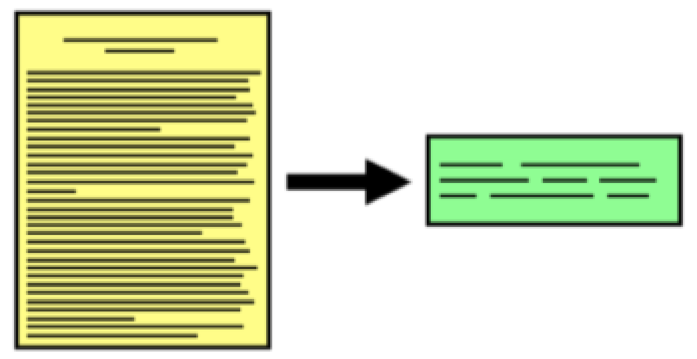
Automatic Text Summarization and Keyword Extraction using Natural Language Processing (NLP)



Abstract

Automatic Text Summarization and Keyword Extraction are one of the most challenging and interesting problems in the field of Natural Language Processing. Text summarization is a process of generating a concise and meaningful summary of text from multiple text resources such as books, news articles, blog posts, research papers, emails, and tweets. Keyword extraction is an important process of highlighting important words, phrases and expressions in a particular content.

Natural Language Processing is an interdisciplinary field of Artificial Intelligence. It is the art of extracting information, hidden insights from unstructured text. NLTK is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets.

It implements website link and text summarization using natural language processing which helps the users to save time and to use the functionality even more efficiently. The keyword extraction feature also plays a vital role in providing the user with the gist of the complete document or website within seconds.

Literature Survey

Research Paper 1:

Title:Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF).

Author(s):Hans Christian,Mikhael Pramodana Agus,Derwin Suhartono

Learning(s):The research paper aimed to produce an automatic text summarizer implemented with TF-IDF algorithm and to compare it with other various online source of automatic text summarizer.

Research Paper 2:

Title:Automatic Text Summarization and keyword Extraction using Natural Language Processing

Author(s):Avinash Payak,Saurabh Rai,Kanishka Shrivastava

Learning(s): This paper gives a brief idea about word frequency algorithm and Text Rank algorithm which are used for text summarization and keyword extraction.

Existing System

The existing system does not focus on Advanced NLP Techniques. End user does not get reliable summaries,which uses Abstractive technique.

Drawbacks in Existing System:-

- Summary is less accurate
- Time constraint is more.
- Computational speed is more.
- It uses Abstractive Summarization.

Proposed System

In the proposed system we are using extractive technique which will provide reliable summary and keywords.In this project input is a web page url and output will be a reliable summary by using various NLP algorithms.

Advantages of Proposed System:

- Summary is more accurate.
- Time Constraint is less.
- Computational speed is more.
- It uses Extractive Summarization.

