'shift has', 'has entailed', 'entailed substa
ntial', 'substantial changes', 'changes in', 'in how', 'how nlp', 'nlp systems', 'system
s are', 'are designed', 'designed ,', ', such', 'such that', 'that deep', 'deep neural',
'neural network-based', 'network-based approaches', 'approaches may', 'may be', 'be view
ed', 'viewed as', 'as a', 'a new', 'new paradigm', 'paradigm distinct', 'distinct from',
'from statistical', 'statistical natural', 'natural language', 'language processing', 'p
rocessing .', '. for', 'for instance', 'instance ,', ', the', 'the term', 'term neural',
'neural machine', 'machine translation', 'translation (', '(nmt', 'nmt)', ') emphasize
s', 'emphasizes the', 'the fact', 'fact that', 'that deep', 'deep learning-based', 'lear
ning-based approaches', 'approaches to', 'to machine', 'machine translation', 'translati
on directly', 'directly learn', 'learn sequence-to-sequence', 'sequence-to-sequence tran
sformations', 'transformations ,', ', obviating', 'obviating the', 'the need', 'need fo
r', 'for intermediate', 'intermediate steps', 'steps such', 'such as', 'as word', 'word
alignment', 'alignment and', 'and language', 'language modeling', 'modeling that', 'that
was', 'was used', 'used in', 'in statistical', 'statistical machine', 'machine translati
on', 'translation (', '(smt', 'smt)', ') .', '. latest', 'latest works', 'works tend',
'tend to', 'to use', 'use non-technical', 'non-technical structure', 'structure of', 'of

a', 'a given', 'given task', 'task to', 'to build', 'build proper', 'proper neural', 'ne
ural network', 'network .', '. [', '[20', '20]', '] the', 'the following', 'following
is', 'is a', 'a list', 'list of', 'of some', 'some of', 'of the', 'the most', 'most comm
only', 'commonly researched', 'researched tasks', 'tasks in', 'in natural', 'natural lan
guage', 'language processing', 'processing .', '. some', 'some of', 'of these', 'these t
asks', 'tasks have', 'have direct', 'direct real-world', 'real-world applications', 'app
lications ,', ', while', 'while others', 'others more', 'more commonly', 'commonly serv
e', 'serve as', 'as subtasks', 'subtasks that', 'that are', 'are used', 'used to', 'to a
id', 'aid in', 'in solving', 'solving larger', 'larger tasks', 'tasks .', '. though', 't
hough natural', 'natural language', 'language processing', 'processing tasks', 'tasks ar
e', 'are closely', 'closely intertwined', 'intertwined ,', ', they', 'they can', 'can b
e', 'be subdivided', 'subdivided into', 'into categories', 'categories for', 'for conven
ience', 'convenience .', '. a', 'a coarse', 'coarse division', 'division is', 'is give
n', 'given below', 'below .', '. based', 'based on', 'on long-standing', 'long-standing
trends', 'trends in', 'in the', 'the field', 'field ,', ', it', 'it is', 'is possible',
'possible to', 'to extrapolate', 'extrapolate future', 'future directions', 'directions
of', 'of nlp', 'nlp .', '. as', 'as of', 'of 2020', '2020 ,', ', three', 'three trends',
'trends among', 'among the', 'the topics', 'topics of', 'of the', 'the long-standing',
'long-standing series', 'series of', 'of conll', 'conll shared', 'shared tasks', 'tasks
can', 'can be', 'be observed', 'observed :', ': [', '[36', '36]', '] most', 'most high
er-level', 'higher-level nlp', 'nlp applications', 'applications involve', 'involve aspe
cts', 'aspects that', 'that emulate', 'emulate intelligent', 'intelligent behaviour', 'b
ehaviour and', 'and apparent', 'apparent comprehension', 'comprehension of', 'of natura
l', 'natural language', 'language .', '. more', 'more broadly', 'broadly speaking', 'spe
aking ,', ', the', 'the technical', 'technical operationalization', 'operationalization
of', 'of increasingly', 'increasingly advanced', 'advanced aspects', 'aspects of', 'of c
ognitive', 'cognitive behaviour', 'behaviour represents', 'represents one', 'one of', 'o
f the', 'the developmental', 'developmental trajectories', 'trajectories of', 'of nlp',
'nlp (', '(see', 'see trends', 'trends among', 'among conll', 'conll shared', 'shared t
asks', 'tasks above', 'above)', ') .', '. cognition', 'cognition refers', 'refers to',
'to ``', '`` the', 'the mental', 'mental action', 'action or', 'or process', 'process o
f', 'of acquiring', 'acquiring knowledge', 'knowledge and', 'and understanding', 'unders
tanding through', 'through thought', 'thought ,', ', experience', 'experience ,', ', an
d', 'and the', 'the senses', 'senses .', '. ``', '`` [', '[37', '37]', '] cognitive',
'cognitive science', 'science is', 'is the', 'the interdisciplinary', 'interdisciplinary
,', ', scientific', 'scientific study', 'study of', 'of the', 'the mind', 'mind and', 'a
nd its', 'its processes', 'processes .', '. [', '[38', '38]', '] cognitive', 'cognitiv
e linguistics', 'linguistics is', 'is an', 'an interdisciplinary', 'interdisciplinary br
anch', 'branch of', 'of linguistics', 'linguistics ,', ', combining', 'combining knowled
ge', 'knowledge and', 'and research', 'research from', 'from both', 'both psychology',
'psychology and', 'and linguistics', 'linguistics .', '. [', '[39', '39]', '] especial
ly', 'especially during', 'during the', 'the age', 'age of', 'of symbolic', 'symbolic nl
p', 'nlp ,', ', the', 'the area', 'area of', 'of computational', 'computational linguist
ics', 'linguistics maintained', 'maintained strong', 'strong ties', 'ties with', 'with c
ognitive', 'cognitive studies', 'studies .', '. as', 'as an', 'an example', 'example ,',
', george', 'george lakoff', 'lakoff offers', 'offers a', 'a methodology', 'methodology
to', 'to build', 'build natural', 'natural language', 'language processing', 'processing
(', '(nlp', 'nlp)', ') algorithms', 'algorithms through', 'through the', 'the perspect
ive', 'perspective of', 'of cognitive', 'cognitive science', 'science ,', ', along', 'al
ong with', 'with the', 'the findings', 'findings of', 'of cognitive', 'cognitive linguis
tics', 'linguistics ,', ', [', '[40', '40]', '] with', 'with two', 'two defining', 'de
fining aspects', 'aspects :', ': ties', 'ties with', 'with cognitive', 'cognitive lingui
stics', 'linguistics are', 'are part', 'part of', 'of the', 'the historical', 'historica
l heritage', 'heritage of', 'of nlp', 'nlp ,', ', but', 'but they', 'they have', 'have b
een', 'been less', 'less frequently', 'frequently addressed', 'addressed since', 'since
the', 'the statistical', 'statistical turn', 'turn during', 'during the', 'the 1990s',
'1990s .', '. nevertheless', 'nevertheless ,', ', approaches', 'approaches to', 'to deve
lop', 'develop cognitive', 'cognitive models', 'models towards', 'towards technically',
'technically operationalizable', 'operationalizable frameworks', 'frameworks have', 'hav
e been', 'been pursued', 'pursued in', 'in the', 'the context', 'context of', 'of variou
s', 'various frameworks', 'frameworks ,', ', e.g.', 'e.g. ,', ', of', 'of cognitive', 'c
ognitive grammar', 'grammar ,', ', [', '[42', '42]', '] functional', 'functional gramm
ar', 'grammar ,', ', [', '[43', '43]', '] construction', 'construction grammar', 'gram
mar ,', ', [', '[44', '44]', '] computational', 'computational psycholinguistics', 'ps
ycholinguistics and', 'and cognitive', 'cognitive neuroscience', 'neuroscience (', '(e.

```
g.', 'e.g. ,', ', act-r', 'act-r )', ') ,', ', however', 'however ,', ', with', 'with li
mited', 'limited uptake', 'uptake in', 'in mainstream', 'mainstream nlp', 'nlp (', '( a
s', 'as measured', 'measured by', 'by presence', 'presence on', 'on major', 'major confe
rences', 'conferences [', '[ 45', '45 ]', '] of', 'of the', 'the acl', 'acl )', ') .',
'. more', 'more recently', 'recently ,', ', ideas', 'ideas of', 'of cognitive', 'cogniti
ve nlp', 'nlp have', 'have been', 'been revived', 'revived as', 'as an', 'an approach',
'approach to', 'to achieve', 'achieve explainability', 'explainability ,', ', e.g.', 'e.
g. ,', ', under', 'under the', 'the notion', 'notion of', 'of ``', '`` cognitive', 'cogn
itive ai', "ai ''", "'' .", '. [', '[ 46', '46 ]', '] likewise', 'likewise ,', ', idea
s', 'ideas of', 'of cognitive', 'cognitive nlp', 'nlp are', 'are inherent', 'inherent t
o', 'to neural', 'neural models', 'models multimodal', 'multimodal nlp', 'nlp (', '( alt
hough', 'although rarely', 'rarely made', 'made explicit', 'explicit )', ') .', '. [',
'[ 47', '47 ]', '] media', 'media related', 'related to', 'to natural', 'natural languag
e', 'language processing', 'processing at', 'at wikimedia', 'wikimedia commons']
```



Out[42]: `<AxesSubplot:xlabel='Samples', ylabel='Counts'>`

Weighted Frequencies Calculation:

In [21]:
```
fd_weighted={}
max_freq=max(freqdist.values())
for i in freqdist:
    fd_weighted[i]=freqdist[i]/max_freq
#pprint(sorted(fd_weighted.items(), key=lambda item: item[1],reverse=True))
```

Sentence Scoring:

