Text_Summarization

September 6, 2023

1 AIT526 Individual Lab 2

[199]:
```
from bs4 import BeautifulSoup
import requests
```

1.1 Task 1 - Text Summarization with Word Frequencies

[200]:
```
# 1.1 - Web Scraping Technique
def get_content_from_page(url):

    final_page_text = ""

    page_response = requests.get(url)
    soup_response = BeautifulSoup(page_response.content, "lxml")
    final_content = soup_response.find(id="content")
    pars = final_content.find_all("p")

    for p in pars:
        final_page_text += p.text

    return final_page_text.lower()

URL = 'https://en.wikipedia.org/wiki/Natural_language_processing'
content = get_content_from_page(URL)
content
```

[200]:
```
'natural language processing (nlp) is an interdisciplinary 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.\nchallenges in natural
language processing frequently involve speech recognition, natural-language
understanding, and natural-language generation.\nnatural 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, though at the time that was not articulated as a problem separate from artificial intelligence. the proposed test includes a task that involves the automated interpretation and generation of natural language.\nthe 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 confronts.\nup 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.[7]\nin the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing. that popularity was due partly to a flurry of results showing that such techniques[8][9] can achieve state-of-the-art results in many natural language tasks, e.g., in language modeling[10] and parsing.[11][12] this is increasingly important in medicine and healthcare, where nlp helps analyze notes and text in electronic health records that would otherwise be inaccessible for study when seeking to improve care.[13]\nin 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:[14][15] such as by writing grammars or devising heuristic rules for stemming.\nmore recent systems based on machine-learning algorithms have many advantages over hand-produced rules:\ndespite the popularity of machine learning in nlp research, symbolic methods are still (2020) commonly used:\nsince the so-called "statistical revolution"[16][17] 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.\nmany 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,[18] and neural networks in general have also been proposed, for e.g. speech[19]). 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.\nsome 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.\nsince 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.\na major drawback of statistical methods is that they require elaborate feature engineering. since 2015,[20] 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).\nthe 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.\nthough natural language processing tasks are closely intertwined, they can be subdivided into categories for convenience. a coarse division is given below.\nbased 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:[41]\nmost 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).\ncognition refers to "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses."[42] cognitive science is the interdisciplinary, scientific study of the mind and its processes.[43] cognitive linguistics is an interdisciplinary branch of linguistics, combining knowledge and research from both psychology and linguistics.[44] especially during the age of symbolic nlp, the area of computational linguistics maintained strong ties with cognitive studies.\nas 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,[45] with two defining

```
aspects:\nties 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,[47] functional grammar,[48]
construction grammar,[49] computational psycholinguistics and cognitive
neuroscience (e.g., act-r), however, with limited uptake in mainstream nlp (as
measured by presence on major conferences[50] 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".[51] likewise, ideas of cognitive nlp
are inherent to neural models multimodal nlp (although rarely made
explicit).[52]\n'
```

[201]:
```
# 1.2

from nltk.tokenize import sent_tokenize, word_tokenize, wordpunct_tokenize
from nltk.corpus import stopwords
from nltk.probability import FreqDist
import string

# Tokenization

# Sentence Tokenizer
my_sentences = sent_tokenize(content)
print('The no of sentences are: ', len(my_sentences), '\n')

# Word Tokenizer
my_words = word_tokenize(content)
print('Total Number of Words: ', len(my_words))
print(my_words[:20], '\n')

my_punct_tokenize = wordpunct_tokenize(content)
print('Total Number of Words: ', len(my_punct_tokenize))
print(my_punct_tokenize[:20], '\n')


# Removing punctuation
words_without_punctuation = []

words_without_punctuation = [''.join(eachcharac for eachcharac in eachword if␣
 ↪eachcharac not in string.punctuation ) for eachword in my_punct_tokenize]

final_words_without_punct = [eachw.lower() for eachw in␣
 ↪words_without_punctuation if eachw!='']

print("The no of words after removing punctuation are:␣
 ↪",len(final_words_without_punct))
```

```
print("The first 20 words are: \n", final_words_without_punct[:20], '\n')



# Removing stop words
stop_words = set(stopwords.words('english'))
without_stop_words = []

for i in final_words_without_punct:
    if i not in stop_words:
        without_stop_words.append(i)


print('No of words without any stopwords: ', len(without_stop_words), '\n')
print("The first 20 words are: \n", without_stop_words[:20], '\n')
```

```
The no of sentences are:  50

Total Number of Words:  1520
['natural', 'language', 'processing', '(', 'nlp', ')', 'is', 'an',
'interdisciplinary', 'subfield', 'of', 'linguistics', ',', 'computer',
'science', ',', 'and', 'artificial', 'intelligence', 'concerned']

Total Number of Words:  1594
['natural', 'language', 'processing', '(', 'nlp', ')', 'is', 'an',
'interdisciplinary', 'subfield', 'of', 'linguistics', ',', 'computer',
'science', ',', 'and', 'artificial', 'intelligence', 'concerned']

The no of words after removing punctuation are:  1340
The first 20 words are:
 ['natural', 'language', 'processing', 'nlp', 'is', 'an', 'interdisciplinary',
'subfield', 'of', 'linguistics', 'computer', 'science', 'and', 'artificial',
'intelligence', 'concerned', 'with', 'the', 'interactions', 'between']

No of words without any stopwords:  862

The first 20 words are:
 ['natural', 'language', 'processing', 'nlp', 'interdisciplinary', 'subfield',
'linguistics', 'computer', 'science', 'artificial', 'intelligence', 'concerned',
'interactions', 'computers', 'human', 'language', 'particular', 'program',
'computers', 'process']
```

[202]:
```
# 1.3

def calc_word_freq(input_words):
```

```
    # to calculate word frequency
    word_freq_list = FreqDist(input_words)

    # finding the max freq
    maximum_cnt = word_freq_list.most_common(1)[0][1]

    # weighted freqencies
    for each_word in word_freq_list.keys():
        word_freq_list[each_word] = word_freq_list[each_word] / maximum_cnt
    return word_freq_list

freq_of_words = calc_word_freq(without_stop_words)
print(freq_of_words.most_common(20))
```

```
[('language', 1.0), ('natural', 0.7142857142857143), ('nlp',
0.6071428571428571), ('processing', 0.5714285714285714), ('machine',
0.4642857142857143), ('learning', 0.4642857142857143), ('cognitive',
0.4642857142857143), ('statistical', 0.42857142857142855), ('e',
0.35714285714285715), ('tasks', 0.35714285714285715), ('linguistics',
0.32142857142857145), ('rules', 0.32142857142857145), ('g',
0.32142857142857145), ('models', 0.32142857142857145), ('neural',
0.2857142857142857), ('based', 0.25), ('systems', 0.21428571428571427),
('algorithms', 0.21428571428571427), ('methods', 0.21428571428571427), ('many',
0.21428571428571427)]
```

[203]:
```
# 1.4
import operator
sent_freq_dict = {}

# calculating the frequency for each sentence based on the individual word␣
 ↪frequency
for each_sentence in my_sentences:
    each_sentence = each_sentence.lower()
    word_list = wordpunct_tokenize(each_sentence)
    sum = 0
    for each_word in word_list:
        sum = sum + freq_of_words[each_word]
    #print(each_sentence, sum)
    sent_freq_dict[each_sentence] = sum

# sorting the dictionary in descending order
ranked_sent = dict(sorted(sent_freq_dict.items(), key=operator.itemgetter(1),␣
 ↪reverse=True))
ranked_sent
```

[203]:
```
{'more recent systems based on machine-learning algorithms have many advantages
over hand-produced rules: \ndespite the popularity of machine learning in nlp
```

research, symbolic methods are still (2020) commonly used:\nsince the so-called "statistical revolution"[16][17] in the late 1980s and mid-1990s, much natural language processing research has relied heavily on machine learning.': 9.607142857142856,
 '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,[45] with two defining aspects:\nties 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.': 7.499999999999997,
 'natural language processing (nlp) is an interdisciplinary 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.': 7.357142857142856,
 'challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.': 6.357142857142858,
 "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 confronts.": 6.071428571428571,
 '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.': 5.964285714285713,
 'nevertheless, approaches to develop cognitive models towards technically operationalizable frameworks have been pursued in the context of various frameworks, e.g., of cognitive grammar,[47] functional grammar,[48] construction grammar,[49] computational psycholinguistics and cognitive neuroscience (e.g., act-r), however, with limited uptake in mainstream nlp (as measured by presence on major conferences[50] of the acl).': 5.607142857142855,
 'starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing.': 5.5,
 '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).': 5.357142857142858,
 'that popularity was due partly to a flurry of results showing that such techniques[8][9] can achieve state-of-the-art results in many natural language tasks, e.g., in language modeling[10] and parsing.': 4.9642857142857135,
 '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.':
4.821428571428571,
 '[7]\nin the 2010s, representation learning and deep neural network-style
machine learning methods became widespread in natural language processing.':
4.571428571428571,
 '[13]\nin 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:[14][15] such as by writing grammars or devising heuristic
rules for stemming.': 4.178571428571428,
 '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,[18] and neural networks in general
have also been proposed, for e.g.': 4.142857142857143,
 'many different classes of machine-learning algorithms have been applied to
natural-language-processing tasks.': 4.142857142857142,
 'as of 2020, three trends among the topics of the long-standing series of conll
shared tasks can be observed:[41]\nmost higher-level nlp applications involve
aspects that emulate intelligent behaviour and apparent comprehension of natural
language.': 3.9999999999999996,
 '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).':
3.928571428571429,
 'up to the 1980s, most natural language processing systems were based on
complex sets of hand-written rules.': 3.5000000000000004,
 'transformational grammar), whose theoretical underpinnings discouraged the
sort of corpus linguistics that underlies the machine-learning approach to
language processing.': 3.3571428571428568,
 '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.': 3.2857142857142856,
 'the cache language models upon which many speech recognition systems now rely
are examples of such statistical models.': 2.928571428571429,
 'though natural language processing tasks are closely intertwined, they can be
subdivided into categories for convenience.': 2.892857142857142,
 'the following is a list of some of the most commonly researched tasks in
natural language processing.': 2.8571428571428568,
 '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.':
2.6428571428571432,
 '[51] likewise, ideas of cognitive nlp are inherent to neural models multimodal
nlp (although rarely made explicit).': 2.642857142857143,
 'more recently, ideas of cognitive nlp have been revived as an approach to
achieve explainability, e.g., under the notion of "cognitive ai".':
2.607142857142857,

'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).': 2.5000000000000004,
'since the neural turn, statistical methods in nlp research have been largely replaced by neural networks.': 2.464285714285714,
'since 2015,[20] the field has thus largely abandoned statistical methods and shifted to neural networks for machine learning.': 2.428571428571429,
'natural language processing has its roots in the 1950s.': 2.3571428571428568,
'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.': 2.214285714285715,
'the proposed test includes a task that involves the automated interpretation and generation of natural language.': 2.1785714285714284,
'[44] especially during the age of symbolic nlp, the area of computational linguistics maintained strong ties with cognitive studies.': 2.0,
'[43] cognitive linguistics is an interdisciplinary branch of linguistics, combining knowledge and research from both psychology and linguistics.': 1.9642857142857142,
'the goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.': 1.6785714285714286,
'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.': 1.6785714285714286,
"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.": 1.6428571428571428,
'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.': 1.5357142857142858,
'[11][12] this is increasingly important in medicine and healthcare, where nlp helps analyze notes and text in electronic health records that would otherwise be inaccessible for study when seeking to improve care.': 1.5000000000000009,
'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, though at the time that was not articulated as a problem separate from artificial intelligence.': 1.4642857142857142,
'based on long-standing trends in the field, it is possible to extrapolate future directions of nlp.': 1.357142857142857,
'these algorithms take as input a large set of "features" that are generated from the input data.': 1.1785714285714284,
'a major drawback of statistical methods is that they require elaborate feature engineering.': 0.9285714285714285,
'"[42] cognitive science is the interdisciplinary, scientific study of the mind