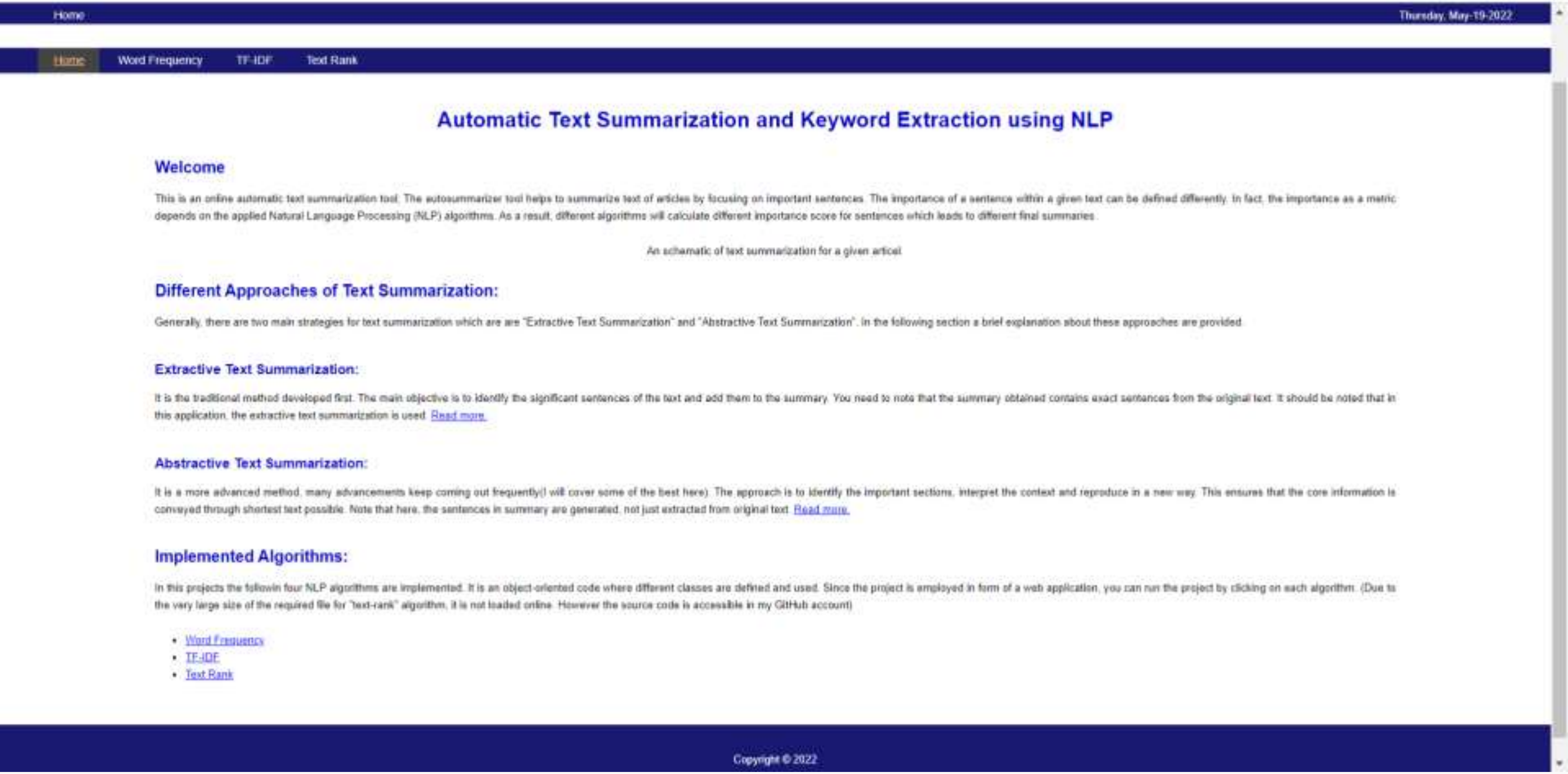
Resources used for the project

Software:

Browser: Google Chrome
OS: Windows
Programming Language: Python
Scripting Language: HTML, CSS, JavaScript
Editor: PyCharm

Hardware:

Laptop with RAM 8 GB
Processor: Intel® Core™ i5
Hard Disk: 500 GB



Automatic Text Summarization and Keyword Extraction using NLP

Welcome

This is an online automatic text summarization tool. The autosummarizer tool helps to summarize text of articles by focusing on important sentences. The importance of a sentence within a given text can be defined differently. In fact, the importance as a metric depends on the applied Natural Language Processing (NLP) algorithms. As a result, different algorithms will calculate different importance score for sentences which leads to different final summaries.

An schematic of text summarization for a given articel

Different Approaches of Text Summarization:

Generally, there are two main strategies for text summarization which are are "Extractive Text Summarization" and "Abstractive Text Summarization". In the following section a brief explanation about these approaches are provided.

Extractive Text Summarization:

It is the traditional method developed first. The main objective is to identify the significant sentences of the text and add them to the summary. You need to note that the summary obtained contains exact sentences from the original text. It should be noted that in this application, the extractive text summarization is used. Read more.

Abstractive Text Summarization:

It is a more advanced method, many advancements keep coming out frequently(I will cover some of the best here). The approach is to identify the important sections, interpret the context and reproduce in a new way. This ensures that the core information is conveyed through shortest text possible. Note that here, the sentences in summary are generated, not just extracted from original text. Read more.