In [47]:
```
sentence_scores2={}

# Function only needs source_text (non-tokenized) and n_grams level (2,3...)

def calculate_sentence_scores_ngram(source_text,n_grams):

    # Computes word freqs
    freqdist = nltk.FreqDist(generate_ngrams(source_text,n_grams))

    # Weighted freqs
    fd_weighted={}
```

```
    max_freq=max(freqdist.values())
    for i in freqdist:
        fd_weighted[i]=freqdist[i]/max_freq

    for sentence in freqdist:
        if sentence in fd_weighted:
            if sentence not in sentence_scores2:
                sentence_scores2[sentence] = fd_weighted[sentence]
            else:
                sentence_scores2[sentence] += fd_weighted[sentence]

calculate_sentence_scores_ngram(text,3)

print(sentence_scores2)
```

```
{'natural language processing': 1.0, 'in natural language': 0.3076923076923077, '( e.g.
,': 0.3076923076923077, 'language processing .': 0.3076923076923077, 'of natural languag
e': 0.23076923076923078, ', however ,': 0.23076923076923078, '] in the': 0.2307692307692
3078, 'grammar , [': 0.23076923076923078, 'language processing (': 0.15384615384615385,
'processing ( nlp': 0.15384615384615385, '( nlp )': 0.15384615384615385, 'of linguistics
,': 0.15384615384615385, 'data . the': 0.15384615384615385, 'of documents ,': 0.15384615
384615385, 'natural language understanding': 0.15384615384615385, 'of symbolic nlp': 0.1
5384615384615385, 'hand-written rules .': 0.15384615384615385, 'in the late': 0.15384615
384615385, 'the late 1980s': 0.15384615384615385, 'of machine learning': 0.1538461538461
5385, 'machine learning algorithms': 0.15384615384615385, 'language processing ,': 0.153
84615384615385, 'a set of': 0.15384615384615385, 'machine-learning algorithms have': 0.1
5384615384615385, 'in nlp research': 0.15384615384615385, 'machine learning .': 0.153846
15384615385, 'the input data': 0.15384615384615385, 'input data .': 0.15384615384615385,
', research has': 0.15384615384615385, 'research has focused': 0.15384615384615385, 'has
focused on': 0.15384615384615385, 'focused on statistical': 0.15384615384615385, 'on sta
tistical models': 0.15384615384615385, 'statistical models ,': 0.15384615384615385, 'mod
els , which': 0.15384615384615385, ', which make': 0.15384615384615385, 'which make sof
t': 0.15384615384615385, 'make soft ,': 0.15384615384615385, 'soft , probabilistic': 0.1
5384615384615385, ', probabilistic decisions': 0.15384615384615385, 'probabilistic decis
ions based': 0.15384615384615385, 'decisions based on': 0.15384615384615385, 'based on a
ttaching': 0.15384615384615385, 'on attaching real-valued': 0.15384615384615385, 'attach
ing real-valued weights': 0.15384615384615385, 'real-valued weights to': 0.1538461538461
5385, '. such models': 0.15384615384615385, 'more reliable results': 0.1538461538461538
5, 'reliable results when': 0.15384615384615385, 'a larger system': 0.15384615384615385,
'. some of': 0.15384615384615385, 'some of the': 0.15384615384615385, '. however ,': 0.1
5384615384615385, ', part-of-speech tagging': 0.15384615384615385, 'the use of': 0.15384
615384615385, 'to natural language': 0.15384615384615385, 'machine translation (': 0.153
84615384615385, 'conll shared tasks': 0.15384615384615385, 'ties with cognitive': 0.1538
4615384615385, ', e.g. ,': 0.15384615384615385, ', ideas of': 0.15384615384615385, 'idea
s of cognitive': 0.15384615384615385, 'of cognitive nlp': 0.15384615384615385, 'nlp ) i
s': 0.07692307692307693, ') is a': 0.07692307692307693, 'is a subfield': 0.0769230769230
7693, 'a subfield of': 0.07692307692307693, 'subfield of linguistics': 0.076923076923076
93, 'linguistics , computer': 0.07692307692307693, ', computer science': 0.0769230769230
7693, 'computer science ,': 0.07692307692307693, 'science , and': 0.07692307692307693,
', and artificial': 0.07692307692307693, 'and artificial intelligence': 0.07692307692307
693, 'artificial intelligence concerned': 0.07692307692307693, 'intelligence concerned w
ith': 0.07692307692307693, 'concerned with the': 0.07692307692307693, 'with the interact
ions': 0.07692307692307693, 'the interactions between': 0.07692307692307693, 'interactio
ns between computers': 0.07692307692307693, 'between computers and': 0.0769230769230769
3, 'computers and human': 0.07692307692307693, 'and human language': 0.0769230769230769
3, 'human language ,': 0.07692307692307693, 'language , in': 0.07692307692307693, ', in
particular': 0.07692307692307693, 'in particular how': 0.07692307692307693, 'particular
how to': 0.07692307692307693, 'how to program': 0.07692307692307693, 'to program compute
rs': 0.07692307692307693, 'program computers to': 0.07692307692307693, 'computers to pro
cess': 0.07692307692307693, 'to process and': 0.07692307692307693, 'process and analyz
e': 0.07692307692307693, 'and analyze large': 0.07692307692307693, 'analyze large amount
s': 0.07692307692307693, 'large amounts of': 0.07692307692307693, 'amounts of natural':
0.07692307692307693, 'natural language data': 0.07692307692307693, 'language data .': 0.
```

07692307692307693, '. the goal': 0.07692307692307693, 'the goal is': 0.0769230769230769
3, 'goal is a': 0.07692307692307693, 'is a computer': 0.07692307692307693, 'a computer c
apable': 0.07692307692307693, 'computer capable of': 0.07692307692307693, 'capable of `
`': 0.07692307692307693, 'of `` understanding': 0.07692307692307693, "`` understanding
''": 0.07692307692307693, "understanding '' the": 0.07692307692307693, "'' the content
s": 0.07692307692307693, 'the contents of': 0.07692307692307693, 'contents of document
s': 0.07692307692307693, 'documents , including': 0.07692307692307693, ', including th
e': 0.07692307692307693, 'including the contextual': 0.07692307692307693, 'the contextua
l nuances': 0.07692307692307693, 'contextual nuances of': 0.07692307692307693, 'nuances
of the': 0.07692307692307693, 'of the language': 0.07692307692307693, 'the language with
in': 0.07692307692307693, 'language within them': 0.07692307692307693, 'within them .':
0.07692307692307693, 'them . the': 0.07692307692307693, '. the technology': 0.0769230769
2307693, 'the technology can': 0.07692307692307693, 'technology can then': 0.07692307692
307693, 'can then accurately': 0.07692307692307693, 'then accurately extract': 0.0769230
7692307693, 'accurately extract information': 0.07692307692307693, 'extract information
and': 0.07692307692307693, 'information and insights': 0.07692307692307693, 'and insight
s contained': 0.07692307692307693, 'insights contained in': 0.07692307692307693, 'contai
ned in the': 0.07692307692307693, 'in the documents': 0.07692307692307693, 'the document
s as': 0.07692307692307693, 'documents as well': 0.07692307692307693, 'as well as': 0.07
692307692307693, 'well as categorize': 0.07692307692307693, 'as categorize and': 0.07692
307692307693, 'categorize and organize': 0.07692307692307693, 'and organize the': 0.0769
2307692307693, 'organize the documents': 0.07692307692307693, 'the documents themselve
s': 0.07692307692307693, 'documents themselves .': 0.07692307692307693, 'themselves . ch
allenges': 0.07692307692307693, '. challenges in': 0.07692307692307693, 'challenges in n
atural': 0.07692307692307693, 'language processing frequently': 0.07692307692307693, 'pr
ocessing frequently involve': 0.07692307692307693, 'frequently involve speech': 0.076923
07692307693, 'involve speech recognition': 0.07692307692307693, 'speech recognition ,':
0.07692307692307693, 'recognition , natural': 0.07692307692307693, ', natural language':
0.07692307692307693, 'language understanding ,': 0.07692307692307693, 'understanding , a
nd': 0.07692307692307693, ', and natural': 0.07692307692307693, 'and natural language':
0.07692307692307693, 'natural language generation': 0.07692307692307693, 'language gener
ation .': 0.07692307692307693, 'generation . natural': 0.07692307692307693, '. natural l
anguage': 0.07692307692307693, 'language processing has': 0.07692307692307693, 'processi
ng has its': 0.07692307692307693, 'has its roots': 0.07692307692307693, 'its roots in':
0.07692307692307693, 'roots in the': 0.07692307692307693, 'in the 1950s': 0.076923076923
07693, 'the 1950s .': 0.07692307692307693, '1950s . already': 0.07692307692307693, '. al
ready in': 0.07692307692307693, 'already in 1950': 0.07692307692307693, 'in 1950 ,': 0.0
7692307692307693, '1950 , alan': 0.07692307692307693, ', alan turing': 0.076923076923076
93, 'alan turing published': 0.07692307692307693, 'turing published an': 0.0769230769230
7693, 'published an article': 0.07692307692307693, 'an article titled': 0.07692307692307
693, 'article titled ``': 0.07692307692307693, 'titled `` computing': 0.0769230769230769
3, '`` computing machinery': 0.07692307692307693, 'computing machinery and': 0.076923076
92307693, 'machinery and intelligence': 0.07692307692307693, "and intelligence ''": 0.07
692307692307693, "intelligence '' which": 0.07692307692307693, "'' which proposed": 0.07
692307692307693, 'which proposed what': 0.07692307692307693, 'proposed what is': 0.07692
307692307693, 'what is now': 0.07692307692307693, 'is now called': 0.07692307692307693,
'now called the': 0.07692307692307693, 'called the turing': 0.07692307692307693, 'the tu
ring test': 0.07692307692307693, 'turing test as': 0.07692307692307693, 'test as a': 0.0
7692307692307693, 'as a criterion': 0.07692307692307693, 'a criterion of': 0.07692307692
307693, 'criterion of intelligence': 0.07692307692307693, 'of intelligence ,': 0.0769230
7692307693, 'intelligence , a': 0.07692307692307693, ', a task': 0.07692307692307693, 'a
task that': 0.07692307692307693, 'task that involves': 0.07692307692307693, 'that involv
es the': 0.07692307692307693, 'involves the automated': 0.07692307692307693, 'the automa
ted interpretation': 0.07692307692307693, 'automated interpretation and': 0.076923076923
07693, 'interpretation and generation': 0.07692307692307693, 'and generation of': 0.0769
2307692307693, 'generation of natural': 0.07692307692307693, 'natural language ,': 0.076
92307692307693, 'language , but': 0.07692307692307693, ', but at': 0.07692307692307693,
'but at the': 0.07692307692307693, 'at the time': 0.07692307692307693, 'the time not':
0.07692307692307693, 'time not articulated': 0.07692307692307693, 'not articulated as':
0.07692307692307693, 'articulated as a': 0.07692307692307693, 'as a problem': 0.07692307
692307693, 'a problem separate': 0.07692307692307693, 'problem separate from': 0.0769230
7692307693, 'separate from artificial': 0.07692307692307693, 'from artificial intelligen
ce': 0.07692307692307693, 'artificial intelligence .': 0.07692307692307693, 'intelligenc
e . the': 0.07692307692307693, '. the premise': 0.07692307692307693, 'the premise of':
0.07692307692307693, 'premise of symbolic': 0.07692307692307693, 'symbolic nlp is': 0.07