```
and its processes.': 0.8928571428571427,
 'however, they continue to be relevant for contexts in which statistical
interpretability and transparency is required.': 0.8214285714285713,
 'the technology can then accurately extract information and insights contained
in the documents as well as categorize and organize the documents themselves.':
0.6428571428571428,
 'cognition refers to "the mental action or process of acquiring knowledge and
understanding through thought, experience, and the senses.': 0.5714285714285714,
 'speech[19]).': 0.2142857142857143,
 'a coarse division is given below.': 0.17857142857142855,
 '[52]': 0.03571428571428571}
```

[204]:
```python
# 1.5 - Summary


def summary_based_on_sent_count(how_many_sent):
    return list(ranked_sent)[:how_many_sent]
    # Returning the first n sentences of the sorted sentence list
    # which were sorted in the descending order based on the freq


def summary_based_word_count(word_count):
    res = ""
    s_index = 0
    w_index = 0
    w_count = 0
    s_words = []
    spc = ""
    new_ln = ""
    sents = list(ranked_sent)

    while w_count < word_count: # reading a new sent and tokenizing it
        if len(s_words) == 0:
            s_words = wordpunct_tokenize(sents[s_index])
            w_index = 0
            spc = ""
            if len(res)>0:
                new_ln = "\n"
        if w_index<len(s_words):
            res += new_ln + spc + s_words[w_index]
            w_index += 1
            w_count += 1
            spc = " "
            new_ln = ""
        else:
            s_words = [] # reading another sentence now
            s_index += 1
```

```
    return res


def summary_based_on_percent(percent):

    w_count = 0
    tot_words = len(my_words)
    w_count = (percent/100)*tot_words
    return summary_based_word_count(w_count)

print('Summary based on sentence count: \n')
print(summary_based_on_sent_count(2))

print('\n\nSummary based on word count: \n')
print(summary_based_word_count(100))

print('\n\nSummary based on Percentage: \n')
print(summary_based_on_percent(15))
```

```
Summary based on sentence count:

['more recent systems based on machine-learning algorithms have many advantages
over hand-produced rules: \ndespite the popularity of machine learning in nlp
research, symbolic methods are still (2020) commonly used:\nsince the so-called
"statistical revolution"[16][17] in the late 1980s and mid-1990s, much natural
language processing research has relied heavily on machine learning.', '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,[45] with two defining aspects:\nties 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.']


Summary based on word count:

more recent 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 "[ 16 ][ 17 ] in the late 1980s and mid - 1990s
, much natural language processing research has relied heavily on machine
learning .
as an example , george lakoff offers a methodology to build natural language
processing ( nlp ) algorithms through the perspective of cognitive science ,
along with the


Summary based on Percentage:
```

```
more recent 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 "[ 16 ][ 17 ] in the late 1980s and mid - 1990s
, much natural language processing research has relied heavily on machine
learning .
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 ,[ 45 ] with two defining
aspects : 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 .
natural language processing ( nlp ) is an interdisciplinary 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 .
challenges in natural language processing frequently involve speech recognition
, natural - language understanding , and natural - language generation .
the premise of symbolic nlp is well - summarized by john searle ' s chinese room
experiment : given a collection
```

[ ]:

[ ]:

1.2 Task 2 - Text Summarization with N-grams

[205]:
```python
from nltk.util import ngrams
```

[206]:
```python
# 2.1 Generating n-grams from the text
def gen_ngrams(text, n):
    n_grams = ngrams(word_tokenize(text.lower()), n)
    return [' '.join(g) for g in n_grams]


grams_res_2 = gen_ngrams(content, 2)
print('bi-grams: \n', grams_res_2[0:20], '\n')

grams_res_3 = gen_ngrams(content, 3)
print('tri-grams: \n', grams_res_3[0:20], '\n')

grams_res_4 = gen_ngrams(content, 4)
print('4-grams: \n', grams_res_4[0:20], '\n')
```

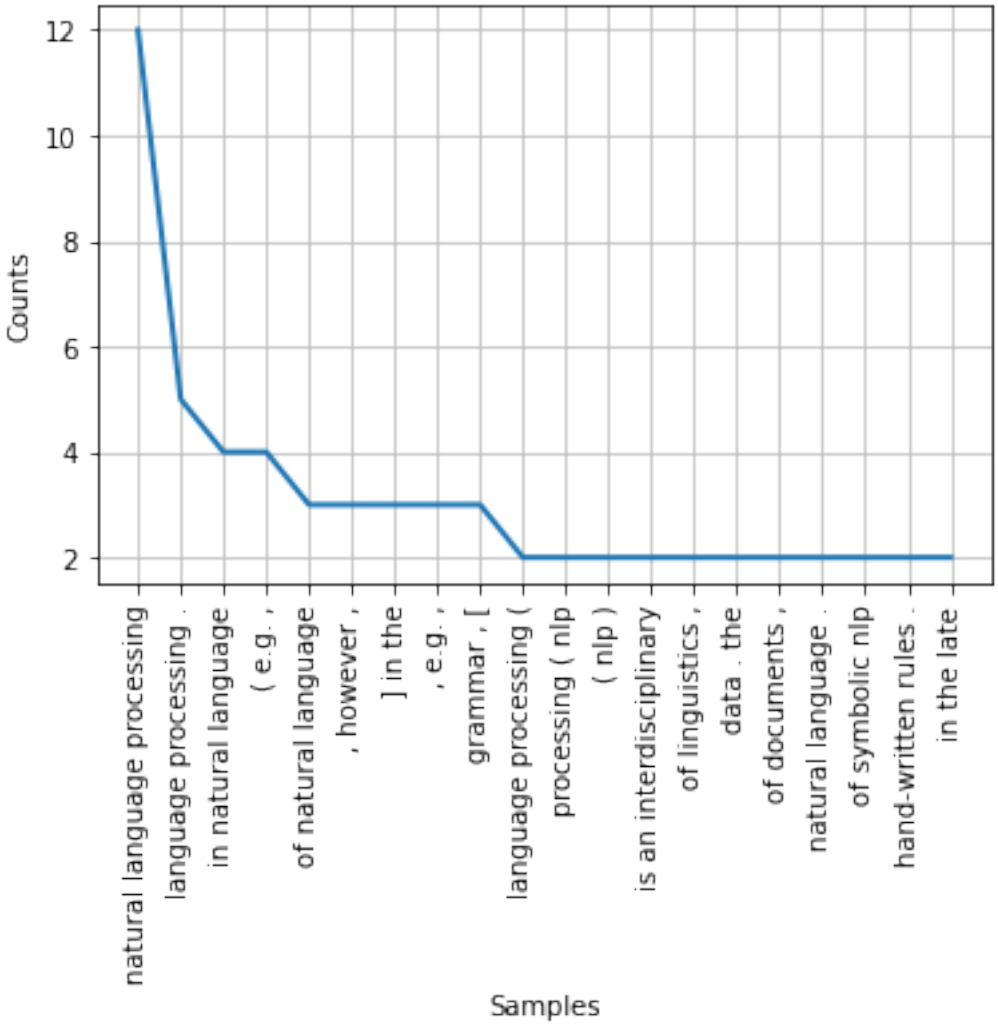
```
bi-grams:
 ['natural language', 'language processing', 'processing (', '( nlp', 'nlp )',
```

```
') is', 'is an', 'an interdisciplinary', 'interdisciplinary subfield', 'subfield
of', 'of linguistics', 'linguistics ,', ', computer', 'computer science',
'science ,', ', and', 'and artificial', 'artificial intelligence', 'intelligence
concerned', 'concerned with']

tri-grams:
 ['natural language processing', 'language processing (', 'processing ( nlp', '(
nlp )', 'nlp ) is', ') is an', 'is an interdisciplinary', 'an interdisciplinary
subfield', 'interdisciplinary subfield of', 'subfield of linguistics', 'of
linguistics ,', 'linguistics , computer', ', computer science', 'computer
science ,', 'science , and', ', and artificial', 'and artificial intelligence',
'artificial intelligence concerned', 'intelligence concerned with', 'concerned
with the']

4-grams:
 ['natural language processing (', 'language processing ( nlp', 'processing (
nlp )', '( nlp ) is', 'nlp ) is an', ') is an interdisciplinary', 'is an
interdisciplinary subfield', 'an interdisciplinary subfield of',
'interdisciplinary subfield of linguistics', 'subfield of linguistics ,', 'of
linguistics , computer', 'linguistics , computer science', ', computer science
,', 'computer science , and', 'science , and artificial', ', and artificial
intelligence', 'and artificial intelligence concerned', 'artificial intelligence
concerned with', 'intelligence concerned with the', 'concerned with the
interactions']
```

[207]:
```python
# 2.2

import matplotlib.pyplot as plt

# a) freq dist to calculate the n-gram frequencies
def display_n_grams(n):
    n_grams = gen_ngrams(content, n)
    freq_grams_res = FreqDist(n_grams)
    print(freq_grams_res)
    print(freq_grams_res.most_common(20))
    freq_grams_res.plot(20)
    plt.show()

display_n_grams(3)
```

```
<FreqDist with 1434 samples and 1518 outcomes>
[('natural language processing', 12), ('language processing .', 5), ('in natural
language', 4), ('( e.g. ,', 4), ('of natural language', 3), (', however ,', 3),
('] in the', 3), (', e.g. ,', 3), ('grammar , [', 3), ('language processing (',
2), ('processing ( nlp', 2), ('( nlp )', 2), ('is an interdisciplinary', 2),
('of linguistics ,', 2), ('data . the', 2), ('of documents ,', 2), ('natural
language .', 2), ('of symbolic nlp', 2), ('hand-written rules .', 2), ('in the
```