Implemented Algorithms:

In this projects the followin four NLP algorithms are implemented. It is an object-oriented code where different classes are defined and used. Since the project is employed in form of a web application, you can run the project by clicking on each algorithm. (Due to the very large size of the required file for "text-rank" algorithm, it is not loaded online. However the source code is accessible in my GitHub account)

- Word Frequency
- TF-IDF
- Text Rank

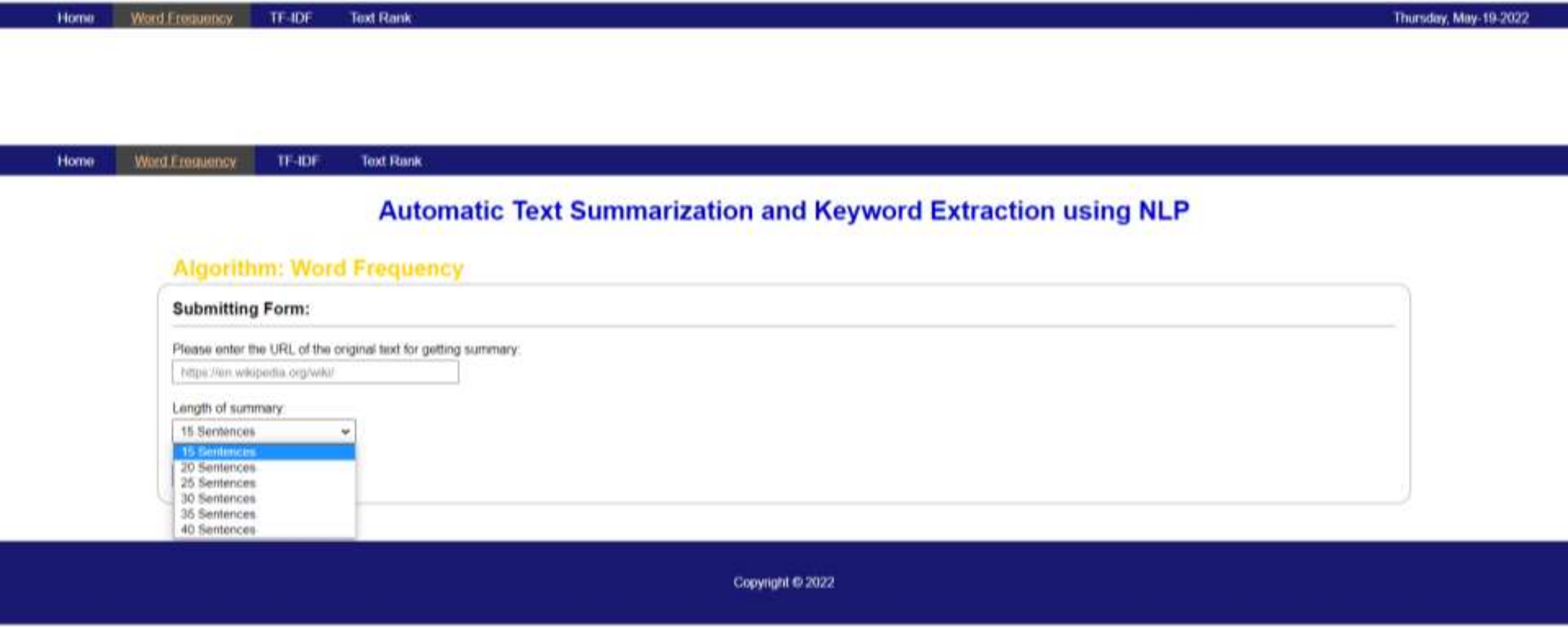
Copyright © 2022

Home Page

Methodology

The project is implemented using three algorithms namely Word Frequency, TF-IDF and Text Rank .

- Text Summarization and Keyword Extraction using Word Frequency algorithm.
- Text Summarization and keyword Extraction using TF-IDF algorithm.
- Text Summarization and Keyword Extraction using Text Rank algorithm.