692307692307693, 'nlp is well-summarized': 0.07692307692307693, 'is well-summarized by': 0.07692307692307693, 'well-summarized by john': 0.07692307692307693, 'by john searle': 0.07692307692307693, "john searle 's": 0.07692307692307693, "searle 's chinese": 0.07692307692307693, "'s chinese room": 0.07692307692307693, 'chinese room experiment': 0.07692307692307693, 'room experiment :': 0.07692307692307693, 'experiment : given': 0.07692307692307693, ': given a': 0.07692307692307693, 'given a collection': 0.07692307692307693, 'a collection of': 0.07692307692307693, 'collection of rules': 0.07692307692307693, 'of rules (': 0.07692307692307693, 'rules (e.g.': 0.07692307692307693, 'e.g. , a': 0.07692307692307693, ', a chinese': 0.07692307692307693, 'a chinese phrasebook': 0.07692307692307693, 'chinese phrasebook ,': 0.07692307692307693, 'phrasebook , with': 0.07692307692307693, ', with questions': 0.07692307692307693, 'with questions and': 0.07692307692307693, 'questions and matching': 0.07692307692307693, 'and matching answers': 0.07692307692307693, 'matching answers)': 0.07692307692307693, 'answers) ,': 0.07692307692307693, ') , the': 0.07692307692307693, ', the computer': 0.07692307692307693, 'the computer emulates': 0.07692307692307693, 'computer emulates natural': 0.07692307692307693, 'emulates natural language': 0.07692307692307693, 'language understanding (': 0.07692307692307693, 'understanding (or': 0.07692307692307693, '(or other': 0.07692307692307693, 'or other nlp': 0.07692307692307693, 'other nlp tasks': 0.07692307692307693, 'nlp tasks)': 0.07692307692307693, 'tasks) by': 0.07692307692307693, ') by applying': 0.07692307692307693, 'by applying those': 0.07692307692307693, 'applying those rules': 0.07692307692307693, 'those rules to': 0.07692307692307693, 'rules to the': 0.07692307692307693, 'to the data': 0.07692307692307693, 'the data it': 0.07692307692307693, 'data it is': 0.07692307692307693, 'it is confronted': 0.07692307692307693, 'is confronted with': 0.07692307692307693, 'confronted with .': 0.07692307692307693, 'with . up': 0.07692307692307693, '. up to': 0.07692307692307693, 'up to the': 0.07692307692307693, 'to the 1980s': 0.07692307692307693, 'the 1980s ,': 0.07692307692307693, '1980s , most': 0.07692307692307693, ', most natural': 0.07692307692307693, 'most natural language': 0.07692307692307693, 'language processing systems': 0.07692307692307693, 'processing systems were': 0.07692307692307693, 'systems were based': 0.07692307692307693, 'were based on': 0.07692307692307693, 'based on complex': 0.07692307692307693, 'on complex sets': 0.07692307692307693, 'complex sets of': 0.07692307692307693, 'sets of hand-written': 0.07692307692307693, 'of hand-written rules': 0.07692307692307693, 'rules . starting': 0.07692307692307693, '. starting in': 0.07692307692307693, 'starting in the': 0.07692307692307693, 'late 1980s ,': 0.07692307692307693, '1980s , however': 0.07692307692307693, 'however , there': 0.07692307692307693, ', there was': 0.07692307692307693, 'there was a': 0.07692307692307693, 'was a revolution': 0.07692307692307693, 'a revolution in': 0.07692307692307693, 'revolution in natural': 0.07692307692307693, 'language processing with': 0.07692307692307693, 'processing with the': 0.07692307692307693, 'with the introduction': 0.07692307692307693, 'the introduction of': 0.07692307692307693, 'introduction of machine': 0.07692307692307693, 'learning algorithms for': 0.07692307692307693, 'algorithms for language': 0.07692307692307693, 'for language processing': 0.07692307692307693, 'processing . this': 0.07692307692307693, '. this was': 0.07692307692307693, 'this was due': 0.07692307692307693, 'was due to': 0.07692307692307693, 'due to both': 0.07692307692307693, 'to both the': 0.07692307692307693, 'both the steady': 0.07692307692307693, 'the steady increase': 0.07692307692307693, 'steady increase in': 0.07692307692307693, 'increase in computational': 0.07692307692307693, 'in computational power': 0.07692307692307693, 'computational power (': 0.07692307692307693, 'power (see': 0.07692307692307693, '(see moore': 0.07692307692307693, "see moore 's": 0.07692307692307693, "moore 's law": 0.07692307692307693, "'s law)": 0.07692307692307693, 'law) and': 0.07692307692307693, ') and the': 0.07692307692307693, 'and the gradual': 0.07692307692307693, 'the gradual lessening': 0.07692307692307693, 'gradual lessening of': 0.07692307692307693, 'lessening of the': 0.07692307692307693, 'of the dominance': 0.07692307692307693, 'the dominance of': 0.07692307692307693, 'dominance of chomskyan': 0.07692307692307693, 'of chomskyan theories': 0.07692307692307693, 'chomskyan theories of': 0.07692307692307693, 'theories of linguistics': 0.07692307692307693, 'of linguistics (': 0.07692307692307693, 'linguistics (e.g': 0.07692307692307693, '(e.g .': 0.07692307692307693, 'e.g . transformational': 0.07692307692307693, '. transformational grammar': 0.07692307692307693, 'transformational grammar)': 0.07692307692307693, 'grammar) ,': 0.07692307692307693, ') , whose': 0.07692307692307693, ', whose theoretical': 0.07692307692307693, 'whose theoretical underpinnings': 0.07692307692307693, 'theoretical underpinnings discouraged': 0.07692307692307693, 'underpinnings discouraged the': 0.07692307692307693, 'discouraged the sort': 0.07692307692307693, 'the sort of': 0.07692307692307693, 'sort of corpus': 0.07692307692307693, 'of corpus linguistics': 0.07692307692307693, 'corpus linguistics that': 0.07692307692307693, 'linguistics that underlies': 0.07692307692307693, 'that underlies the': 0.07692307692307693, 'underlies the machine-learning': 0.07692307692307693, 'the machine-learning approach': 0.07692307692307693, 'machi