```
late', 2)]
```



[211]:
```
# 2.2 b

n_grams = 3

ngram_freq = FreqDist(gen_ngrams(content, n_grams))
sent_score = {}
maximum_cnt = ngram_freq.most_common(1)[0][1]
for each_word in ngram_freq.keys():
    ngram_freq[each_word] = ngram_freq[each_word] / maximum_cnt


print(ngrams_freq_res.items())
```

dict_items([('natural language processing', 1.0), ('language processing (', 0.16666666666666666), ('processing (nlp', 0.16666666666666666), ('(nlp)', 0.16666666666666666), ('nlp) is', 0.08333333333333333), (') is an', 0.08333333333333333), ('is an interdisciplinary', 0.16666666666666666), ('an interdisciplinary subfield', 0.08333333333333333), ('interdisciplinary subfield of', 0.08333333333333333), ('subfield of linguistics', 0.08333333333333333), ('of linguistics ,', 0.16666666666666666), ('linguistics , computer', 0.08333333333333333), (', computer science', 0.08333333333333333), ('computer science ,', 0.08333333333333333), ('science , and', 0.08333333333333333), (', and artificial', 0.08333333333333333), ('and artificial intelligence', 0.08333333333333333), ('artificial intelligence concerned', 0.08333333333333333), ('intelligence concerned with', 0.08333333333333333), ('concerned with the', 0.08333333333333333), ('with the interactions', 0.08333333333333333), ('the interactions between', 0.08333333333333333), ('interactions between computers', 0.08333333333333333), ('between computers and', 0.08333333333333333), ('computers and human', 0.08333333333333333), ('and human language', 0.08333333333333333), ('human language ,', 0.08333333333333333), ('language , in', 0.08333333333333333), (', in particular', 0.08333333333333333), ('in particular how', 0.08333333333333333), ('particular how to', 0.08333333333333333), ('how to program', 0.08333333333333333), ('to program computers', 0.08333333333333333), ('program computers to', 0.08333333333333333), ('computers to process', 0.08333333333333333), ('to process and', 0.08333333333333333), ('process and analyze', 0.08333333333333333), ('and analyze large', 0.08333333333333333), ('analyze large amounts', 0.08333333333333333), ('large amounts of', 0.08333333333333333), ('amounts of natural', 0.08333333333333333), ('of natural language', 0.25), ('natural language data', 0.08333333333333333), ('language data .', 0.08333333333333333), ('data . the', 0.16666666666666666), ('. the goal', 0.08333333333333333), ('the goal is', 0.08333333333333333), ('goal is a', 0.08333333333333333), ('is a computer', 0.08333333333333333), ('a computer capable', 0.08333333333333333), ('computer capable of', 0.08333333333333333), ('capable of ``', 0.08333333333333333), ('of `` understanding', 0.08333333333333333), ("`` understanding ''", 0.08333333333333333), ("understanding '' the", 0.08333333333333333), ("'' the contents", 0.08333333333333333), ('the contents of', 0.08333333333333333), ('contents of documents', 0.08333333333333333), ('of documents ,', 0.16666666666666666), ('documents , including', 0.08333333333333333), (', including the', 0.08333333333333333), ('including the contextual', 0.08333333333333333), ('the contextual nuances', 0.08333333333333333), ('contextual nuances of', 0.08333333333333333), ('nuances of the', 0.08333333333333333), ('of the language', 0.08333333333333333), ('the language within', 0.08333333333333333), ('language within them', 0.08333333333333333), ('within them .', 0.08333333333333333), ('them . the', 0.08333333333333333), ('. the technology', 0.08333333333333333), ('the technology can', 0.08333333333333333), ('technology can then', 0.08333333333333333), ('can then accurately', 0.08333333333333333), ('then accurately extract', 0.08333333333333333), ('accurately extract information', 0.08333333333333333), ('extract information and', 0.08333333333333333), ('information and insights', 0.08333333333333333), ('and

insights contained', 0.08333333333333333), ('insights contained in', 0.08333333333333333), ('contained in the', 0.08333333333333333), ('in the documents', 0.08333333333333333), ('the documents as', 0.08333333333333333), ('documents as well', 0.08333333333333333), ('as well as', 0.08333333333333333), ('well as categorize', 0.08333333333333333), ('as categorize and', 0.08333333333333333), ('categorize and organize', 0.08333333333333333), ('and organize the', 0.08333333333333333), ('organize the documents', 0.08333333333333333), ('the documents themselves', 0.08333333333333333), ('documents themselves .', 0.08333333333333333), ('themselves . challenges', 0.08333333333333333), ('. challenges in', 0.08333333333333333), ('challenges in natural', 0.08333333333333333), ('in natural language', 0.3333333333333333), ('language processing frequently', 0.08333333333333333), ('processing frequently involve', 0.08333333333333333), ('frequently involve speech', 0.08333333333333333), ('involve speech recognition', 0.08333333333333333), ('speech recognition ,', 0.08333333333333333), ('recognition , natural-language', 0.08333333333333333), (', natural-language understanding', 0.08333333333333333), ('natural-language understanding ,', 0.08333333333333333), ('understanding , and', 0.08333333333333333), (', and natural-language', 0.08333333333333333), ('and natural-language generation', 0.08333333333333333), ('natural-language generation .', 0.08333333333333333), ('generation . natural', 0.08333333333333333), ('. natural language', 0.08333333333333333), ('language processing has', 0.08333333333333333), ('processing has its', 0.08333333333333333), ('has its roots', 0.08333333333333333), ('its roots in', 0.08333333333333333), ('roots in the', 0.08333333333333333), ('in the 1950s', 0.08333333333333333), ('the 1950s .', 0.08333333333333333), ('1950s . already', 0.08333333333333333), ('. already in', 0.08333333333333333), ('already in 1950', 0.08333333333333333), ('in 1950 ,', 0.08333333333333333), ('1950 , alan', 0.08333333333333333), (', alan turing', 0.08333333333333333), ('alan turing published', 0.08333333333333333), ('turing published an', 0.08333333333333333), ('published an article', 0.08333333333333333), ('an article titled', 0.08333333333333333), ('article titled ``', 0.08333333333333333), ('titled `` computing', 0.08333333333333333), ('`` computing machinery', 0.08333333333333333), ('computing machinery and', 0.08333333333333333), ('machinery and intelligence', 0.08333333333333333), ("and intelligence ''", 0.08333333333333333), ("intelligence '' which", 0.08333333333333333), ("'' which proposed", 0.08333333333333333), ('which proposed what', 0.08333333333333333), ('proposed what is', 0.08333333333333333), ('what is now', 0.08333333333333333), ('is now called', 0.08333333333333333), ('now called the', 0.08333333333333333), ('called the turing', 0.08333333333333333), ('the turing test', 0.08333333333333333), ('turing test as', 0.08333333333333333), ('test as a', 0.08333333333333333), ('as a criterion', 0.08333333333333333), ('a criterion of', 0.08333333333333333), ('criterion of intelligence', 0.08333333333333333), ('of intelligence ,', 0.08333333333333333), ('intelligence , though', 0.08333333333333333), (', though at', 0.08333333333333333), ('though at the', 0.08333333333333333), ('at the time', 0.08333333333333333), ('the time that', 0.08333333333333333), ('time that was', 0.08333333333333333), ('that was not', 0.08333333333333333), ('was not articulated', 0.08333333333333333), ('not articulated as', 0.08333333333333333), ('articulated as a',

0.08333333333333333), ('as a problem', 0.08333333333333333), ('a problem
separate', 0.08333333333333333), ('problem separate from', 0.08333333333333333),
('separate from artificial', 0.08333333333333333), ('from artificial
intelligence', 0.08333333333333333), ('artificial intelligence .',
0.08333333333333333), ('intelligence . the', 0.08333333333333333), ('. the
proposed', 0.08333333333333333), ('the proposed test', 0.08333333333333333),
('proposed test includes', 0.08333333333333333), ('test includes a',
0.08333333333333333), ('includes a task', 0.08333333333333333), ('a task that',
0.08333333333333333), ('task that involves', 0.08333333333333333), ('that
involves the', 0.08333333333333333), ('involves the automated',
0.08333333333333333), ('the automated interpretation', 0.08333333333333333),
('automated interpretation and', 0.08333333333333333), ('interpretation and
generation', 0.08333333333333333), ('and generation of', 0.08333333333333333),
('generation of natural', 0.08333333333333333), ('natural language .',
0.16666666666666666), ('language . the', 0.08333333333333333), ('. the premise',
0.08333333333333333), ('the premise of', 0.08333333333333333), ('premise of
symbolic', 0.08333333333333333), ('of symbolic nlp', 0.16666666666666666),
('symbolic nlp is', 0.08333333333333333), ('nlp is well-summarized',
0.08333333333333333), ('is well-summarized by', 0.08333333333333333), ('well-
summarized by john', 0.08333333333333333), ('by john searle',
0.08333333333333333), ("john searle 's", 0.08333333333333333), ("searle 's
chinese", 0.08333333333333333), ("'s chinese room", 0.08333333333333333),
('chinese room experiment', 0.08333333333333333), ('room experiment :',
0.08333333333333333), ('experiment : given', 0.08333333333333333), (': given a',
0.08333333333333333), ('given a collection', 0.08333333333333333), ('a
collection of', 0.08333333333333333), ('collection of rules',
0.08333333333333333), ('of rules (', 0.08333333333333333), ('rules (e.g.',
0.08333333333333333), ('(e.g. ,', 0.3333333333333333), ('e.g. , a',
0.08333333333333333), (', a chinese', 0.08333333333333333), ('a chinese
phrasebook', 0.08333333333333333), ('chinese phrasebook ,',
0.08333333333333333), ('phrasebook , with', 0.08333333333333333), (', with
questions', 0.08333333333333333), ('with questions and', 0.08333333333333333),
('questions and matching', 0.08333333333333333), ('and matching answers',
0.08333333333333333), ('matching answers)', 0.08333333333333333), ('answers)
,', 0.08333333333333333), (') , the', 0.08333333333333333), (', the computer',
0.08333333333333333), ('the computer emulates', 0.08333333333333333), ('computer
emulates natural', 0.08333333333333333), ('emulates natural language',
0.08333333333333333), ('natural language understanding', 0.08333333333333333),
('language understanding (', 0.08333333333333333), ('understanding (or',
0.08333333333333333), ('(or other', 0.08333333333333333), ('or other nlp',
0.08333333333333333), ('other nlp tasks', 0.08333333333333333), ('nlp tasks)',
0.08333333333333333), ('tasks) by', 0.08333333333333333), (') by applying',
0.08333333333333333), ('by applying those', 0.08333333333333333), ('applying
those rules', 0.08333333333333333), ('those rules to', 0.08333333333333333),
('rules to the', 0.08333333333333333), ('to the data', 0.08333333333333333),
('the data it', 0.08333333333333333), ('data it confronts',
0.08333333333333333), ('it confronts .', 0.08333333333333333), ('confronts .
up', 0.08333333333333333), ('. up to', 0.08333333333333333), ('up to the',

0.08333333333333333), ('to the 1980s', 0.08333333333333333), ('the 1980s ,',
0.08333333333333333), ('1980s , most', 0.08333333333333333), (', most natural',
0.08333333333333333), ('most natural language', 0.08333333333333333), ('language
processing systems', 0.08333333333333333), ('processing systems were',
0.08333333333333333), ('systems were based', 0.08333333333333333), ('were based
on', 0.08333333333333333), ('based on complex', 0.08333333333333333), ('on
complex sets', 0.08333333333333333), ('complex sets of', 0.08333333333333333),
('sets of hand-written', 0.08333333333333333), ('of hand-written rules',
0.08333333333333333), ('hand-written rules .', 0.16666666666666666), ('rules .
starting', 0.08333333333333333), ('. starting in', 0.08333333333333333),
('starting in the', 0.08333333333333333), ('in the late', 0.16666666666666666),
('the late 1980s', 0.16666666666666666), ('late 1980s ,', 0.08333333333333333),
('1980s , however', 0.08333333333333333), (', however ,', 0.25), ('however ,
there', 0.08333333333333333), (', there was', 0.08333333333333333), ('there was
a', 0.08333333333333333), ('was a revolution', 0.08333333333333333), ('a
revolution in', 0.08333333333333333), ('revolution in natural',
0.08333333333333333), ('language processing with', 0.08333333333333333),
('processing with the', 0.08333333333333333), ('with the introduction',
0.08333333333333333), ('the introduction of', 0.08333333333333333),
('introduction of machine', 0.08333333333333333), ('of machine learning',
0.16666666666666666), ('machine learning algorithms', 0.16666666666666666),
('learning algorithms for', 0.08333333333333333), ('algorithms for language',
0.08333333333333333), ('for language processing', 0.08333333333333333),
('language processing .', 0.4166666666666667), ('processing . this',
0.08333333333333333), ('. this was', 0.08333333333333333), ('this was due',
0.08333333333333333), ('was due to', 0.08333333333333333), ('due to both',
0.08333333333333333), ('to both the', 0.08333333333333333), ('both the steady',
0.08333333333333333), ('the steady increase', 0.08333333333333333), ('steady
increase in', 0.08333333333333333), ('increase in computational',
0.08333333333333333), ('in computational power', 0.08333333333333333),
('computational power (', 0.08333333333333333), ('power (see',
0.08333333333333333), ('(see moore', 0.08333333333333333), ("see moore 's",
0.08333333333333333), ("moore 's law", 0.08333333333333333), ("'s law)",
0.08333333333333333), ('law) and', 0.08333333333333333), (') and the',
0.08333333333333333), ('and the gradual', 0.08333333333333333), ('the gradual
lessening', 0.08333333333333333), ('gradual lessening of', 0.08333333333333333),
('lessening of the', 0.08333333333333333), ('of the dominance',
0.08333333333333333), ('the dominance of', 0.08333333333333333), ('dominance of
chomskyan', 0.08333333333333333), ('of chomskyan theories',
0.08333333333333333), ('chomskyan theories of', 0.08333333333333333), ('theories
of linguistics', 0.08333333333333333), ('of linguistics (',
0.08333333333333333), ('linguistics (e.g', 0.08333333333333333), ('(e.g .',
0.08333333333333333), ('e.g . transformational', 0.08333333333333333), ('.
transformational grammar', 0.08333333333333333), ('transformational grammar)',
0.08333333333333333), ('grammar) ,', 0.08333333333333333), (') , whose',
0.08333333333333333), (', whose theoretical', 0.08333333333333333), ('whose
theoretical underpinnings', 0.08333333333333333), ('theoretical underpinnings
discouraged', 0.08333333333333333), ('underpinnings discouraged the',

0.08333333333333333), ('discouraged the sort', 0.08333333333333333), ('the sort
of', 0.08333333333333333), ('sort of corpus', 0.08333333333333333), ('of corpus
linguistics', 0.08333333333333333), ('corpus linguistics that',
0.08333333333333333), ('linguistics that underlies', 0.08333333333333333),
('that underlies the', 0.08333333333333333), ('underlies the machine-learning',
0.08333333333333333), ('the machine-learning approach', 0.08333333333333333),
('machine-learning approach to', 0.08333333333333333), ('approach to language',
0.08333333333333333), ('to language processing', 0.08333333333333333),