Automatic Text Summarization and Keyword Extraction using NLP

Algorithm: Word Frequency

Submitting Form:

Please enter the URL of the original text for getting summary:

https://en.wikipedia.org/wiki/

Length of summary:

15 Sentences

- 15 Sentences
- 20 Sentences
- 25 Sentences
- 30 Sentences
- 35 Sentences
- 40 Sentences

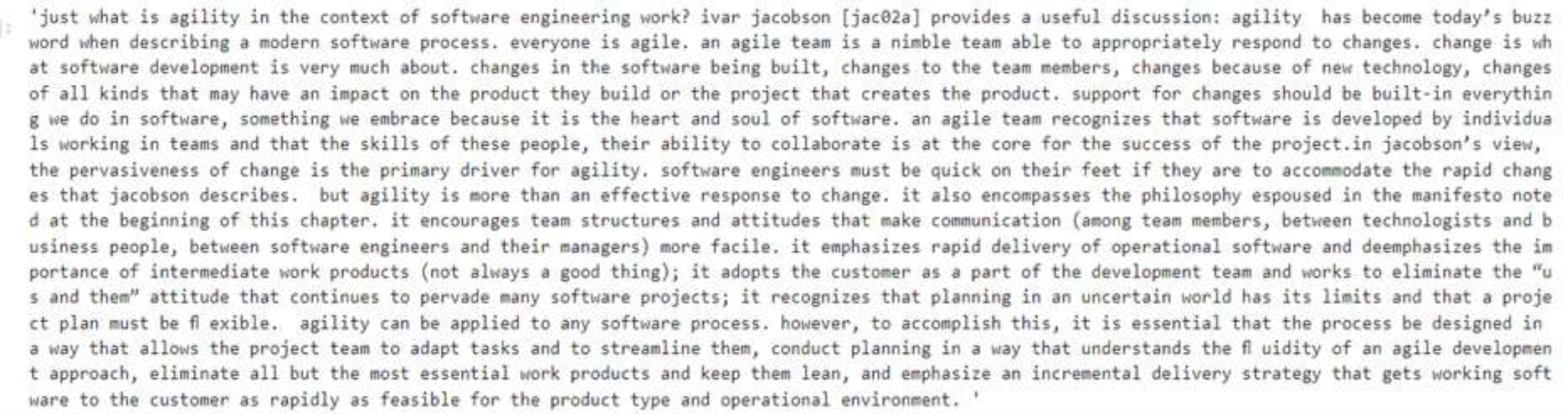
Copyright © 2022

User is given the choice of algorithm and length of summary to be generated.

Word frequency algorithm

Text Summarization and Keyword Extraction using Word Frequency algorithm.