ne-learning approach to': 0.07692307692307693, 'approach to language': 0.076923076923076
93, 'to language processing': 0.07692307692307693, 'processing . [': 0.0769230769230769
3, '. [6': 0.07692307692307693, '[6]': 0.07692307692307693, '6] in': 0.0769230769230
7693, 'in the 2010s': 0.07692307692307693, 'the 2010s ,': 0.07692307692307693, '2010s ,
representation': 0.07692307692307693, ', representation learning': 0.07692307692307693,
'representation learning and': 0.07692307692307693, 'learning and deep': 0.0769230769230
7693, 'and deep neural': 0.07692307692307693, 'deep neural network-style': 0.07692307692
307693, 'neural network-style machine': 0.07692307692307693, 'network-style machine lear
ning': 0.07692307692307693, 'machine learning methods': 0.07692307692307693, 'learning m
ethods became': 0.07692307692307693, 'methods became widespread': 0.07692307692307693,
'became widespread in': 0.07692307692307693, 'widespread in natural': 0.0769230769230769
3, 'processing , due': 0.07692307692307693, ', due in': 0.07692307692307693, 'due in par
t': 0.07692307692307693, 'in part to': 0.07692307692307693, 'part to a': 0.0769230769230
7693, 'to a flurry': 0.07692307692307693, 'a flurry of': 0.07692307692307693, 'flurry of
results': 0.07692307692307693, 'of results showing': 0.07692307692307693, 'results showi
ng that': 0.07692307692307693, 'showing that such': 0.07692307692307693, 'that such tech
niques': 0.07692307692307693, 'such techniques [': 0.07692307692307693, 'techniques [
7': 0.07692307692307693, '[7]': 0.07692307692307693, '7] [': 0.07692307692307693, ']
[8': 0.07692307692307693, '[8]': 0.07692307692307693, '8] can': 0.07692307692307693,
'] can achieve': 0.07692307692307693, 'can achieve state-of-the-art': 0.0769230769230769
3, 'achieve state-of-the-art results': 0.07692307692307693, 'state-of-the-art results i
n': 0.07692307692307693, 'results in many': 0.07692307692307693, 'in many natural': 0.07
692307692307693, 'many natural language': 0.07692307692307693, 'natural language tasks':
0.07692307692307693, 'language tasks ,': 0.07692307692307693, 'tasks , for': 0.076923076
92307693, ', for example': 0.07692307692307693, 'for example in': 0.07692307692307693,
'example in language': 0.07692307692307693, 'in language modeling': 0.07692307692307693,
'language modeling ,': 0.07692307692307693, 'modeling , [': 0.07692307692307693, ', [
9': 0.07692307692307693, '[9]': 0.07692307692307693, '9] parsing': 0.0769230769230769
3, '] parsing ,': 0.07692307692307693, 'parsing , [': 0.07692307692307693, ', [10': 0.0
7692307692307693, '[10]': 0.07692307692307693, '10] [': 0.07692307692307693, '] [1
1': 0.07692307692307693, '[11]': 0.07692307692307693, '11] and': 0.07692307692307693,
'] and many': 0.07692307692307693, 'and many others': 0.07692307692307693, 'many others
.': 0.07692307692307693, 'others . this': 0.07692307692307693, '. this is': 0.0769230769
2307693, 'this is increasingly': 0.07692307692307693, 'is increasingly important': 0.076
92307692307693, 'increasingly important in': 0.07692307692307693, 'important in medicin
e': 0.07692307692307693, 'in medicine and': 0.07692307692307693, 'medicine and healthcar
e': 0.07692307692307693, 'and healthcare ,': 0.07692307692307693, 'healthcare , where':
0.07692307692307693, ', where nlp': 0.07692307692307693, 'where nlp is': 0.0769230769230
7693, 'nlp is being': 0.07692307692307693, 'is being used': 0.07692307692307693, 'being
used to': 0.07692307692307693, 'used to analyze': 0.07692307692307693, 'to analyze note
s': 0.07692307692307693, 'analyze notes and': 0.07692307692307693, 'notes and text': 0.0
7692307692307693, 'and text in': 0.07692307692307693, 'text in electronic': 0.0769230769
2307693, 'in electronic health': 0.07692307692307693, 'electronic health records': 0.076
92307692307693, 'health records that': 0.07692307692307693, 'records that would': 0.0769
2307692307693, 'that would otherwise': 0.07692307692307693, 'would otherwise be': 0.0769
2307692307693, 'otherwise be inaccessible': 0.07692307692307693, 'be inaccessible for':
0.07692307692307693, 'inaccessible for study': 0.07692307692307693, 'for study when': 0.
07692307692307693, 'study when seeking': 0.07692307692307693, 'when seeking to': 0.07692
307692307693, 'seeking to improve': 0.07692307692307693, 'to improve care': 0.0769230769
2307693, 'improve care .': 0.07692307692307693, 'care . [': 0.07692307692307693, '. [1
2': 0.07692307692307693, '[12]': 0.07692307692307693, '12] in': 0.07692307692307693,
'in the early': 0.07692307692307693, 'the early days': 0.07692307692307693, 'early days
,': 0.07692307692307693, 'days , many': 0.07692307692307693, ', many language-processin
g': 0.07692307692307693, 'many language-processing systems': 0.07692307692307693, 'langu
age-processing systems were': 0.07692307692307693, 'systems were designed': 0.0769230769
2307693, 'were designed by': 0.07692307692307693, 'designed by symbolic': 0.076923076923
07693, 'by symbolic methods': 0.07692307692307693, 'symbolic methods ,': 0.0769230769230
7693, 'methods , i.e.': 0.07692307692307693, ', i.e. ,': 0.07692307692307693, 'i.e. , th
e': 0.07692307692307693, ', the hand-coding': 0.07692307692307693, 'the hand-coding of':
0.07692307692307693, 'hand-coding of a': 0.07692307692307693, 'of a set': 0.076923076923
07693, 'set of rules': 0.07692307692307693, 'of rules ,': 0.07692307692307693, 'rules ,
coupled': 0.07692307692307693, ', coupled with': 0.07692307692307693, 'coupled with a':
0.07692307692307693, 'with a dictionary': 0.07692307692307693, 'a dictionary lookup': 0.
07692307692307693, 'dictionary lookup :': 0.07692307692307693, 'lookup : [': 0.076923076
92307693, ': [13': 0.07692307692307693, '[13]': 0.07692307692307693, '13] [': 0.0769

2307692307693, '] [14': 0.07692307692307693, '[14]': 0.07692307692307693, '14] such': 0.07692307692307693, '] such as': 0.07692307692307693, 'such as by': 0.07692307692307693, 'as by writing': 0.07692307692307693, 'by writing grammars': 0.07692307692307693, 'writing grammars or': 0.07692307692307693, 'grammars or devising': 0.07692307692307693, 'or devising heuristic': 0.07692307692307693, 'devising heuristic rules': 0.07692307692307693, 'heuristic rules for': 0.07692307692307693, 'rules for stemming': 0.07692307692307693, 'for stemming .': 0.07692307692307693, 'stemming . more': 0.07692307692307693, '. more recent': 0.07692307692307693, 'more recent systems': 0.07692307692307693, 'recent systems based': 0.07692307692307693, 'systems based on': 0.07692307692307693, 'based on machine-learning': 0.07692307692307693, 'on machine-learning algorithms': 0.07692307692307693, 'algorithms have many': 0.07692307692307693, 'have many advantages': 0.07692307692307693, 'many advantages over': 0.07692307692307693, 'advantages over hand-produced': 0.07692307692307693, 'over hand-produced rules': 0.07692307692307693, 'hand-produced rules :': 0.07692307692307693, 'rules : despite': 0.07692307692307693, ': despite the': 0.07692307692307693, 'despite the popularity': 0.07692307692307693, 'the popularity of': 0.07692307692307693, 'popularity of machine': 0.07692307692307693, 'machine learning in': 0.07692307692307693, 'learning in nlp': 0.07692307692307693, 'nlp research ,': 0.07692307692307693, 'research , symbolic': 0.07692307692307693, ', symbolic methods': 0.07692307692307693, 'symbolic methods are': 0.07692307692307693, 'methods are still': 0.07692307692307693, 'are still (': 0.07692307692307693, 'still (2020': 0.07692307692307693, '(2020)': 0.07692307692307693, '2020) commonly': 0.07692307692307693, ') commonly used': 0.07692307692307693, 'commonly used :': 0.07692307692307693, 'used : since': 0.07692307692307693, ': since the': 0.07692307692307693, 'since the so-called': 0.07692307692307693, 'the so-called ``': 0.07692307692307693, 'so-called `` statistical': 0.07692307692307693, '`` statistical revolution': 0.07692307692307693, "statistical revolution ''": 0.07692307692307693, "revolution '' [": 0.07692307692307693, "'' [15": 0.07692307692307693, '[15]': 0.07692307692307693, '15] [': 0.07692307692307693, '] [16': 0.07692307692307693, '[16]': 0.07692307692307693, '16] in': 0.07692307692307693, 'late 1980s and': 0.07692307692307693, '1980s and mid-1990s': 0.07692307692307693, 'and mid-1990s ,': 0.07692307692307693, 'mid-1990s , much': 0.07692307692307693, ', much natural': 0.07692307692307693, 'much natural language': 0.07692307692307693, 'language processing research': 0.07692307692307693, 'processing research has': 0.07692307692307693, 'research has relied': 0.07692307692307693, 'has relied heavily': 0.07692307692307693, 'relied heavily on': 0.07692307692307693, 'heavily on machine': 0.07692307692307693, 'on machine learning': 0.07692307692307693, 'learning . the': 0.07692307692307693, '. the machine-learning': 0.07692307692307693, 'the machine-learning paradigm': 0.07692307692307693, 'machine-learning paradigm calls': 0.07692307692307693, 'paradigm calls instead': 0.07692307692307693, 'calls instead for': 0.07692307692307693, 'instead for using': 0.07692307692307693, 'for using statistical': 0.07692307692307693, 'using statistical inference': 0.07692307692307693, 'statistical inference to': 0.07692307692307693, 'inference to automatically': 0.07692307692307693, 'to automatically learn': 0.07692307692307693, 'automatically learn such': 0.07692307692307693, 'learn such rules': 0.07692307692307693, 'such rules through': 0.07692307692307693, 'rules through the': 0.07692307692307693, 'through the analysis': 0.07692307692307693, 'the analysis of': 0.07692307692307693, 'analysis of large': 0.07692307692307693, 'of large corpora': 0.07692307692307693, 'large corpora (': 0.07692307692307693, 'corpora (the': 0.07692307692307693, '(the plural': 0.07692307692307693, 'the plural form': 0.07692307692307693, 'plural form of': 0.07692307692307693, 'form of corpus': 0.07692307692307693, 'of corpus ,': 0.07692307692307693, 'corpus , is': 0.07692307692307693, ', is a': 0.07692307692307693, 'is a set': 0.07692307692307693, 'set of documents': 0.07692307692307693, 'documents , possibly': 0.07692307692307693, ', possibly with': 0.07692307692307693, 'possibly with human': 0.07692307692307693, 'with human or': 0.07692307692307693, 'human or computer': 0.07692307692307693, 'or computer annotations': 0.07692307692307693, 'computer annotations)': 0.07692307692307693, 'annotations) of': 0.07692307692307693, ') of typical': 0.07692307692307693, 'of typical real-world': 0.07692307692307693, 'typical real-world examples': 0.07692307692307693, 'real-world examples .': 0.07692307692307693, 'examples . many': 0.07692307692307693, '. many different': 0.07692307692307693, 'many different classes': 0.07692307692307693, 'different classes of': 0.07692307692307693, 'classes of machine-learning': 0.07692307692307693, 'of machine-learning algorithms': 0.07692307692307693, 'algorithms have been': 0.07692307692307693, 'have been applied': 0.07692307692307693, 'been applied to': 0.07692307692307693, 'applied to natural-language-processing': 0.07692307692307693, 'to natural-language-processing tasks': 0.07692307692307693, 'natural-language-processing tasks .': 0.07692307692307693, 'tasks . these': 0.07692307692307693, '. these algorithms': 0.07692307692307693, 'these algorithms take': 0.07692307692307693, 'algorithms take as': 0.07692307692307693, 'take as input': 0.07692307692307693, 'as input a': 0.07692307692307693, 'input a large': 0.07692307