('processing . [', 0.08333333333333333), ('. [7', 0.08333333333333333), ('[7
]', 0.08333333333333333), ('7] in', 0.08333333333333333), ('] in the', 0.25),
('in the 2010s', 0.08333333333333333), ('the 2010s ,', 0.08333333333333333),
('2010s , representation', 0.08333333333333333), (', representation learning',
0.08333333333333333), ('representation learning and', 0.08333333333333333),
('learning and deep', 0.08333333333333333), ('and deep neural',
0.08333333333333333), ('deep neural network-style', 0.08333333333333333),
('neural network-style machine', 0.08333333333333333), ('network-style machine
learning', 0.08333333333333333), ('machine learning methods',
0.08333333333333333), ('learning methods became', 0.08333333333333333),
('methods became widespread', 0.08333333333333333), ('became widespread in',
0.08333333333333333), ('widespread in natural', 0.08333333333333333),
('processing . that', 0.08333333333333333), ('. that popularity',
0.08333333333333333), ('that popularity was', 0.08333333333333333), ('popularity
was due', 0.08333333333333333), ('was due partly', 0.08333333333333333), ('due
partly to', 0.08333333333333333), ('partly to a', 0.08333333333333333), ('to a
flurry', 0.08333333333333333), ('a flurry of', 0.08333333333333333), ('flurry of
results', 0.08333333333333333), ('of results showing', 0.08333333333333333),
('results showing that', 0.08333333333333333), ('showing that such',
0.08333333333333333), ('that such techniques', 0.08333333333333333), ('such
techniques [', 0.08333333333333333), ('techniques [8', 0.08333333333333333),
('[8]', 0.08333333333333333), ('8] [', 0.08333333333333333), ('] [9',
0.08333333333333333), ('[9]', 0.08333333333333333), ('9] can',
0.08333333333333333), ('] can achieve', 0.08333333333333333), ('can achieve
state-of-the-art', 0.08333333333333333), ('achieve state-of-the-art results',
0.08333333333333333), ('state-of-the-art results in', 0.08333333333333333),
('results in many', 0.08333333333333333), ('in many natural',
0.08333333333333333), ('many natural language', 0.08333333333333333), ('natural
language tasks', 0.08333333333333333), ('language tasks ,',
0.08333333333333333), ('tasks , e.g.', 0.08333333333333333), (', e.g. ,', 0.25),
('e.g. , in', 0.08333333333333333), (', in language', 0.08333333333333333), ('in
language modeling', 0.08333333333333333), ('language modeling [',
0.08333333333333333), ('modeling [10', 0.08333333333333333), ('[10]',
0.08333333333333333), ('10] and', 0.08333333333333333), ('] and parsing',
0.08333333333333333), ('and parsing .', 0.08333333333333333), ('parsing . [',
0.08333333333333333), ('. [11', 0.08333333333333333), ('[11]',
0.08333333333333333), ('11] [', 0.08333333333333333), ('] [12',
0.08333333333333333), ('[12]', 0.08333333333333333), ('12] this',
0.08333333333333333), ('] this is', 0.08333333333333333), ('this is
increasingly', 0.08333333333333333), ('is increasingly important',

0.08333333333333333), ('increasingly important in', 0.08333333333333333), ('important in medicine', 0.08333333333333333), ('in medicine and', 0.08333333333333333), ('medicine and healthcare', 0.08333333333333333), ('and healthcare ,', 0.08333333333333333), ('healthcare , where', 0.08333333333333333), (', where nlp', 0.08333333333333333), ('where nlp helps', 0.08333333333333333), ('nlp helps analyze', 0.08333333333333333), ('helps analyze notes', 0.08333333333333333), ('analyze notes and', 0.08333333333333333), ('notes and text', 0.08333333333333333), ('and text in', 0.08333333333333333), ('text in electronic', 0.08333333333333333), ('in electronic health', 0.08333333333333333), ('electronic health records', 0.08333333333333333), ('health records that', 0.08333333333333333), ('records that would', 0.08333333333333333), ('that would otherwise', 0.08333333333333333), ('would otherwise be', 0.08333333333333333), ('otherwise be inaccessible', 0.08333333333333333), ('be inaccessible for', 0.08333333333333333), ('inaccessible for study', 0.08333333333333333), ('for study when', 0.08333333333333333), ('study when seeking', 0.08333333333333333), ('when seeking to', 0.08333333333333333), ('seeking to improve', 0.08333333333333333), ('to improve care', 0.08333333333333333), ('improve care .', 0.08333333333333333), ('care . [', 0.08333333333333333), ('. [13', 0.08333333333333333), ('[13]', 0.08333333333333333), ('13] in', 0.08333333333333333), ('in the early', 0.08333333333333333), ('the early days', 0.08333333333333333), ('early days ,', 0.08333333333333333), ('days , many', 0.08333333333333333), (', many language-processing', 0.08333333333333333), ('many language-processing systems', 0.08333333333333333), ('language-processing systems were', 0.08333333333333333), ('systems were designed', 0.08333333333333333), ('were designed by', 0.08333333333333333), ('designed by symbolic', 0.08333333333333333), ('by symbolic methods', 0.08333333333333333), ('symbolic methods ,', 0.08333333333333333), ('methods , i.e.', 0.08333333333333333), (', i.e. ,', 0.08333333333333333), ('i.e. , the', 0.08333333333333333), (', the hand-coding', 0.08333333333333333), ('the hand-coding of', 0.08333333333333333), ('hand-coding of a', 0.08333333333333333), ('of a set', 0.08333333333333333), ('a set of', 0.16666666666666666), ('set of rules', 0.08333333333333333), ('of rules ,', 0.08333333333333333), ('rules , coupled', 0.08333333333333333), (', coupled with', 0.08333333333333333), ('coupled with a', 0.08333333333333333), ('with a dictionary', 0.08333333333333333), ('a dictionary lookup', 0.08333333333333333), ('dictionary lookup :', 0.08333333333333333), ('lookup : [', 0.08333333333333333), (': [14', 0.08333333333333333), ('[14]', 0.08333333333333333), ('14] [', 0.08333333333333333), ('] [15', 0.08333333333333333), ('[15]', 0.08333333333333333), ('15] such', 0.08333333333333333), ('] such as', 0.08333333333333333), ('such as by', 0.08333333333333333), ('as by writing', 0.08333333333333333), ('by writing grammars', 0.08333333333333333), ('writing grammars or', 0.08333333333333333), ('grammars or devising', 0.08333333333333333), ('or devising heuristic', 0.08333333333333333), ('devising heuristic rules', 0.08333333333333333), ('heuristic rules for', 0.08333333333333333), ('rules for stemming', 0.08333333333333333), ('for stemming .', 0.08333333333333333), ('stemming . more', 0.08333333333333333), ('. more recent', 0.08333333333333333), ('more recent systems',

0.08333333333333333), ('recent systems based', 0.08333333333333333), ('systems based on', 0.08333333333333333), ('based on machine-learning', 0.08333333333333333), ('on machine-learning algorithms', 0.08333333333333333), ('machine-learning algorithms have', 0.16666666666666666), ('algorithms have many', 0.08333333333333333), ('have many advantages', 0.08333333333333333), ('many advantages over', 0.08333333333333333), ('advantages over hand-produced', 0.08333333333333333), ('over hand-produced rules', 0.08333333333333333), ('hand-produced rules :', 0.08333333333333333), ('rules : despite', 0.08333333333333333), (': despite the', 0.08333333333333333), ('despite the popularity', 0.08333333333333333), ('the popularity of', 0.08333333333333333), ('popularity of machine', 0.08333333333333333), ('machine learning in', 0.08333333333333333), ('learning in nlp', 0.08333333333333333), ('in nlp research', 0.16666666666666666), ('nlp research ,', 0.08333333333333333), ('research , symbolic', 0.08333333333333333), (', symbolic methods', 0.08333333333333333), ('symbolic methods are', 0.08333333333333333), ('methods are still', 0.08333333333333333), ('are still (', 0.08333333333333333), ('still (2020', 0.08333333333333333), ('(2020)', 0.08333333333333333), ('2020) commonly', 0.08333333333333333), (') commonly used', 0.08333333333333333), ('commonly used :', 0.08333333333333333), ('used : since', 0.08333333333333333), (': since the', 0.08333333333333333), ('since the so-called', 0.08333333333333333), ('the so-called ``', 0.08333333333333333), ('so-called `` statistical', 0.08333333333333333), ('`` statistical revolution', 0.08333333333333333), ("statistical revolution ''", 0.08333333333333333), ("revolution '' [", 0.08333333333333333), ("'' [16", 0.08333333333333333), ('[16]', 0.08333333333333333), ('16] [', 0.08333333333333333), ('] [17', 0.08333333333333333), ('[17]', 0.08333333333333333), ('17] in', 0.08333333333333333), ('late 1980s and', 0.08333333333333333), ('1980s and mid-1990s', 0.08333333333333333), ('and mid-1990s ,', 0.08333333333333333), ('mid-1990s , much', 0.08333333333333333), (', much natural', 0.08333333333333333), ('much natural language', 0.08333333333333333), ('language processing research', 0.08333333333333333), ('processing research has', 0.08333333333333333), ('research has relied', 0.08333333333333333), ('has relied heavily', 0.08333333333333333), ('relied heavily on', 0.08333333333333333), ('heavily on machine', 0.08333333333333333), ('on machine learning', 0.08333333333333333), ('machine learning .', 0.16666666666666666), ('learning . the', 0.08333333333333333), ('. the machine-learning', 0.08333333333333333), ('the machine-learning paradigm', 0.08333333333333333), ('machine-learning paradigm calls', 0.08333333333333333), ('paradigm calls instead', 0.08333333333333333), ('calls instead for', 0.08333333333333333), ('instead for using', 0.08333333333333333), ('for using statistical', 0.08333333333333333), ('using statistical inference', 0.08333333333333333), ('statistical inference to', 0.08333333333333333), ('inference to automatically', 0.08333333333333333), ('to automatically learn', 0.08333333333333333), ('automatically learn such', 0.08333333333333333), ('learn such rules', 0.08333333333333333), ('such rules through', 0.08333333333333333), ('rules through the', 0.08333333333333333), ('through the analysis', 0.08333333333333333), ('the analysis of', 0.08333333333333333), ('analysis of large', 0.08333333333333333), ('of large corpora', 0.08333333333333333), ('large corpora (', 0.08333333333333333),

('corpora (the', 0.08333333333333333), ('(the plural', 0.08333333333333333), ('the plural form', 0.08333333333333333), ('plural form of', 0.08333333333333333), ('form of corpus', 0.08333333333333333), ('of corpus ,', 0.08333333333333333), ('corpus , is', 0.08333333333333333), (', is a', 0.08333333333333333), ('is a set', 0.08333333333333333), ('set of documents', 0.08333333333333333), ('documents , possibly', 0.08333333333333333), (', possibly with', 0.08333333333333333), ('possibly with human', 0.08333333333333333), ('with human or', 0.08333333333333333), ('human or computer', 0.08333333333333333), ('or computer annotations', 0.08333333333333333), ('computer annotations)', 0.08333333333333333), ('annotations) of', 0.08333333333333333), (') of typical', 0.08333333333333333), ('of typical real-world', 0.08333333333333333), ('typical real-world examples', 0.08333333333333333), ('real-world examples .', 0.08333333333333333), ('examples . many', 0.08333333333333333), ('. many different', 0.08333333333333333), ('many different classes', 0.08333333333333333), ('different classes of', 0.08333333333333333), ('classes of machine-learning', 0.08333333333333333), ('of machine-learning algorithms', 0.08333333333333333), ('algorithms have been', 0.08333333333333333), ('have been applied', 0.08333333333333333), ('been applied to', 0.08333333333333333), ('applied to natural-language-processing', 0.08333333333333333), ('to natural-language-processing tasks', 0.08333333333333333), ('natural-language-processing tasks .', 0.08333333333333333), ('tasks . these', 0.08333333333333333), ('. these algorithms', 0.08333333333333333), ('these algorithms take', 0.08333333333333333), ('algorithms take as', 0.08333333333333333), ('take as input', 0.08333333333333333), ('as input a', 0.08333333333333333), ('input a large', 0.08333333333333333), ('a large set', 0.08333333333333333), ('large set of', 0.08333333333333333), ('set of ``', 0.08333333333333333), ('of `` features', 0.08333333333333333), ("`` features ''", 0.08333333333333333), ("features '' that", 0.08333333333333333), ("'' that are", 0.08333333333333333), ('that are generated', 0.08333333333333333), ('are generated from', 0.08333333333333333), ('generated from the', 0.08333333333333333), ('from the input', 0.08333333333333333), ('the input data', 0.16666666666666666), ('input data .', 0.16666666666666666), ('data . increasingly', 0.08333333333333333), ('. increasingly ,', 0.08333333333333333), ('increasingly , however', 0.08333333333333333), ('however , research', 0.08333333333333333), (', research has', 0.16666666666666666), ('research has focused', 0.16666666666666666), ('has focused on', 0.16666666666666666), ('focused on statistical', 0.16666666666666666), ('on statistical models', 0.16666666666666666), ('statistical models ,', 0.16666666666666666), ('models , which', 0.16666666666666666), (', which make', 0.16666666666666666), ('which make soft', 0.16666666666666666), ('make soft ,', 0.16666666666666666), ('soft , probabilistic', 0.16666666666666666), (', probabilistic decisions', 0.16666666666666666), ('probabilistic decisions based', 0.16666666666666666), ('decisions based on', 0.16666666666666666), ('based on attaching', 0.16666666666666666), ('on attaching real-valued', 0.16666666666666666), ('attaching real-valued weights', 0.16666666666666666), ('real-valued weights to', 0.16666666666666666), ('weights to each', 0.08333333333333333), ('to each input', 0.08333333333333333), ('each input feature', 0.08333333333333333),