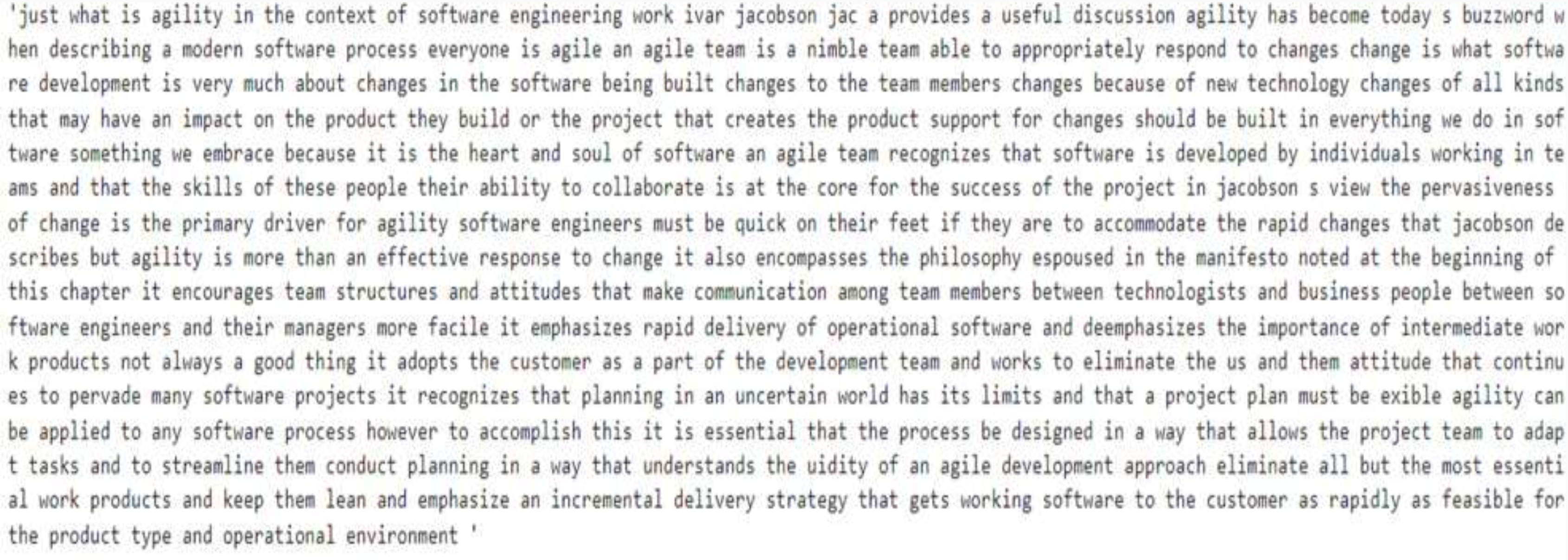
First a url of the website which contains the text is given as input.For example the text is like this from which the summary and keywords must be drawn :



```
'just what is agility in the context of software engineering work? ivar jacobson [jac02a] provides a useful discussion: agility  has become today's buzz
word when describing a modern software process. everyone is agile. an agile team is a nimble team able to appropriately respond to changes. change is wh
at software development is very much about. changes in the software being built, changes to the team members, changes because of new technology, changes
of all kinds that may have an impact on the product they build or the project that creates the product. support for changes should be built-in everythin
g we do in software, something we embrace because it is the heart and soul of software. an agile team recognizes that software is developed by individua
ls working in teams and that the skills of these people, their ability to collaborate is at the core for the success of the project.in jacobson's view,
the pervasiveness of change is the primary driver for agility. software engineers must be quick on their feet if they are to accommodate the rapid chang
es that jacobson describes.  but agility is more than an effective response to change. it also encompasses the philosophy espoused in the manifesto note
d at the beginning of this chapter. it encourages team structures and attitudes that make communication (among team members, between technologists and b
usiness people, between software engineers and their managers) more facile. it emphasizes rapid delivery of operational software and deemphasizes the im
portance of intermediate work products (not always a good thing); it adopts the customer as a part of the development team and works to eliminate the "u
s and them" attitude that continues to pervade many software projects; it recognizes that planning in an uncertain world has its limits and that a proje
ct plan must be fl exible.  agility can be applied to any software process. however, to accomplish this, it is essential that the process be designed in
a way that allows the project team to adapt tasks and to streamline them, conduct planning in a way that understands the fl uidity of an agile developmen
t approach, eliminate all but the most essential work products and keep them lean, and emphasize an incremental delivery strategy that gets working soft
ware to the customer as rapidly as feasible for the product type and operational environment. '
```

How the word frequency algorithm works:

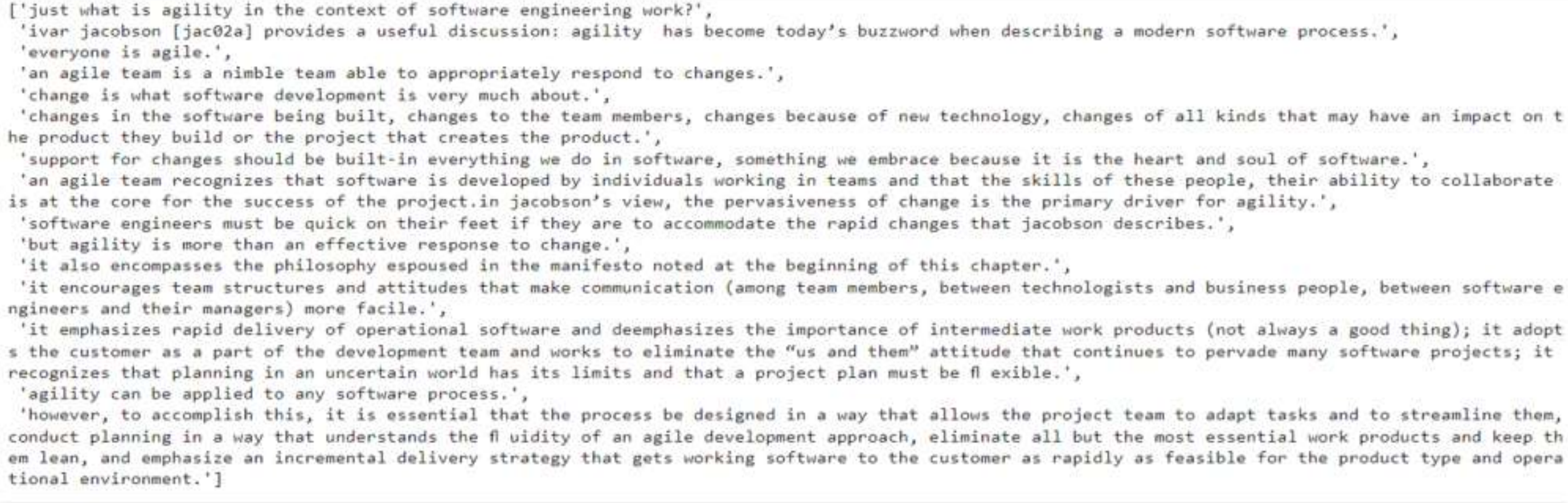
Step 1 : Cleaning the text.



```
'just what is agility in the context of software engineering work ivar jacobson jac a provides a useful discussion agility has become today s buzzword w
hen describing a modern software process everyone is agile an agile team is a nimble team able to appropriately respond to changes change is what softwa
re development is very much about changes in the software being built changes to the team members changes because of new technology changes of all kinds
that may have an impact on the product they build or the project that creates the product support for changes should be built in everything we do in sof
tware something we embrace because it is the heart and soul of software an agile team recognizes that software is developed by individuals working in te
ams and that the skills of these people their ability to collaborate is at the core for the success of the project in jacobson s view the pervasiveness
of change is the primary driver for agility software engineers must be quick on their feet if they are to accommodate the rapid changes that jacobson de
scribes but agility is more than an effective response to change it also encompasses the philosophy espoused in the manifesto noted at the beginning of
this chapter it encourages team structures and attitudes that make communication among team members between technologists and business people between so
ftware engineers and their managers more facile it emphasizes rapid delivery of operational software and deemphasizes the importance of intermediate wor
k products not always a good thing it adopts the customer as a part of the development team and works to eliminate the us and them attitude that continu
es to pervade many software projects it recognizes that planning in an uncertain world has its limits and that a project plan must be exible agility can
be applied to any software process however to accomplish this it is essential that the process be designed in a way that allows the project team to adap
t tasks and to streamline them conduct planning in a way that understands the uidity of an agile development approach eliminate all but the most essenti
al work products and keep them lean and emphasize an incremental delivery strategy that gets working software to the customer as rapidly as feasible for
the product type and operational environment '
```

How the word frequency algorithm works:

Step 2: Splitting the text into sentences.