92307693, 'a large set': 0.07692307692307693, 'large set of': 0.07692307692307693, 'set
of ``': 0.07692307692307693, 'of `` features': 0.07692307692307693, "`` features ''": 0.
07692307692307693, "features '' that": 0.07692307692307693, "'' that are": 0.07692307692
307693, 'that are generated': 0.07692307692307693, 'are generated from': 0.0769230769230
7693, 'generated from the': 0.07692307692307693, 'from the input': 0.07692307692307693,
'data . increasingly': 0.07692307692307693, '. increasingly ,': 0.07692307692307693, 'in
creasingly , however': 0.07692307692307693, 'however , research': 0.07692307692307693,
'weights to each': 0.07692307692307693, 'to each input': 0.07692307692307693, 'each inpu
t feature': 0.07692307692307693, 'input feature (': 0.07692307692307693, 'feature (comp
lex-valued': 0.07692307692307693, '(complex-valued embeddings': 0.07692307692307693, 'c
omplex-valued embeddings ,': 0.07692307692307693, 'embeddings , [': 0.07692307692307693,
', [17': 0.07692307692307693, '[17]': 0.07692307692307693, '17] and': 0.076923076923
07693, '] and neural': 0.07692307692307693, 'and neural networks': 0.07692307692307693,
'neural networks in': 0.07692307692307693, 'networks in general': 0.07692307692307693,
'in general have': 0.07692307692307693, 'general have also': 0.07692307692307693, 'have
also been': 0.07692307692307693, 'also been proposed': 0.07692307692307693, 'been propos
ed ,': 0.07692307692307693, 'proposed , for': 0.07692307692307693, ', for e.g': 0.076923
07692307693, 'for e.g .': 0.07692307692307693, 'e.g . speech': 0.07692307692307693, '. s
peech [': 0.07692307692307693, 'speech [18': 0.07692307692307693, '[18]': 0.076923076
92307693, '18])': 0.07692307692307693, ']) .': 0.07692307692307693, ') . such': 0.076
92307692307693, 'such models have': 0.07692307692307693, 'models have the': 0.0769230769
2307693, 'have the advantage': 0.07692307692307693, 'the advantage that': 0.076923076923
07693, 'advantage that they': 0.07692307692307693, 'that they can': 0.07692307692307693,
'they can express': 0.07692307692307693, 'can express the': 0.07692307692307693, 'expres
s the relative': 0.07692307692307693, 'the relative certainty': 0.07692307692307693, 're
lative certainty of': 0.07692307692307693, 'certainty of many': 0.07692307692307693, 'of
many different': 0.07692307692307693, 'many different possible': 0.07692307692307693, 'd
ifferent possible answers': 0.07692307692307693, 'possible answers rather': 0.0769230769
2307693, 'answers rather than': 0.07692307692307693, 'rather than only': 0.0769230769230
7693, 'than only one': 0.07692307692307693, 'only one ,': 0.07692307692307693, 'one , pr
oducing': 0.07692307692307693, ', producing more': 0.07692307692307693, 'producing more
reliable': 0.07692307692307693, 'results when such': 0.07692307692307693, 'when such a':
0.07692307692307693, 'such a model': 0.07692307692307693, 'a model is': 0.07692307692307
693, 'model is included': 0.07692307692307693, 'is included as': 0.07692307692307693, 'i
ncluded as a': 0.07692307692307693, 'as a component': 0.07692307692307693, 'a component
of': 0.07692307692307693, 'component of a': 0.07692307692307693, 'of a larger': 0.076923
07692307693, 'larger system .': 0.07692307692307693, 'system . some': 0.0769230769230769
3, 'of the earliest-used': 0.07692307692307693, 'the earliest-used machine': 0.076923076
92307693, 'earliest-used machine learning': 0.07692307692307693, 'learning algorithms
,': 0.07692307692307693, 'algorithms , such': 0.07692307692307693, ', such as': 0.076923
07692307693, 'such as decision': 0.07692307692307693, 'as decision trees': 0.07692307692
307693, 'decision trees ,': 0.07692307692307693, 'trees , produced': 0.0769230769230769
3, ', produced systems': 0.07692307692307693, 'produced systems of': 0.0769230769230769
3, 'systems of hard': 0.07692307692307693, 'of hard if-then': 0.07692307692307693, 'hard
if-then rules': 0.07692307692307693, 'if-then rules similar': 0.07692307692307693, 'rule
s similar to': 0.07692307692307693, 'similar to existing': 0.07692307692307693, 'to exis
ting hand-written': 0.07692307692307693, 'existing hand-written rules': 0.07692307692307
693, 'rules . however': 0.07692307692307693, 'however , part-of-speech': 0.0769230769230
7693, 'part-of-speech tagging introduced': 0.07692307692307693, 'tagging introduced th
e': 0.07692307692307693, 'introduced the use': 0.07692307692307693, 'use of hidden': 0.0
7692307692307693, 'of hidden markov': 0.07692307692307693, 'hidden markov models': 0.076
92307692307693, 'markov models to': 0.07692307692307693, 'models to natural': 0.07692307
692307693, 'processing , and': 0.07692307692307693, ', and increasingly': 0.076923076923
07693, 'and increasingly ,': 0.07692307692307693, 'increasingly , research': 0.076923076
92307693, 'weights to the': 0.07692307692307693, 'to the features': 0.07692307692307693,
'the features making': 0.07692307692307693, 'features making up': 0.07692307692307693,
'making up the': 0.07692307692307693, 'up the input': 0.07692307692307693, '. the cach
e': 0.07692307692307693, 'the cache language': 0.07692307692307693, 'cache language mode
ls': 0.07692307692307693, 'language models upon': 0.07692307692307693, 'models upon whic
h': 0.07692307692307693, 'upon which many': 0.07692307692307693, 'which many speech': 0.
07692307692307693, 'many speech recognition': 0.07692307692307693, 'speech recognition s
ystems': 0.07692307692307693, 'recognition systems now': 0.07692307692307693, 'systems n
ow rely': 0.07692307692307693, 'now rely are': 0.07692307692307693, 'rely are examples':
0.07692307692307693, 'are examples of': 0.07692307692307693, 'examples of such': 0.07692
307692307693, 'of such statistical': 0.07692307692307693, 'such statistical models': 0.0

7692307692307693, 'statistical models .': 0.07692307692307693, 'models . such': 0.076923
07692307693, 'such models are': 0.07692307692307693, 'models are generally': 0.076923076
92307693, 'are generally more': 0.07692307692307693, 'generally more robust': 0.07692307
692307693, 'more robust when': 0.07692307692307693, 'robust when given': 0.0769230769230
7693, 'when given unfamiliar': 0.07692307692307693, 'given unfamiliar input': 0.07692307
692307693, 'unfamiliar input ,': 0.07692307692307693, 'input , especially': 0.0769230769
2307693, ', especially input': 0.07692307692307693, 'especially input that': 0.076923076
92307693, 'input that contains': 0.07692307692307693, 'that contains errors': 0.07692307
692307693, 'contains errors (': 0.07692307692307693, 'errors (as': 0.07692307692307693,
'(as is': 0.07692307692307693, 'as is very': 0.07692307692307693, 'is very common': 0.0
7692307692307693, 'very common for': 0.07692307692307693, 'common for real-world': 0.076
92307692307693, 'for real-world data': 0.07692307692307693, 'real-world data)': 0.07692
307692307693, 'data) ,': 0.07692307692307693, ') , and': 0.07692307692307693, ', and pr
oduce': 0.07692307692307693, 'and produce more': 0.07692307692307693, 'produce more reli
able': 0.07692307692307693, 'results when integrated': 0.07692307692307693, 'when integr
ated into': 0.07692307692307693, 'integrated into a': 0.07692307692307693, 'into a large
r': 0.07692307692307693, 'larger system comprising': 0.07692307692307693, 'system compri
sing multiple': 0.07692307692307693, 'comprising multiple subtasks': 0.0769230769230769
3, 'multiple subtasks .': 0.07692307692307693, 'subtasks . since': 0.07692307692307693,
'. since the': 0.07692307692307693, 'since the neural': 0.07692307692307693, 'the neural
turn': 0.07692307692307693, 'neural turn ,': 0.07692307692307693, 'turn , statistical':
0.07692307692307693, ', statistical methods': 0.07692307692307693, 'statistical methods
in': 0.07692307692307693, 'methods in nlp': 0.07692307692307693, 'nlp research have': 0.
07692307692307693, 'research have been': 0.07692307692307693, 'have been largely': 0.076
92307692307693, 'been largely replaced': 0.07692307692307693, 'largely replaced by': 0.0
7692307692307693, 'replaced by neural': 0.07692307692307693, 'by neural networks': 0.076
92307692307693, 'neural networks .': 0.07692307692307693, 'networks . however': 0.076923
07692307693, 'however , they': 0.07692307692307693, ', they continue': 0.076923076923076
93, 'they continue to': 0.07692307692307693, 'continue to be': 0.07692307692307693, 'to
be relevant': 0.07692307692307693, 'be relevant for': 0.07692307692307693, 'relevant for
contexts': 0.07692307692307693, 'for contexts in': 0.07692307692307693, 'contexts in whi
ch': 0.07692307692307693, 'in which statistical': 0.07692307692307693, 'which statistica
l interpretability': 0.07692307692307693, 'statistical interpretability and': 0.07692307
692307693, 'interpretability and transparency': 0.07692307692307693, 'and transparency i
s': 0.07692307692307693, 'transparency is required': 0.07692307692307693, 'is required
.': 0.07692307692307693, 'required . a': 0.07692307692307693, '. a major': 0.07692307692
307693, 'a major drawback': 0.07692307692307693, 'major drawback of': 0.0769230769230769
3, 'drawback of statistical': 0.07692307692307693, 'of statistical methods': 0.076923076
92307693, 'statistical methods is': 0.07692307692307693, 'methods is that': 0.0769230769
2307693, 'is that they': 0.07692307692307693, 'that they require': 0.07692307692307693,
'they require elaborate': 0.07692307692307693, 'require elaborate feature': 0.0769230769
2307693, 'elaborate feature engineering': 0.07692307692307693, 'feature engineering .':
0.07692307692307693, 'engineering . since': 0.07692307692307693, '. since 2015': 0.07692
307692307693, 'since 2015 ,': 0.07692307692307693, '2015 , [': 0.07692307692307693, ', [
19': 0.07692307692307693, '[19]': 0.07692307692307693, '19] the': 0.0769230769230769
3, '] the field': 0.07692307692307693, 'the field has': 0.07692307692307693, 'field has
thus': 0.07692307692307693, 'has thus largely': 0.07692307692307693, 'thus largely aband
oned': 0.07692307692307693, 'largely abandoned statistical': 0.07692307692307693, 'aband
oned statistical methods': 0.07692307692307693, 'statistical methods and': 0.07692307692
307693, 'methods and shifted': 0.07692307692307693, 'and shifted to': 0.0769230769230769
3, 'shifted to neural': 0.07692307692307693, 'to neural networks': 0.07692307692307693,
'neural networks for': 0.07692307692307693, 'networks for machine': 0.07692307692307693,
'for machine learning': 0.07692307692307693, 'learning . popular': 0.07692307692307693,
'. popular techniques': 0.07692307692307693, 'popular techniques include': 0.07692307692
307693, 'techniques include the': 0.07692307692307693, 'include the use': 0.076923076923
07693, 'use of word': 0.07692307692307693, 'of word embeddings': 0.07692307692307693, 'w
ord embeddings to': 0.07692307692307693, 'embeddings to capture': 0.07692307692307693,
'to capture semantic': 0.07692307692307693, 'capture semantic properties': 0.07692307692
307693, 'semantic properties of': 0.07692307692307693, 'properties of words': 0.07692307
692307693, 'of words ,': 0.07692307692307693, 'words , and': 0.07692307692307693, ', and
an': 0.07692307692307693, 'and an increase': 0.07692307692307693, 'an increase in': 0.07
692307692307693, 'increase in end-to-end': 0.07692307692307693, 'in end-to-end learnin
g': 0.07692307692307693, 'end-to-end learning of': 0.07692307692307693, 'learning of a':
0.07692307692307693, 'of a higher-level': 0.07692307692307693, 'a higher-level task': 0.
07692307692307693, 'higher-level task (': 0.07692307692307693, 'task (e.g.': 0.07692307