('input feature (', 0.08333333333333333), ('feature (complex-valued', 0.08333333333333333), ('(complex-valued embeddings', 0.08333333333333333), ('complex-valued embeddings ,', 0.08333333333333333), ('embeddings , [', 0.08333333333333333), (', [18', 0.08333333333333333), ('[18]', 0.08333333333333333), ('18] and', 0.08333333333333333), ('] and neural', 0.08333333333333333), ('and neural networks', 0.08333333333333333), ('neural networks in', 0.08333333333333333), ('networks in general', 0.08333333333333333), ('in general have', 0.08333333333333333), ('general have also', 0.08333333333333333), ('have also been', 0.08333333333333333), ('also been proposed', 0.08333333333333333), ('been proposed ,', 0.08333333333333333), ('proposed , for', 0.08333333333333333), (', for e.g', 0.08333333333333333), ('for e.g .', 0.08333333333333333), ('e.g . speech', 0.08333333333333333), ('. speech [', 0.08333333333333333), ('speech [19', 0.08333333333333333), ('[19]', 0.08333333333333333), ('19])', 0.08333333333333333), (']) .', 0.08333333333333333), (') . such', 0.08333333333333333), ('. such models', 0.16666666666666666), ('such models have', 0.08333333333333333), ('models have the', 0.08333333333333333), ('have the advantage', 0.08333333333333333), ('the advantage that', 0.08333333333333333), ('advantage that they', 0.08333333333333333), ('that they can', 0.08333333333333333), ('they can express', 0.08333333333333333), ('can express the', 0.08333333333333333), ('express the relative', 0.08333333333333333), ('the relative certainty', 0.08333333333333333), ('relative certainty of', 0.08333333333333333), ('certainty of many', 0.08333333333333333), ('of many different', 0.08333333333333333), ('many different possible', 0.08333333333333333), ('different possible answers', 0.08333333333333333), ('possible answers rather', 0.08333333333333333), ('answers rather than', 0.08333333333333333), ('rather than only', 0.08333333333333333), ('than only one', 0.08333333333333333), ('only one ,', 0.08333333333333333), ('one , producing', 0.08333333333333333), (', producing more', 0.08333333333333333), ('producing more reliable', 0.08333333333333333), ('more reliable results', 0.16666666666666666), ('reliable results when', 0.16666666666666666), ('results when such', 0.08333333333333333), ('when such a', 0.08333333333333333), ('such a model', 0.08333333333333333), ('a model is', 0.08333333333333333), ('model is included', 0.08333333333333333), ('is included as', 0.08333333333333333), ('included as a', 0.08333333333333333), ('as a component', 0.08333333333333333), ('a component of', 0.08333333333333333), ('component of a', 0.08333333333333333), ('of a larger', 0.08333333333333333), ('a larger system', 0.16666666666666666), ('larger system .', 0.08333333333333333), ('system . some', 0.08333333333333333), ('. some of', 0.16666666666666666), ('some of the', 0.16666666666666666), ('of the earliest-used', 0.08333333333333333), ('the earliest-used machine', 0.08333333333333333), ('earliest-used machine learning', 0.08333333333333333), ('learning algorithms ,', 0.08333333333333333), ('algorithms , such', 0.08333333333333333), (', such as', 0.08333333333333333), ('such as decision', 0.08333333333333333), ('as decision trees', 0.08333333333333333), ('decision trees ,', 0.08333333333333333), ('trees , produced', 0.08333333333333333), (', produced systems', 0.08333333333333333), ('produced systems of', 0.08333333333333333), ('systems of hard', 0.08333333333333333), ('of hard if-then', 0.08333333333333333), ('hard if-then rules', 0.08333333333333333), ('if-then

rules similar', 0.08333333333333333), ('rules similar to', 0.08333333333333333), ('similar to existing', 0.08333333333333333), ('to existing hand-written', 0.08333333333333333), ('existing hand-written rules', 0.08333333333333333), ('rules . however', 0.08333333333333333), ('. however ,', 0.16666666666666666), ('however , part-of-speech', 0.08333333333333333), (', part-of-speech tagging', 0.16666666666666666), ('part-of-speech tagging introduced', 0.08333333333333333), ('tagging introduced the', 0.08333333333333333), ('introduced the use', 0.08333333333333333), ('the use of', 0.16666666666666666), ('use of hidden', 0.08333333333333333), ('of hidden markov', 0.08333333333333333), ('hidden markov models', 0.08333333333333333), ('markov models to', 0.08333333333333333), ('models to natural', 0.08333333333333333), ('to natural language', 0.08333333333333333), ('language processing ,', 0.08333333333333333), ('processing , and', 0.08333333333333333), (', and increasingly', 0.08333333333333333), ('and increasingly ,', 0.08333333333333333), ('increasingly , research', 0.08333333333333333), ('weights to the', 0.08333333333333333), ('to the features', 0.08333333333333333), ('the features making', 0.08333333333333333), ('features making up', 0.08333333333333333), ('making up the', 0.08333333333333333), ('up the input', 0.08333333333333333), ('. the cache', 0.08333333333333333), ('the cache language', 0.08333333333333333), ('cache language models', 0.08333333333333333), ('language models upon', 0.08333333333333333), ('models upon which', 0.08333333333333333), ('upon which many', 0.08333333333333333), ('which many speech', 0.08333333333333333), ('many speech recognition', 0.08333333333333333), ('speech recognition systems', 0.08333333333333333), ('recognition systems now', 0.08333333333333333), ('systems now rely', 0.08333333333333333), ('now rely are', 0.08333333333333333), ('rely are examples', 0.08333333333333333), ('are examples of', 0.08333333333333333), ('examples of such', 0.08333333333333333), ('of such statistical', 0.08333333333333333), ('such statistical models', 0.08333333333333333), ('statistical models .', 0.08333333333333333), ('models . such', 0.08333333333333333), ('such models are', 0.08333333333333333), ('models are generally', 0.08333333333333333), ('are generally more', 0.08333333333333333), ('generally more robust', 0.08333333333333333), ('more robust when', 0.08333333333333333), ('robust when given', 0.08333333333333333), ('when given unfamiliar', 0.08333333333333333), ('given unfamiliar input', 0.08333333333333333), ('unfamiliar input ,', 0.08333333333333333), ('input , especially', 0.08333333333333333), (', especially input', 0.08333333333333333), ('especially input that', 0.08333333333333333), ('input that contains', 0.08333333333333333), ('that contains errors', 0.08333333333333333), ('contains errors (', 0.08333333333333333), ('errors (as', 0.08333333333333333), ('(as is', 0.08333333333333333), ('as is very', 0.08333333333333333), ('is very common', 0.08333333333333333), ('very common for', 0.08333333333333333), ('common for real-world', 0.08333333333333333), ('for real-world data', 0.08333333333333333), ('real-world data)', 0.08333333333333333), ('data) ,', 0.08333333333333333), (') , and', 0.08333333333333333), (', and produce', 0.08333333333333333), ('and produce more', 0.08333333333333333), ('produce more reliable', 0.08333333333333333), ('results when integrated', 0.08333333333333333), ('when integrated into', 0.08333333333333333),

('integrated into a', 0.08333333333333333), ('into a larger',
0.08333333333333333), ('larger system comprising', 0.08333333333333333),
('system comprising multiple', 0.08333333333333333), ('comprising multiple
subtasks', 0.08333333333333333), ('multiple subtasks .', 0.08333333333333333),
('subtasks . since', 0.08333333333333333), ('. since the', 0.08333333333333333),
('since the neural', 0.08333333333333333), ('the neural turn',
0.08333333333333333), ('neural turn ,', 0.08333333333333333), ('turn ,
statistical', 0.08333333333333333), (', statistical methods',
0.08333333333333333), ('statistical methods in', 0.08333333333333333), ('methods
in nlp', 0.08333333333333333), ('nlp research have', 0.08333333333333333),
('research have been', 0.08333333333333333), ('have been largely',
0.08333333333333333), ('been largely replaced', 0.08333333333333333), ('largely
replaced by', 0.08333333333333333), ('replaced by neural', 0.08333333333333333),
('by neural networks', 0.08333333333333333), ('neural networks .',
0.08333333333333333), ('networks . however', 0.08333333333333333), ('however ,
they', 0.08333333333333333), (', they continue', 0.08333333333333333), ('they
continue to', 0.08333333333333333), ('continue to be', 0.08333333333333333),
('to be relevant', 0.08333333333333333), ('be relevant for',
0.08333333333333333), ('relevant for contexts', 0.08333333333333333), ('for
contexts in', 0.08333333333333333), ('contexts in which', 0.08333333333333333),
('in which statistical', 0.08333333333333333), ('which statistical
interpretability', 0.08333333333333333), ('statistical interpretability and',
0.08333333333333333), ('interpretability and transparency',
0.08333333333333333), ('and transparency is', 0.08333333333333333),
('transparency is required', 0.08333333333333333), ('is required .',
0.08333333333333333), ('required . a', 0.08333333333333333), ('. a major',
0.08333333333333333), ('a major drawback', 0.08333333333333333), ('major
drawback of', 0.08333333333333333), ('drawback of statistical',
0.08333333333333333), ('of statistical methods', 0.08333333333333333),
('statistical methods is', 0.08333333333333333), ('methods is that',
0.08333333333333333), ('is that they', 0.08333333333333333), ('that they
require', 0.08333333333333333), ('they require elaborate', 0.08333333333333333),
('require elaborate feature', 0.08333333333333333), ('elaborate feature
engineering', 0.08333333333333333), ('feature engineering .',
0.08333333333333333), ('engineering . since', 0.08333333333333333), ('. since
2015', 0.08333333333333333), ('since 2015 ,', 0.08333333333333333), ('2015 , [',
0.08333333333333333), (', [20', 0.08333333333333333), ('[20]',
0.08333333333333333), ('20] the', 0.08333333333333333), ('] the field',
0.08333333333333333), ('the field has', 0.08333333333333333), ('field has thus',
0.08333333333333333), ('has thus largely', 0.08333333333333333), ('thus largely
abandoned', 0.08333333333333333), ('largely abandoned statistical',
0.08333333333333333), ('abandoned statistical methods', 0.08333333333333333),
('statistical methods and', 0.08333333333333333), ('methods and shifted',
0.08333333333333333), ('and shifted to', 0.08333333333333333), ('shifted to
neural', 0.08333333333333333), ('to neural networks', 0.08333333333333333),
('neural networks for', 0.08333333333333333), ('networks for machine',
0.08333333333333333), ('for machine learning', 0.08333333333333333), ('learning
. popular', 0.08333333333333333), ('. popular techniques', 0.08333333333333333),

('popular techniques include', 0.08333333333333333), ('techniques include the',
0.08333333333333333), ('include the use', 0.08333333333333333), ('use of word',
0.08333333333333333), ('of word embeddings', 0.08333333333333333), ('word
embeddings to', 0.08333333333333333), ('embeddings to capture',
0.08333333333333333), ('to capture semantic', 0.08333333333333333), ('capture
semantic properties', 0.08333333333333333), ('semantic properties of',
0.08333333333333333), ('properties of words', 0.08333333333333333), ('of words
,', 0.08333333333333333), ('words , and', 0.08333333333333333), (', and an',
0.08333333333333333), ('and an increase', 0.08333333333333333), ('an increase
in', 0.08333333333333333), ('increase in end-to-end', 0.08333333333333333), ('in
end-to-end learning', 0.08333333333333333), ('end-to-end learning of',
0.08333333333333333), ('learning of a', 0.08333333333333333), ('of a higher-
level', 0.08333333333333333), ('a higher-level task', 0.08333333333333333),
('higher-level task (', 0.08333333333333333), ('task (e.g.',
0.08333333333333333), ('e.g. , question', 0.08333333333333333), (', question
answering', 0.08333333333333333), ('question answering)', 0.08333333333333333),
('answering) instead', 0.08333333333333333), (') instead of',
0.08333333333333333), ('instead of relying', 0.08333333333333333), ('of relying
on', 0.08333333333333333), ('relying on a', 0.08333333333333333), ('on a
pipeline', 0.08333333333333333), ('a pipeline of', 0.08333333333333333),
('pipeline of separate', 0.08333333333333333), ('of separate intermediate',
0.08333333333333333), ('separate intermediate tasks', 0.08333333333333333),
('intermediate tasks (', 0.08333333333333333), ('tasks (e.g.',
0.08333333333333333), ('e.g. , part-of-speech', 0.08333333333333333), ('part-of-
speech tagging and', 0.08333333333333333), ('tagging and dependency',
0.08333333333333333), ('and dependency parsing', 0.08333333333333333),
('dependency parsing)', 0.08333333333333333), ('parsing) .',
0.08333333333333333), (') . in', 0.08333333333333333), ('. in some',
0.08333333333333333), ('in some areas', 0.08333333333333333), ('some areas ,',
0.08333333333333333), ('areas , this', 0.08333333333333333), (', this shift',
0.08333333333333333), ('this shift has', 0.08333333333333333), ('shift has
entailed', 0.08333333333333333), ('has entailed substantial',
0.08333333333333333), ('entailed substantial changes', 0.08333333333333333),
('substantial changes in', 0.08333333333333333), ('changes in how',
0.08333333333333333), ('in how nlp', 0.08333333333333333), ('how nlp systems',
0.08333333333333333), ('nlp systems are', 0.08333333333333333), ('systems are
designed', 0.08333333333333333), ('are designed ,', 0.08333333333333333),
('designed , such', 0.08333333333333333), (', such that', 0.08333333333333333),
('such that deep', 0.08333333333333333), ('that deep neural',
0.08333333333333333), ('deep neural network-based', 0.08333333333333333),
('neural network-based approaches', 0.08333333333333333), ('network-based
approaches may', 0.08333333333333333), ('approaches may be',
0.08333333333333333), ('may be viewed', 0.08333333333333333), ('be viewed as',
0.08333333333333333), ('viewed as a', 0.08333333333333333), ('as a new',
0.08333333333333333), ('a new paradigm', 0.08333333333333333), ('new paradigm
distinct', 0.08333333333333333), ('paradigm distinct from',
0.08333333333333333), ('distinct from statistical', 0.08333333333333333), ('from
statistical natural', 0.08333333333333333), ('statistical natural language',

0.08333333333333333), ('processing . for', 0.08333333333333333), ('. for instance', 0.08333333333333333), ('for instance ,', 0.08333333333333333), ('instance , the', 0.08333333333333333), (', the term', 0.08333333333333333), ('the term neural', 0.08333333333333333), ('term neural machine', 0.08333333333333333), ('neural machine translation', 0.08333333333333333), ('machine translation (', 0.16666666666666666), ('translation (nmt', 0.08333333333333333), ('(nmt)', 0.08333333333333333), ('nmt) emphasizes', 0.08333333333333333), (') emphasizes the', 0.08333333333333333), ('emphasizes the fact', 0.08333333333333333), ('the fact that', 0.08333333333333333), ('fact that deep', 0.08333333333333333), ('that deep learning-based', 0.08333333333333333), ('deep learning-based approaches', 0.08333333333333333), ('learning-based approaches to', 0.08333333333333333), ('approaches to machine', 0.08333333333333333), ('to machine translation', 0.08333333333333333), ('machine translation directly', 0.08333333333333333), ('translation directly learn', 0.08333333333333333), ('directly learn sequence-to-sequence', 0.08333333333333333), ('learn sequence-to-sequence transformations', 0.08333333333333333), ('sequence-to-sequence transformations ,', 0.08333333333333333), ('transformations , obviating', 0.08333333333333333), (', obviating the', 0.08333333333333333), ('obviating the need', 0.08333333333333333), ('the need for', 0.08333333333333333), ('need for intermediate', 0.08333333333333333), ('for intermediate steps', 0.08333333333333333), ('intermediate steps such', 0.08333333333333333), ('steps such as', 0.08333333333333333), ('such as word', 0.08333333333333333), ('as word alignment', 0.08333333333333333), ('word alignment and', 0.08333333333333333), ('alignment and language', 0.08333333333333333), ('and language modeling', 0.08333333333333333), ('language modeling that', 0.08333333333333333), ('modeling that was', 0.08333333333333333), ('that was used', 0.08333333333333333), ('was used in', 0.08333333333333333), ('used in statistical', 0.08333333333333333), ('in statistical machine', 0.08333333333333333), ('statistical machine translation', 0.08333333333333333), ('translation (smt', 0.08333333333333333), ('(smt)', 0.08333333333333333), ('smt) .', 0.08333333333333333), (') . the', 0.08333333333333333), ('. the following', 0.08333333333333333), ('the following is', 0.08333333333333333), ('following is a', 0.08333333333333333), ('is a list', 0.08333333333333333), ('a list of', 0.08333333333333333), ('list of some', 0.08333333333333333), ('of some of', 0.08333333333333333), ('of the most', 0.08333333333333333), ('the most commonly', 0.08333333333333333), ('most commonly researched', 0.08333333333333333), ('commonly researched tasks', 0.08333333333333333), ('researched tasks in', 0.08333333333333333), ('tasks in natural', 0.08333333333333333), ('processing . some', 0.08333333333333333), ('some of these', 0.08333333333333333), ('of these tasks', 0.08333333333333333), ('these tasks have', 0.08333333333333333), ('tasks have direct', 0.08333333333333333), ('have direct real-world', 0.08333333333333333), ('direct real-world applications', 0.08333333333333333), ('real-world applications ,', 0.08333333333333333), ('applications , while', 0.08333333333333333), (', while others', 0.08333333333333333), ('while others more', 0.08333333333333333), ('others more commonly', 0.08333333333333333), ('more commonly serve', 0.08333333333333333), ('commonly serve as', 0.08333333333333333), ('serve as