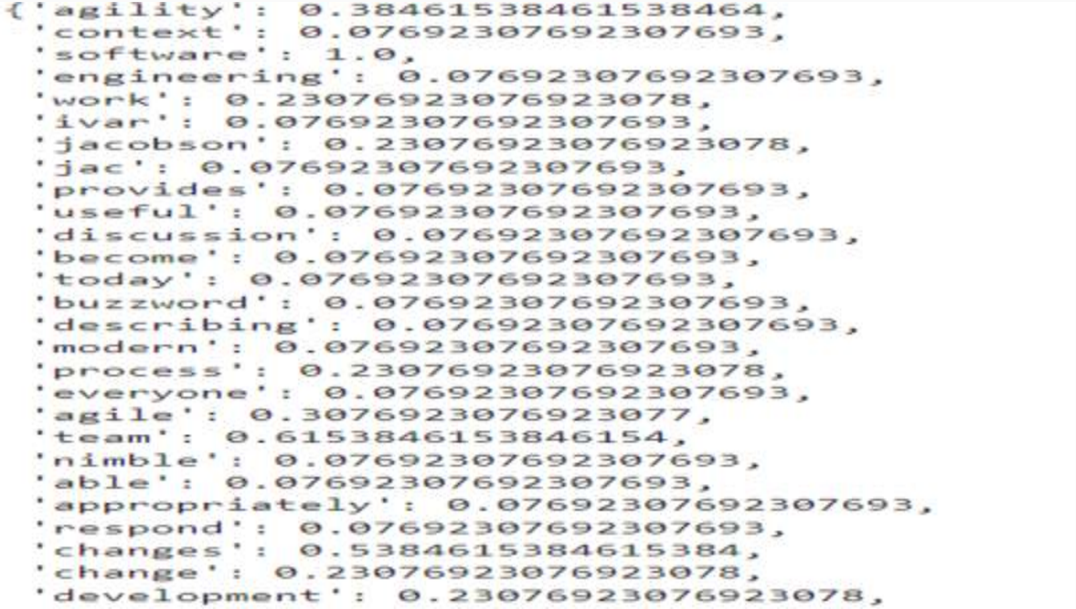


```
['just what is agility in the context of software engineering work?',
 'ivar jacobson [jac02a] provides a useful discussion: agility  has become today’s buzzword when describing a modern software process.',
 'everyone is agile.',
 'an agile team is a nimble team able to appropriately respond to changes.',
 'change is what software development is very much about.',
 'changes in the software being built, changes to the team members, changes because of new technology, changes of all kinds that may have an impact on t
he product they build or the project that creates the product.',
 'support for changes should be built-in everything we do in software, something we embrace because it is the heart and soul of software.',
 'an agile team recognizes that software is developed by individuals working in teams and that the skills of these people, their ability to collaborate
is at the core for the success of the project.in jacobson’s view, the pervasiveness of change is the primary driver for agility.',
 'software engineers must be quick on their feet if they are to accommodate the rapid changes that jacobson describes.',
 'but agility is more than an effective response to change.',
 'it also encompasses the philosophy espoused in the manifesto noted at the beginning of this chapter.',
 'it encourages team structures and attitudes that make communication (among team members, between technologists and business people, between software e
ngineers and their managers) more facile.',
 'it emphasizes rapid delivery of operational software and deemphasizes the importance of intermediate work products (not always a good thing); it adopt
s the customer as a part of the development team and works to eliminate the “us and them” attitude that continues to pervade many software projects; it
recognizes that planning in an uncertain world has its limits and that a project plan must be fl exible.',
 'agility can be applied to any software process.',
 'however, to accomplish this, it is essential that the process be designed in a way that allows the project team to adapt tasks and to streamline them,
conduct planning in a way that understands the fl uidity of an agile development approach, eliminate all but the most essential work products and keep th
em lean, and emphasize an incremental delivery strategy that gets working software to the customer as rapidly as feasible for the product type and opera
tional environment.']
```

How the word frequency algorithm works:

Step 3: Tokenizing the sentences into tokens and finding the frequency(F) of each word.

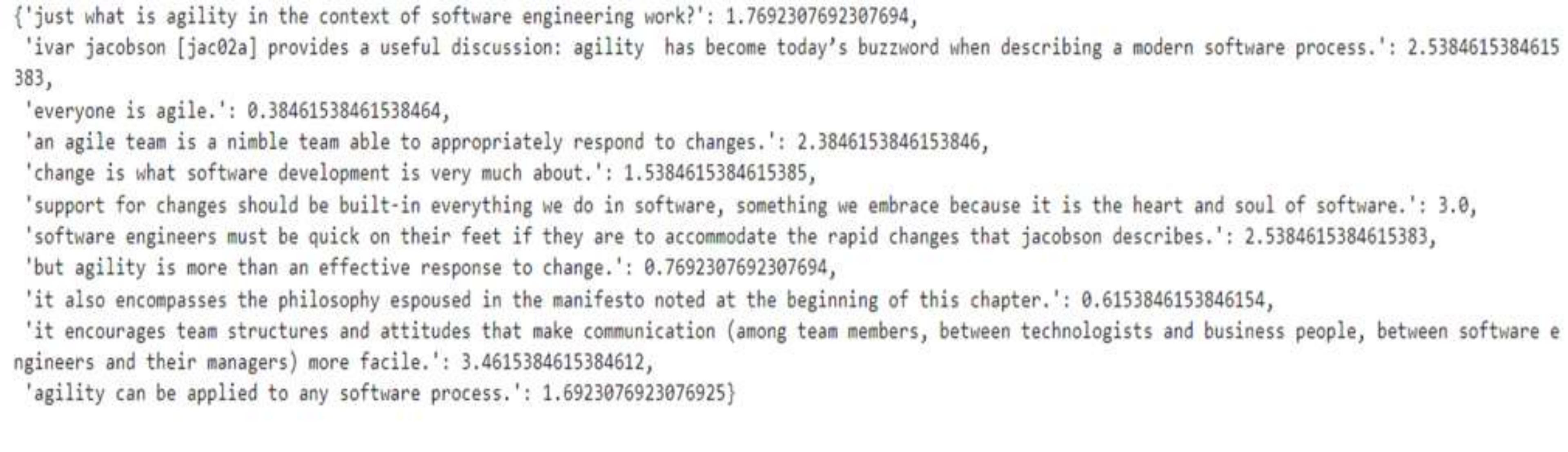
After calculating the frequency, weighted frequency(WF) is calculated.{WF=F/(max)F}.



```
{'agility': 0.38461538461538464,
 'context': 0.07692307692307693,
 'software': 1.0,
 'engineering': 0.07692307692307693,
 'work': 0.23076923076923078,
 'ivar': 0.07692307692307693,
 'jacobson': 0.23076923076923078,
 'jac': 0.07692307692307693,
 'provides': 0.07692307692307693,
 'useful': 0.07692307692307693,
 'discussion': 0.07692307692307693,
 'become': 0.07692307692307693,
 'today': 0.07692307692307693,
 'buzzword': 0.07692307692307693,
 'describing': 0.07692307692307693,
 'modern': 0.07692307692307693,
 'process': 0.23076923076923078,
 'everyone': 0.07692307692307693,
 'agile': 0.3076923076923077,
 'team': 0.6153846153846154,
 'nimble': 0.07692307692307693,
 'able': 0.07692307692307693,
 'appropriately': 0.07692307692307693,
 'respond': 0.07692307692307693,
 'changes': 0.5384615384615384,
 'change': 0.23076923076923078,
 'development': 0.23076923076923078,
```