692307693, 'e.g. , question': 0.07692307692307693, ', question answering': 0.07692307692
307693, 'question answering)': 0.07692307692307693, 'answering) instead': 0.0769230769
2307693, ') instead of': 0.07692307692307693, 'instead of relying': 0.07692307692307693,
'of relying on': 0.07692307692307693, 'relying on a': 0.07692307692307693, 'on a pipelin
e': 0.07692307692307693, 'a pipeline of': 0.07692307692307693, 'pipeline of separate':
0.07692307692307693, 'of separate intermediate': 0.07692307692307693, 'separate intermed
iate tasks': 0.07692307692307693, 'intermediate tasks (': 0.07692307692307693, 'tasks (
e.g.': 0.07692307692307693, 'e.g. , part-of-speech': 0.07692307692307693, 'part-of-speec
h tagging and': 0.07692307692307693, 'tagging and dependency': 0.07692307692307693, 'and
dependency parsing': 0.07692307692307693, 'dependency parsing)': 0.07692307692307693,
'parsing) .': 0.07692307692307693, ') . in': 0.07692307692307693, '. in some': 0.076923
07692307693, 'in some areas': 0.07692307692307693, 'some areas ,': 0.07692307692307693,
'areas , this': 0.07692307692307693, ', this shift': 0.07692307692307693, 'this shift ha
s': 0.07692307692307693, 'shift has entailed': 0.07692307692307693, 'has entailed substa
ntial': 0.07692307692307693, 'entailed substantial changes': 0.07692307692307693, 'subst
antial changes in': 0.07692307692307693, 'changes in how': 0.07692307692307693, 'in how
nlp': 0.07692307692307693, 'how nlp systems': 0.07692307692307693, 'nlp systems are': 0.
07692307692307693, 'systems are designed': 0.07692307692307693, 'are designed ,': 0.0769
2307692307693, 'designed , such': 0.07692307692307693, ', such that': 0.0769230769230769
3, 'such that deep': 0.07692307692307693, 'that deep neural': 0.07692307692307693, 'deep
neural network-based': 0.07692307692307693, 'neural network-based approaches': 0.0769230
7692307693, 'network-based approaches may': 0.07692307692307693, 'approaches may be': 0.
07692307692307693, 'may be viewed': 0.07692307692307693, 'be viewed as': 0.0769230769230
7693, 'viewed as a': 0.07692307692307693, 'as a new': 0.07692307692307693, 'a new paradi
gm': 0.07692307692307693, 'new paradigm distinct': 0.07692307692307693, 'paradigm distin
ct from': 0.07692307692307693, 'distinct from statistical': 0.07692307692307693, 'from s
tatistical natural': 0.07692307692307693, 'statistical natural language': 0.076923076923
07693, 'processing . for': 0.07692307692307693, '. for instance': 0.07692307692307693,
'for instance ,': 0.07692307692307693, 'instance , the': 0.07692307692307693, ', the ter
m': 0.07692307692307693, 'the term neural': 0.07692307692307693, 'term neural machine':
0.07692307692307693, 'neural machine translation': 0.07692307692307693, 'translation (n
mt': 0.07692307692307693, '(nmt)': 0.07692307692307693, 'nmt) emphasizes': 0.07692307
692307693, ') emphasizes the': 0.07692307692307693, 'emphasizes the fact': 0.07692307692
307693, 'the fact that': 0.07692307692307693, 'fact that deep': 0.07692307692307693, 'th
at deep learning-based': 0.07692307692307693, 'deep learning-based approaches': 0.076923
07692307693, 'learning-based approaches to': 0.07692307692307693, 'approaches to machin
e': 0.07692307692307693, 'to machine translation': 0.07692307692307693, 'machine transla
tion directly': 0.07692307692307693, 'translation directly learn': 0.07692307692307693,
'directly learn sequence-to-sequence': 0.07692307692307693, 'learn sequence-to-sequence
transformations': 0.07692307692307693, 'sequence-to-sequence transformations ,': 0.07692
307692307693, 'transformations , obviating': 0.07692307692307693, ', obviating the': 0.0
7692307692307693, 'obviating the need': 0.07692307692307693, 'the need for': 0.076923076
92307693, 'need for intermediate': 0.07692307692307693, 'for intermediate steps': 0.0769
2307692307693, 'intermediate steps such': 0.07692307692307693, 'steps such as': 0.076923
07692307693, 'such as word': 0.07692307692307693, 'as word alignment': 0.076923076923076
93, 'word alignment and': 0.07692307692307693, 'alignment and language': 0.0769230769230
7693, 'and language modeling': 0.07692307692307693, 'language modeling that': 0.07692307
692307693, 'modeling that was': 0.07692307692307693, 'that was used': 0.0769230769230769
3, 'was used in': 0.07692307692307693, 'used in statistical': 0.07692307692307693, 'in s
tatistical machine': 0.07692307692307693, 'statistical machine translation': 0.076923076
92307693, 'translation (smt': 0.07692307692307693, '(smt)': 0.07692307692307693, 'smt
) .': 0.07692307692307693, ') . latest': 0.07692307692307693, '. latest works': 0.076923
07692307693, 'latest works tend': 0.07692307692307693, 'works tend to': 0.07692307692307
693, 'tend to use': 0.07692307692307693, 'to use non-technical': 0.07692307692307693, 'u
se non-technical structure': 0.07692307692307693, 'non-technical structure of': 0.076923
07692307693, 'structure of a': 0.07692307692307693, 'of a given': 0.07692307692307693,
'a given task': 0.07692307692307693, 'given task to': 0.07692307692307693, 'task to buil
d': 0.07692307692307693, 'to build proper': 0.07692307692307693, 'build proper neural':
0.07692307692307693, 'proper neural network': 0.07692307692307693, 'neural network .':
0.07692307692307693, 'network . [': 0.07692307692307693, '. [20': 0.07692307692307693,
'[20]': 0.07692307692307693, '20] the': 0.07692307692307693, '] the following': 0.076
92307692307693, 'the following is': 0.07692307692307693, 'following is a': 0.07692307692
307693, 'is a list': 0.07692307692307693, 'a list of': 0.07692307692307693, 'list of som
e': 0.07692307692307693, 'of some of': 0.07692307692307693, 'of the most': 0.07692307692
307693, 'the most commonly': 0.07692307692307693, 'most commonly researched': 0.07692307