subtasks', 0.08333333333333333), ('as subtasks that', 0.08333333333333333), ('subtasks that are', 0.08333333333333333), ('that are used', 0.08333333333333333), ('are used to', 0.08333333333333333), ('used to aid', 0.08333333333333333), ('to aid in', 0.08333333333333333), ('aid in solving', 0.08333333333333333), ('in solving larger', 0.08333333333333333), ('solving larger tasks', 0.08333333333333333), ('larger tasks .', 0.08333333333333333), ('tasks . though', 0.08333333333333333), ('. though natural', 0.08333333333333333), ('though natural language', 0.08333333333333333), ('language processing tasks', 0.08333333333333333), ('processing tasks are', 0.08333333333333333), ('tasks are closely', 0.08333333333333333), ('are closely intertwined', 0.08333333333333333), ('closely intertwined ,', 0.08333333333333333), ('intertwined , they', 0.08333333333333333), (', they can', 0.08333333333333333), ('they can be', 0.08333333333333333), ('can be subdivided', 0.08333333333333333), ('be subdivided into', 0.08333333333333333), ('subdivided into categories', 0.08333333333333333), ('into categories for', 0.08333333333333333), ('categories for convenience', 0.08333333333333333), ('for convenience .', 0.08333333333333333), ('convenience . a', 0.08333333333333333), ('. a coarse', 0.08333333333333333), ('a coarse division', 0.08333333333333333), ('coarse division is', 0.08333333333333333), ('division is given', 0.08333333333333333), ('is given below', 0.08333333333333333), ('given below .', 0.08333333333333333), ('below . based', 0.08333333333333333), ('. based on', 0.08333333333333333), ('based on long-standing', 0.08333333333333333), ('on long-standing trends', 0.08333333333333333), ('long-standing trends in', 0.08333333333333333), ('trends in the', 0.08333333333333333), ('in the field', 0.08333333333333333), ('the field ,', 0.08333333333333333), ('field , it', 0.08333333333333333), (', it is', 0.08333333333333333), ('it is possible', 0.08333333333333333), ('is possible to', 0.08333333333333333), ('possible to extrapolate', 0.08333333333333333), ('to extrapolate future', 0.08333333333333333), ('extrapolate future directions', 0.08333333333333333), ('future directions of', 0.08333333333333333), ('directions of nlp', 0.08333333333333333), ('of nlp .', 0.08333333333333333), ('nlp . as', 0.08333333333333333), ('. as of', 0.08333333333333333), ('as of 2020', 0.08333333333333333), ('of 2020 ,', 0.08333333333333333), ('2020 , three', 0.08333333333333333), (', three trends', 0.08333333333333333), ('three trends among', 0.08333333333333333), ('trends among the', 0.08333333333333333), ('among the topics', 0.08333333333333333), ('the topics of', 0.08333333333333333), ('topics of the', 0.08333333333333333), ('of the long-standing', 0.08333333333333333), ('the long-standing series', 0.08333333333333333), ('long-standing series of', 0.08333333333333333), ('series of conll', 0.08333333333333333), ('of conll shared', 0.08333333333333333), ('conll shared tasks', 0.16666666666666666), ('shared tasks can', 0.08333333333333333), ('tasks can be', 0.08333333333333333), ('can be observed', 0.08333333333333333), ('be observed :', 0.08333333333333333), ('observed : [', 0.08333333333333333), (': [41', 0.08333333333333333), ('[41]', 0.08333333333333333), ('41] most', 0.08333333333333333), ('] most higher-level', 0.08333333333333333), ('most higher-level nlp', 0.08333333333333333), ('higher-level nlp applications', 0.08333333333333333), ('nlp applications involve', 0.08333333333333333), ('applications involve aspects', 0.08333333333333333), ('involve aspects that',

0.08333333333333333), ('aspects that emulate', 0.08333333333333333), ('that emulate intelligent', 0.08333333333333333), ('emulate intelligent behaviour', 0.08333333333333333), ('intelligent behaviour and', 0.08333333333333333), ('behaviour and apparent', 0.08333333333333333), ('and apparent comprehension', 0.08333333333333333), ('apparent comprehension of', 0.08333333333333333), ('comprehension of natural', 0.08333333333333333), ('language . more', 0.08333333333333333), ('. more broadly', 0.08333333333333333), ('more broadly speaking', 0.08333333333333333), ('broadly speaking ,', 0.08333333333333333), ('speaking , the', 0.08333333333333333), (', the technical', 0.08333333333333333), ('the technical operationalization', 0.08333333333333333), ('technical operationalization of', 0.08333333333333333), ('operationalization of increasingly', 0.08333333333333333), ('of increasingly advanced', 0.08333333333333333), ('increasingly advanced aspects', 0.08333333333333333), ('advanced aspects of', 0.08333333333333333), ('aspects of cognitive', 0.08333333333333333), ('of cognitive behaviour', 0.08333333333333333), ('cognitive behaviour represents', 0.08333333333333333), ('behaviour represents one', 0.08333333333333333), ('represents one of', 0.08333333333333333), ('one of the', 0.08333333333333333), ('of the developmental', 0.08333333333333333), ('the developmental trajectories', 0.08333333333333333), ('developmental trajectories of', 0.08333333333333333), ('trajectories of nlp', 0.08333333333333333), ('of nlp (', 0.08333333333333333), ('nlp (see', 0.08333333333333333), ('(see trends', 0.08333333333333333), ('see trends among', 0.08333333333333333), ('trends among conll', 0.08333333333333333), ('among conll shared', 0.08333333333333333), ('shared tasks above', 0.08333333333333333), ('tasks above)', 0.08333333333333333), ('above) .', 0.08333333333333333), (') . cognition', 0.08333333333333333), ('. cognition refers', 0.08333333333333333), ('cognition refers to', 0.08333333333333333), ('refers to ``', 0.08333333333333333), ('to `` the', 0.08333333333333333), ('`` the mental', 0.08333333333333333), ('the mental action', 0.08333333333333333), ('mental action or', 0.08333333333333333), ('action or process', 0.08333333333333333), ('or process of', 0.08333333333333333), ('process of acquiring', 0.08333333333333333), ('of acquiring knowledge', 0.08333333333333333), ('acquiring knowledge and', 0.08333333333333333), ('knowledge and understanding', 0.08333333333333333), ('and understanding through', 0.08333333333333333), ('understanding through thought', 0.08333333333333333), ('through thought ,', 0.08333333333333333), ('thought , experience', 0.08333333333333333), (', experience ,', 0.08333333333333333), ('experience , and', 0.08333333333333333), (', and the', 0.08333333333333333), ('and the senses', 0.08333333333333333), ('the senses .', 0.08333333333333333), ('senses . ``', 0.08333333333333333), ('. `` [', 0.08333333333333333), ('`` [42', 0.08333333333333333), ('[42]', 0.08333333333333333), ('42] cognitive', 0.08333333333333333), ('] cognitive science', 0.08333333333333333), ('cognitive science is', 0.08333333333333333), ('science is the', 0.08333333333333333), ('is the interdisciplinary', 0.08333333333333333), ('the interdisciplinary ,', 0.08333333333333333), ('interdisciplinary , scientific', 0.08333333333333333), (', scientific study', 0.08333333333333333), ('scientific study of', 0.08333333333333333), ('study of the', 0.08333333333333333), ('of the mind', 0.08333333333333333), ('the mind and', 0.08333333333333333), ('mind and its', 0.08333333333333333), ('and its

processes', 0.08333333333333333), ('its processes .', 0.08333333333333333), ('processes . [', 0.08333333333333333), ('. [43', 0.08333333333333333), ('[43]', 0.08333333333333333), ('43] cognitive', 0.08333333333333333), ('] cognitive linguistics', 0.08333333333333333), ('cognitive linguistics is', 0.08333333333333333), ('linguistics is an', 0.08333333333333333), ('an interdisciplinary branch', 0.08333333333333333), ('interdisciplinary branch of', 0.08333333333333333), ('branch of linguistics', 0.08333333333333333), ('linguistics , combining', 0.08333333333333333), (', combining knowledge', 0.08333333333333333), ('combining knowledge and', 0.08333333333333333), ('knowledge and research', 0.08333333333333333), ('and research from', 0.08333333333333333), ('research from both', 0.08333333333333333), ('from both psychology', 0.08333333333333333), ('both psychology and', 0.08333333333333333), ('psychology and linguistics', 0.08333333333333333), ('and linguistics .', 0.08333333333333333), ('linguistics . [', 0.08333333333333333), ('. [44', 0.08333333333333333), ('[44]', 0.08333333333333333), ('44] especially', 0.08333333333333333), ('] especially during', 0.08333333333333333), ('especially during the', 0.08333333333333333), ('during the age', 0.08333333333333333), ('the age of', 0.08333333333333333), ('age of symbolic', 0.08333333333333333), ('symbolic nlp ,', 0.08333333333333333), ('nlp , the', 0.08333333333333333), (', the area', 0.08333333333333333), ('the area of', 0.08333333333333333), ('area of computational', 0.08333333333333333), ('of computational linguistics', 0.08333333333333333), ('computational linguistics maintained', 0.08333333333333333), ('linguistics maintained strong', 0.08333333333333333), ('maintained strong ties', 0.08333333333333333), ('strong ties with', 0.08333333333333333), ('ties with cognitive', 0.16666666666666666), ('with cognitive studies', 0.08333333333333333), ('cognitive studies .', 0.08333333333333333), ('studies . as', 0.08333333333333333), ('. as an', 0.08333333333333333), ('as an example', 0.08333333333333333), ('an example ,', 0.08333333333333333), ('example , george', 0.08333333333333333), (', george lakoff', 0.08333333333333333), ('george lakoff offers', 0.08333333333333333), ('lakoff offers a', 0.08333333333333333), ('offers a methodology', 0.08333333333333333), ('a methodology to', 0.08333333333333333), ('methodology to build', 0.08333333333333333), ('to build natural', 0.08333333333333333), ('build natural language', 0.08333333333333333), ('nlp) algorithms', 0.08333333333333333), (') algorithms through', 0.08333333333333333), ('algorithms through the', 0.08333333333333333), ('through the perspective', 0.08333333333333333), ('the perspective of', 0.08333333333333333), ('perspective of cognitive', 0.08333333333333333), ('of cognitive science', 0.08333333333333333), ('cognitive science ,', 0.08333333333333333), ('science , along', 0.08333333333333333), (', along with', 0.08333333333333333), ('along with the', 0.08333333333333333), ('with the findings', 0.08333333333333333), ('the findings of', 0.08333333333333333), ('findings of cognitive', 0.08333333333333333), ('of cognitive linguistics', 0.08333333333333333), ('cognitive linguistics ,', 0.08333333333333333), ('linguistics , [', 0.08333333333333333), (', [45', 0.08333333333333333), ('[45]', 0.08333333333333333), ('45] with', 0.08333333333333333), ('] with two', 0.08333333333333333), ('with two defining', 0.08333333333333333), ('two defining aspects', 0.08333333333333333), ('defining aspects :', 0.08333333333333333),

('aspects : ties', 0.08333333333333333), (': ties with', 0.08333333333333333), ('with cognitive linguistics', 0.08333333333333333), ('cognitive linguistics are', 0.08333333333333333), ('linguistics are part', 0.08333333333333333), ('are part of', 0.08333333333333333), ('part of the', 0.08333333333333333), ('of the historical', 0.08333333333333333), ('the historical heritage', 0.08333333333333333), ('historical heritage of', 0.08333333333333333), ('heritage of nlp', 0.08333333333333333), ('of nlp ,', 0.08333333333333333), ('nlp , but', 0.08333333333333333), (', but they', 0.08333333333333333), ('but they have', 0.08333333333333333), ('they have been', 0.08333333333333333), ('have been less', 0.08333333333333333), ('been less frequently', 0.08333333333333333), ('less frequently addressed', 0.08333333333333333), ('frequently addressed since', 0.08333333333333333), ('addressed since the', 0.08333333333333333), ('since the statistical', 0.08333333333333333), ('the statistical turn', 0.08333333333333333), ('statistical turn during', 0.08333333333333333), ('turn during the', 0.08333333333333333), ('during the 1990s', 0.08333333333333333), ('the 1990s .', 0.08333333333333333), ('1990s . nevertheless', 0.08333333333333333), ('. nevertheless ,', 0.08333333333333333), ('nevertheless , approaches', 0.08333333333333333), (', approaches to', 0.08333333333333333), ('approaches to develop', 0.08333333333333333), ('to develop cognitive', 0.08333333333333333), ('develop cognitive models', 0.08333333333333333), ('cognitive models towards', 0.08333333333333333), ('models towards technically', 0.08333333333333333), ('towards technically operationalizable', 0.08333333333333333), ('technically operationalizable frameworks', 0.08333333333333333), ('operationalizable frameworks have', 0.08333333333333333), ('frameworks have been', 0.08333333333333333), ('have been pursued', 0.08333333333333333), ('been pursued in', 0.08333333333333333), ('pursued in the', 0.08333333333333333), ('in the context', 0.08333333333333333), ('the context of', 0.08333333333333333), ('context of various', 0.08333333333333333), ('of various frameworks', 0.08333333333333333), ('various frameworks ,', 0.08333333333333333), ('frameworks , e.g.', 0.08333333333333333), ('e.g. , of', 0.08333333333333333), (', of cognitive', 0.08333333333333333), ('of cognitive grammar', 0.08333333333333333), ('cognitive grammar ,', 0.08333333333333333), ('grammar , [', 0.25), (', [47', 0.08333333333333333), ('[47]', 0.08333333333333333), ('47] functional', 0.08333333333333333), ('] functional grammar', 0.08333333333333333), ('functional grammar ,', 0.08333333333333333), (', [48', 0.08333333333333333), ('[48]', 0.08333333333333333), ('48] construction', 0.08333333333333333), ('] construction grammar', 0.08333333333333333), ('construction grammar ,', 0.08333333333333333), (', [49', 0.08333333333333333), ('[49]', 0.08333333333333333), ('49] computational', 0.08333333333333333), ('] computational psycholinguistics', 0.08333333333333333), ('computational psycholinguistics and', 0.08333333333333333), ('psycholinguistics and cognitive', 0.08333333333333333), ('and cognitive neuroscience', 0.08333333333333333), ('cognitive neuroscience (', 0.08333333333333333), ('neuroscience (e.g.', 0.08333333333333333), ('e.g. , act-r', 0.08333333333333333), (', act-r)', 0.08333333333333333), ('act-r) ,', 0.08333333333333333), (') , however', 0.08333333333333333), ('however , with', 0.08333333333333333), (', with limited', 0.08333333333333333), ('with limited