692307693, 'commonly researched tasks': 0.07692307692307693, 'researched tasks in': 0.07692307692307693, 'tasks in natural': 0.07692307692307693, 'processing . some': 0.07692307692307693, 'some of these': 0.07692307692307693, 'of these tasks': 0.07692307692307693, 'these tasks have': 0.07692307692307693, 'tasks have direct': 0.07692307692307693, 'have direct real-world': 0.07692307692307693, 'direct real-world applications': 0.07692307692307693, 'real-world applications ,': 0.07692307692307693, 'applications , while': 0.07692307692307693, ', while others': 0.07692307692307693, 'while others more': 0.07692307692307693, 'others more commonly': 0.07692307692307693, 'more commonly serve': 0.07692307692307693, 'commonly serve as': 0.07692307692307693, 'serve as subtasks': 0.07692307692307693, 'as subtasks that': 0.07692307692307693, 'subtasks that are': 0.07692307692307693, 'that are used': 0.07692307692307693, 'are used to': 0.07692307692307693, 'used to aid': 0.07692307692307693, 'to aid in': 0.07692307692307693, 'aid in solving': 0.07692307692307693, 'in solving larger': 0.07692307692307693, 'solving larger tasks': 0.07692307692307693, 'larger tasks .': 0.07692307692307693, 'tasks . though': 0.07692307692307693, '. though natural': 0.07692307692307693, 'though natural language': 0.07692307692307693, 'language processing tasks': 0.07692307692307693, 'processing tasks are': 0.07692307692307693, 'tasks are closely': 0.07692307692307693, 'are closely intertwined': 0.07692307692307693, 'closely intertwined ,': 0.07692307692307693, 'intertwined , they': 0.07692307692307693, ', they can': 0.07692307692307693, 'they can be': 0.07692307692307693, 'can be subdivided': 0.07692307692307693, 'be subdivided into': 0.07692307692307693, 'subdivided into categories': 0.07692307692307693, 'into categories for': 0.07692307692307693, 'categories for convenience': 0.07692307692307693, 'for convenience .': 0.07692307692307693, 'convenience . a': 0.07692307692307693, '. a coarse': 0.07692307692307693, 'a coarse division': 0.07692307692307693, 'coarse division is': 0.07692307692307693, 'division is given': 0.07692307692307693, 'is given below': 0.07692307692307693, 'given below .': 0.07692307692307693, 'below . based': 0.07692307692307693, '. based on': 0.07692307692307693, 'based on long-standing': 0.07692307692307693, 'on long-standing trends': 0.07692307692307693, 'long-standing trends in': 0.07692307692307693, 'trends in the': 0.07692307692307693, 'in the field': 0.07692307692307693, 'the field ,': 0.07692307692307693, 'field , it': 0.07692307692307693, ', it is': 0.07692307692307693, 'it is possible': 0.07692307692307693, 'is possible to': 0.07692307692307693, 'possible to extrapolate': 0.07692307692307693, 'to extrapolate future': 0.07692307692307693, 'extrapolate future directions': 0.07692307692307693, 'future directions of': 0.07692307692307693, 'directions of nlp': 0.07692307692307693, 'of nlp .': 0.07692307692307693, 'nlp . as': 0.07692307692307693, '. as of': 0.07692307692307693, 'as of 2020': 0.07692307692307693, 'of 2020 ,': 0.07692307692307693, '2020 , three': 0.07692307692307693, ', three trends': 0.07692307692307693, 'three trends among': 0.07692307692307693, 'trends among the': 0.07692307692307693, 'among the topics': 0.07692307692307693, 'the topics of': 0.07692307692307693, 'topics of the': 0.07692307692307693, 'of the long-standing': 0.07692307692307693, 'the long-standing series': 0.07692307692307693, 'long-standing series of': 0.07692307692307693, 'series of conll': 0.07692307692307693, 'of conll shared': 0.07692307692307693, 'shared tasks can': 0.07692307692307693, 'tasks can be': 0.07692307692307693, 'can be observed': 0.07692307692307693, 'be observed :': 0.07692307692307693, 'observed : [': 0.07692307692307693, ': [36': 0.07692307692307693, '[36]': 0.07692307692307693, '36] most': 0.07692307692307693, '] most higher-level': 0.07692307692307693, 'most higher-level nlp': 0.07692307692307693, 'higher-level nlp applications': 0.07692307692307693, 'nlp applications involve': 0.07692307692307693, 'applications involve aspects': 0.07692307692307693, 'involve aspects that': 0.07692307692307693, 'aspects that emulate': 0.07692307692307693, 'that emulate intelligent': 0.07692307692307693, 'emulate intelligent behaviour': 0.07692307692307693, 'intelligent behaviour and': 0.07692307692307693, 'behaviour and apparent': 0.07692307692307693, 'and apparent comprehension': 0.07692307692307693, 'apparent comprehension of': 0.07692307692307693, 'comprehension of natural': 0.07692307692307693, 'natural language .': 0.07692307692307693, 'language . more': 0.07692307692307693, '. more broadly': 0.07692307692307693, 'more broadly speaking': 0.07692307692307693, 'broadly speaking ,': 0.07692307692307693, 'speaking , the': 0.07692307692307693, ', the technical': 0.07692307692307693, 'the technical operationalization': 0.07692307692307693, 'technical operationalization of': 0.07692307692307693, 'operationalization of increasingly': 0.07692307692307693, 'of increasingly advanced': 0.07692307692307693, 'increasingly advanced aspects': 0.07692307692307693, 'advanced aspects of': 0.07692307692307693, 'aspects of cognitive': 0.07692307692307693, 'of cognitive behaviour': 0.07692307692307693, 'cognitive behaviour represents': 0.07692307692307693, 'behaviour represents one': 0.07692307692307693, 'represents one of': 0.07692307692307693, 'one of the': 0.07692307692307693, 'of the developmental': 0.07692307692307693, 'the developmental trajectories': 0.07692307692307693, 'developmental trajectories of': 0.07692307692307693, 'trajectories of nlp': 0.07692307692307693, 'of nlp (': 0.07692307692307693, 'nlp (see': 0.07692307692307693, '(see tren

ds': 0.07692307692307693, 'see trends among': 0.07692307692307693, 'trends among conll': 0.07692307692307693, 'among conll shared': 0.07692307692307693, 'shared tasks above': 0.07692307692307693, 'tasks above)': 0.07692307692307693, 'above) .': 0.07692307692307693, ') . cognition': 0.07692307692307693, '. cognition refers': 0.07692307692307693, 'cognition refers to': 0.07692307692307693, 'refers to ``': 0.07692307692307693, 'to `` the': 0.07692307692307693, '`` the mental': 0.07692307692307693, 'the mental action': 0.07692307692307693, 'mental action or': 0.07692307692307693, 'action or process': 0.07692307692307693, 'or process of': 0.07692307692307693, 'process of acquiring': 0.07692307692307693, 'of acquiring knowledge': 0.07692307692307693, 'acquiring knowledge and': 0.07692307692307693, 'knowledge and understanding': 0.07692307692307693, 'and understanding through': 0.07692307692307693, 'understanding through thought': 0.07692307692307693, 'through thought ,': 0.07692307692307693, 'thought , experience': 0.07692307692307693, ', experience ,': 0.07692307692307693, 'experience , and': 0.07692307692307693, ', and the': 0.07692307692307693, 'and the senses': 0.07692307692307693, 'the senses .': 0.07692307692307693, 'senses . ``': 0.07692307692307693, '. `` [': 0.07692307692307693, '`` [37': 0.07692307692307693, '[37]': 0.07692307692307693, '37] cognitive': 0.07692307692307693, '] cognitive science': 0.07692307692307693, 'cognitive science is': 0.07692307692307693, 'science is the': 0.07692307692307693, 'is the interdisciplinary': 0.07692307692307693, 'the interdisciplinary ,': 0.07692307692307693, 'interdisciplinary , scientific': 0.07692307692307693, ', scientific study': 0.07692307692307693, 'scientific study of': 0.07692307692307693, 'study of the': 0.07692307692307693, 'of the mind': 0.07692307692307693, 'the mind and': 0.07692307692307693, 'mind and its': 0.07692307692307693, 'and its processes': 0.07692307692307693, 'its processes .': 0.07692307692307693, 'processes . [': 0.07692307692307693, '. [38': 0.07692307692307693, '[38]': 0.07692307692307693, '38] cognitive': 0.07692307692307693, '] cognitive linguistics': 0.07692307692307693, 'cognitive linguistics is': 0.07692307692307693, 'linguistics is an': 0.07692307692307693, 'is an interdisciplinary': 0.07692307692307693, 'an interdisciplinary branch': 0.07692307692307693, 'interdisciplinary branch of': 0.07692307692307693, 'branch of linguistics': 0.07692307692307693, 'linguistics , combining': 0.07692307692307693, ', combining knowledge': 0.07692307692307693, 'combining knowledge and': 0.07692307692307693, 'knowledge and research': 0.07692307692307693, 'and research from': 0.07692307692307693, 'research from both': 0.07692307692307693, 'from both psychology': 0.07692307692307693, 'both psychology and': 0.07692307692307693, 'psychology and linguistics': 0.07692307692307693, 'and linguistics .': 0.07692307692307693, 'linguistics . [': 0.07692307692307693, '. [39': 0.07692307692307693, '[39]': 0.07692307692307693, '39] especially': 0.07692307692307693, '] especially during': 0.07692307692307693, 'especially during the': 0.07692307692307693, 'during the age': 0.07692307692307693, 'the age of': 0.07692307692307693, 'age of symbolic': 0.07692307692307693, 'symbolic nlp ,': 0.07692307692307693, 'nlp , the': 0.07692307692307693, ', the area': 0.07692307692307693, 'the area of': 0.07692307692307693, 'area of computational': 0.07692307692307693, 'of computational linguistics': 0.07692307692307693, 'computational linguistics maintained': 0.07692307692307693, 'linguistics maintained strong': 0.07692307692307693, 'maintained strong ties': 0.07692307692307693, 'strong ties with': 0.07692307692307693, 'with cognitive studies': 0.07692307692307693, 'cognitive studies .': 0.07692307692307693, 'studies . as': 0.07692307692307693, '. as an': 0.07692307692307693, 'as an example': 0.07692307692307693, 'an example ,': 0.07692307692307693, 'example , george': 0.07692307692307693, ', george lakoff': 0.07692307692307693, 'george lakoff offers': 0.07692307692307693, 'lakoff offers a': 0.07692307692307693, 'offers a methodology': 0.07692307692307693, 'a methodology to': 0.07692307692307693, 'methodology to build': 0.07692307692307693, 'to build natural': 0.07692307692307693, 'build natural language': 0.07692307692307693, 'nlp) algorithms': 0.07692307692307693, ') algorithms through': 0.07692307692307693, 'algorithms through the': 0.07692307692307693, 'through the perspective': 0.07692307692307693, 'the perspective of': 0.07692307692307693, 'perspective of cognitive': 0.07692307692307693, 'of cognitive science': 0.07692307692307693, 'cognitive science ,': 0.07692307692307693, 'science , along': 0.07692307692307693, ', along with': 0.07692307692307693, 'along with the': 0.07692307692307693, 'with the findings': 0.07692307692307693, 'the findings of': 0.07692307692307693, 'findings of cognitive': 0.07692307692307693, 'of cognitive linguistics': 0.07692307692307693, 'cognitive linguistics ,': 0.07692307692307693, 'linguistics , [': 0.07692307692307693, ', [40': 0.07692307692307693, '[40]': 0.07692307692307693, '40] with': 0.07692307692307693, '] with two': 0.07692307692307693, 'with two defining': 0.07692307692307693, 'two defining aspects': 0.07692307692307693, 'defining aspects :': 0.07692307692307693, 'aspects : ties': 0.07692307692307693, ': ties with': 0.07692307692307693, 'with cognitive linguistics': 0.07692307692307693, 'cognitive linguistics are': 0.07692307692307693, 'linguistics are part': 0.07692307692307693, 'are part of': 0.07692307692307693, 'part of the': 0.07692307692307693, 'of the historical': 0.07692307692307693, 'the historical heritage': 0.07692

307692307693, 'historical heritage of': 0.07692307692307693, 'heritage of nlp': 0.076923
07692307693, 'of nlp ,': 0.07692307692307693, 'nlp , but': 0.07692307692307693, ', but t
hey': 0.07692307692307693, 'but they have': 0.07692307692307693, 'they have been': 0.076
92307692307693, 'have been less': 0.07692307692307693, 'been less frequently': 0.0769230
7692307693, 'less frequently addressed': 0.07692307692307693, 'frequently addressed sinc
e': 0.07692307692307693, 'addressed since the': 0.07692307692307693, 'since the statisti
cal': 0.07692307692307693, 'the statistical turn': 0.07692307692307693, 'statistical tur
n during': 0.07692307692307693, 'turn during the': 0.07692307692307693, 'during the 1990
s': 0.07692307692307693, 'the 1990s .': 0.07692307692307693, '1990s . nevertheless': 0.0
7692307692307693, '. nevertheless ,': 0.07692307692307693, 'nevertheless , approaches':
0.07692307692307693, ', approaches to': 0.07692307692307693, 'approaches to develop': 0.
07692307692307693, 'to develop cognitive': 0.07692307692307693, 'develop cognitive model
s': 0.07692307692307693, 'cognitive models towards': 0.07692307692307693, 'models toward
s technically': 0.07692307692307693, 'towards technically operationalizable': 0.07692307
692307693, 'technically operationalizable frameworks': 0.07692307692307693, 'operational
izable frameworks have': 0.07692307692307693, 'frameworks have been': 0.0769230769230769
3, 'have been pursued': 0.07692307692307693, 'been pursued in': 0.07692307692307693, 'pu
rsued in the': 0.07692307692307693, 'in the context': 0.07692307692307693, 'the context
of': 0.07692307692307693, 'context of various': 0.07692307692307693, 'of various framewo
rks': 0.07692307692307693, 'various frameworks ,': 0.07692307692307693, 'frameworks , e.
g.': 0.07692307692307693, 'e.g. , of': 0.07692307692307693, ', of cognitive': 0.07692307
692307693, 'of cognitive grammar': 0.07692307692307693, 'cognitive grammar ,': 0.0769230
7692307693, ', [42': 0.07692307692307693, '[42]': 0.07692307692307693, '42] function
al': 0.07692307692307693, '] functional grammar': 0.07692307692307693, 'functional gramm
ar ,': 0.07692307692307693, ', [43': 0.07692307692307693, '[43]': 0.0769230769230769
3, '43] construction': 0.07692307692307693, '] construction grammar': 0.076923076923076
93, 'construction grammar ,': 0.07692307692307693, ', [44': 0.07692307692307693, '[44
]': 0.07692307692307693, '44] computational': 0.07692307692307693, '] computational psy
cholinguistics': 0.07692307692307693, 'computational psycholinguistics and': 0.076923076
92307693, 'psycholinguistics and cognitive': 0.07692307692307693, 'and cognitive neurosc
ience': 0.07692307692307693, 'cognitive neuroscience (': 0.07692307692307693, 'neurosci
nce (e.g.': 0.07692307692307693, 'e.g. , act-r': 0.07692307692307693, ', act-r)': 0.07
692307692307693, 'act-r) ,': 0.07692307692307693, ') , however': 0.07692307692307693,
'however , with': 0.07692307692307693, ', with limited': 0.07692307692307693, 'with limi
ted uptake': 0.07692307692307693, 'limited uptake in': 0.07692307692307693, 'uptake in m
ainstream': 0.07692307692307693, 'in mainstream nlp': 0.07692307692307693, 'mainstream n
lp (': 0.07692307692307693, 'nlp (as': 0.07692307692307693, '(as measured': 0.07692307
692307693, 'as measured by': 0.07692307692307693, 'measured by presence': 0.076923076923
07693, 'by presence on': 0.07692307692307693, 'presence on major': 0.07692307692307693,
'on major conferences': 0.07692307692307693, 'major conferences [': 0.07692307692307693,
'conferences [45': 0.07692307692307693, '[45]': 0.07692307692307693, '45] of': 0.076
92307692307693, '] of the': 0.07692307692307693, 'of the acl': 0.07692307692307693, 'the
acl)': 0.07692307692307693, 'acl) .': 0.07692307692307693, ') . more': 0.0769230769230
7693, '. more recently': 0.07692307692307693, 'more recently ,': 0.07692307692307693, 'r
ecently , ideas': 0.07692307692307693, 'cognitive nlp have': 0.07692307692307693, 'nlp h
ave been': 0.07692307692307693, 'have been revived': 0.07692307692307693, 'been revived
as': 0.07692307692307693, 'revived as an': 0.07692307692307693, 'as an approach': 0.0769
2307692307693, 'an approach to': 0.07692307692307693, 'approach to achieve': 0.076923076
92307693, 'to achieve explainability': 0.07692307692307693, 'achieve explainability ,':
0.07692307692307693, 'explainability , e.g.': 0.07692307692307693, 'e.g. , under': 0.076
92307692307693, ', under the': 0.07692307692307693, 'under the notion': 0.07692307692307
693, 'the notion of': 0.07692307692307693, 'notion of ``': 0.07692307692307693, 'of `` c
ognitive': 0.07692307692307693, '`` cognitive ai': 0.07692307692307693, "cognitive ai
''": 0.07692307692307693, "ai '' .": 0.07692307692307693, "'' . [": 0.07692307692307693,
'. [46': 0.07692307692307693, '[46]': 0.07692307692307693, '46] likewise': 0.0769230
7692307693, '] likewise ,': 0.07692307692307693, 'likewise , ideas': 0.0769230769230769
3, 'cognitive nlp are': 0.07692307692307693, 'nlp are inherent': 0.07692307692307693, 'a
re inherent to': 0.07692307692307693, 'inherent to neural': 0.07692307692307693, 'to neu
ral models': 0.07692307692307693, 'neural models multimodal': 0.07692307692307693, 'mode
ls multimodal nlp': 0.07692307692307693, 'multimodal nlp (': 0.07692307692307693, 'nlp (
although': 0.07692307692307693, '(although rarely': 0.07692307692307693, 'although rare
ly made': 0.07692307692307693, 'rarely made explicit': 0.07692307692307693, 'made explic
it)': 0.07692307692307693, 'explicit) .': 0.07692307692307693, ') . [': 0.076923076923
07693, '. [47': 0.07692307692307693, '[47]': 0.07692307692307693, '47] media': 0.076
92307692307693, '] media related': 0.07692307692307693, 'media related to': 0.0769230769

```
2307693, 'related to natural': 0.07692307692307693, 'language processing at': 0.07692307
692307693, 'processing at wikimedia': 0.07692307692307693, 'at wikimedia commons': 0.076
92307692307693}
```

Summary:

In [48]:

```
sentences=sent_tokenize(text)
sentence_scores={}

for sentence in sentences:
    for word in fd_weighted:
        if word in fd_weighted:
            if sentence not in sentence_scores:
                sentence_scores[sentence] = fd_weighted[word]
            else:
                sentence_scores[sentence] += fd_weighted[word]


summary_sentences = heapq.nlargest(6, sentence_scores, key=sentence_scores.get)
summary = ' '.join(summary_sentences)

print(summary)
```

```
Natural language processing (NLP) is a subfield of linguistics, computer science, and ar
tificial intelligence concerned with the interactions between computers and human langua
ge, in particular how to program computers to process and analyze large amounts of natur
al language data. The goal is a computer capable of "understanding" the contents of docu
ments, including the contextual nuances of the language within them. The technology can
then accurately extract information and insights contained in the documents as well as c
ategorize and organize the documents themselves. Challenges in natural language processi
ng frequently involve speech recognition, natural language understanding, and natural la
nguage generation. Natural language processing has its roots in the 1950s. Already in 19
50, Alan Turing published an article titled "Computing Machinery and Intelligence" which
proposed what is now called the Turing test as a criterion of intelligence, a task that
involves the automated interpretation and generation of natural language, but at the tim
e not articulated as a problem separate from artificial intelligence.
```

3 - Comparisons:

In conclusion, introducing n-grams produces superior results. Compared to the first method, the final summary using n-grams is more grammatically precise, and descriptive, and thus achieves human-like summarization. In the below example, we compare both summaries limited to 5 sentences for each method, and it is evident that the first summary references a list without a proper description (sentence #4):

Regular 5 sentences: “Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. Up to the 1980s, most natural language processing systems were based on complex sets of hand-written rules. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing. [20] The following is a list of some of the most commonly researched tasks in natural language processing. Though natural language processing tasks are closely intertwined, they can be subdivided into categories for convenience.”

N-Grams 5 sentences:

“Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. Natural language processing has its roots in the 1950s.”