```
uptake', 0.08333333333333333), ('limited uptake in', 0.08333333333333333),
('uptake in mainstream', 0.08333333333333333), ('in mainstream nlp',
0.08333333333333333), ('mainstream nlp (', 0.08333333333333333), ('nlp ( as',
0.08333333333333333), ('( as measured', 0.08333333333333333), ('as measured by',
0.08333333333333333), ('measured by presence', 0.08333333333333333), ('by
presence on', 0.08333333333333333), ('presence on major', 0.08333333333333333),
('on major conferences', 0.08333333333333333), ('major conferences [',
0.08333333333333333), ('conferences [ 50', 0.08333333333333333), ('[ 50 ]',
0.08333333333333333), ('50 ] of', 0.08333333333333333), ('] of the',
0.08333333333333333), ('of the acl', 0.08333333333333333), ('the acl )',
0.08333333333333333), ('acl ) .', 0.08333333333333333), (') . more',
0.08333333333333333), ('. more recently', 0.08333333333333333), ('more recently
,', 0.08333333333333333), ('recently , ideas', 0.08333333333333333), (', ideas
of', 0.16666666666666666), ('ideas of cognitive', 0.16666666666666666), ('of
cognitive nlp', 0.16666666666666666), ('cognitive nlp have',
0.08333333333333333), ('nlp have been', 0.08333333333333333), ('have been
revived', 0.08333333333333333), ('been revived as', 0.08333333333333333),
('revived as an', 0.08333333333333333), ('as an approach', 0.08333333333333333),
('an approach to', 0.08333333333333333), ('approach to achieve',
0.08333333333333333), ('to achieve explainability', 0.08333333333333333),
('achieve explainability ,', 0.08333333333333333), ('explainability , e.g.',
0.08333333333333333), ('e.g. , under', 0.08333333333333333), (', under the',
0.08333333333333333), ('under the notion', 0.08333333333333333), ('the notion
of', 0.08333333333333333), ('notion of ``', 0.08333333333333333), ('of ``
cognitive', 0.08333333333333333), ('`` cognitive ai', 0.08333333333333333),
("cognitive ai ''", 0.08333333333333333), ("ai '' .", 0.08333333333333333), ("''
. [", 0.08333333333333333), ('. [ 51', 0.08333333333333333), ('[ 51 ]',
0.08333333333333333), ('51 ] likewise', 0.08333333333333333), ('] likewise ,',
0.08333333333333333), ('likewise , ideas', 0.08333333333333333), ('cognitive nlp
are', 0.08333333333333333), ('nlp are inherent', 0.08333333333333333), ('are
inherent to', 0.08333333333333333), ('inherent to neural', 0.08333333333333333),
('to neural models', 0.08333333333333333), ('neural models multimodal',
0.08333333333333333), ('models multimodal nlp', 0.08333333333333333),
('multimodal nlp (', 0.08333333333333333), ('nlp ( although',
0.08333333333333333), ('( although rarely', 0.08333333333333333), ('although
rarely made', 0.08333333333333333), ('rarely made explicit',
0.08333333333333333), ('made explicit )', 0.08333333333333333), ('explicit ) .',
0.08333333333333333), (') . [', 0.08333333333333333), ('. [ 52',
0.08333333333333333), ('[ 52 ]', 0.08333333333333333)])
```

[ ]:
```
# 2.2 c - Calculating the senetence scores
def calc_sent_scores_ngrams(sent_tokens, ngram_freq, n_grams):

    sent_score = {}
    maximum_cnt = ngram_freq.most_common(1)[0][1]
    for each_word in ngram_freq.keys():
        ngram_freq[each_word] = ngram_freq[each_word] / maximum_cnt
```

```
    for eachsentence in sent_tokens:
        sum = 0
        sent_n_grams = gen_ngrams(eachsentence.lower(), n_grams)
        for each_sent_ngram in sent_n_grams:
            sum = sum + ngram_freq[each_sent_ngram]
            sent_score[eachsentence] = sum

    # ranked_sent = dict(sorted(sent_score.items(), key=operator.itemgetter(1),␣
 ↪reverse=True))
    return sent_score




n_grams = 3
ngram_freq = FreqDist(gen_ngrams(content, n_grams))
sent_score = calc_sent_scores_ngrams(my_sentences, ngram_freq, n_grams)
sent_score
```

[ ]:
```
# 2.2 d

from heapq import nlargest
def n_gram_summary_based_on_sent_count(n_grams, sent_count):
    ngrams_frequency = FreqDist(gen_ngrams(content, n_grams))
    sent_scores = calc_sent_scores_ngrams(my_sentences, ngrams_frequency,␣
 ↪n_grams)
    sent_summary = nlargest(sent_count, sent_scores, key=sent_scores.get)
    final_res = " ".join(sent_summary)
    return final_res

print(n_gram_summary_based_on_sent_count(3,3))
```

1.3 Task 3 - Comparison

In the first task, we divided the sentences word by word and found out which sentence had the highest score based on their weighted frequencies.

In the second task, we split the sentences into two or more words and then found out which one occurred the most based on their frequencies.

The second task produced more meaningful and trustworthy results as we have the chance to consider phrases instead of just using a single word. For example, the phrase 'natural processing language' has a higher probability of appearing together as it is a more coherent and also commonly used phrase. Whereas the individual words - natural, processing and language would mean differently in different contexts.

115 words