How the word frequency algorithm works:

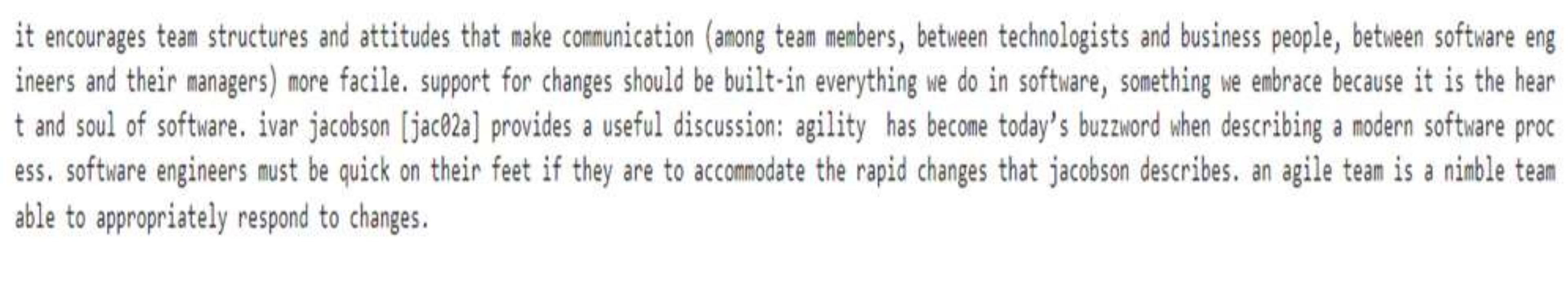
Step 4 : calculating the sentence scores (Renewing the words in the sentence by the weighted frequency).



```
{'just what is agility in the context of software engineering work?': 1.7692307692307694,
 'ivar jacobson [jac02a] provides a useful discussion: agility  has become today’s buzzword when describing a modern software process.': 2.5384615384615
383,
 'everyone is agile.': 0.38461538461538464,
 'an agile team is a nimble team able to appropriately respond to changes.': 2.3846153846153846,
 'change is what software development is very much about.': 1.5384615384615385,
 'support for changes should be built-in everything we do in software, something we embrace because it is the heart and soul of software.': 3.0,
 'software engineers must be quick on their feet if they are to accommodate the rapid changes that jacobson describes.': 2.5384615384615383,
 'but agility is more than an effective response to change.': 0.7692307692307694,
 'it also encompasses the philosophy espoused in the manifesto noted at the beginning of this chapter.': 0.6153846153846154,
 'it encourages team structures and attitudes that make communication (among team members, between technologists and business people, between software e
ngineers and their managers) more facile.': 3.4615384615384612,
 'agility can be applied to any software process.': 1.6923076923076925}
```

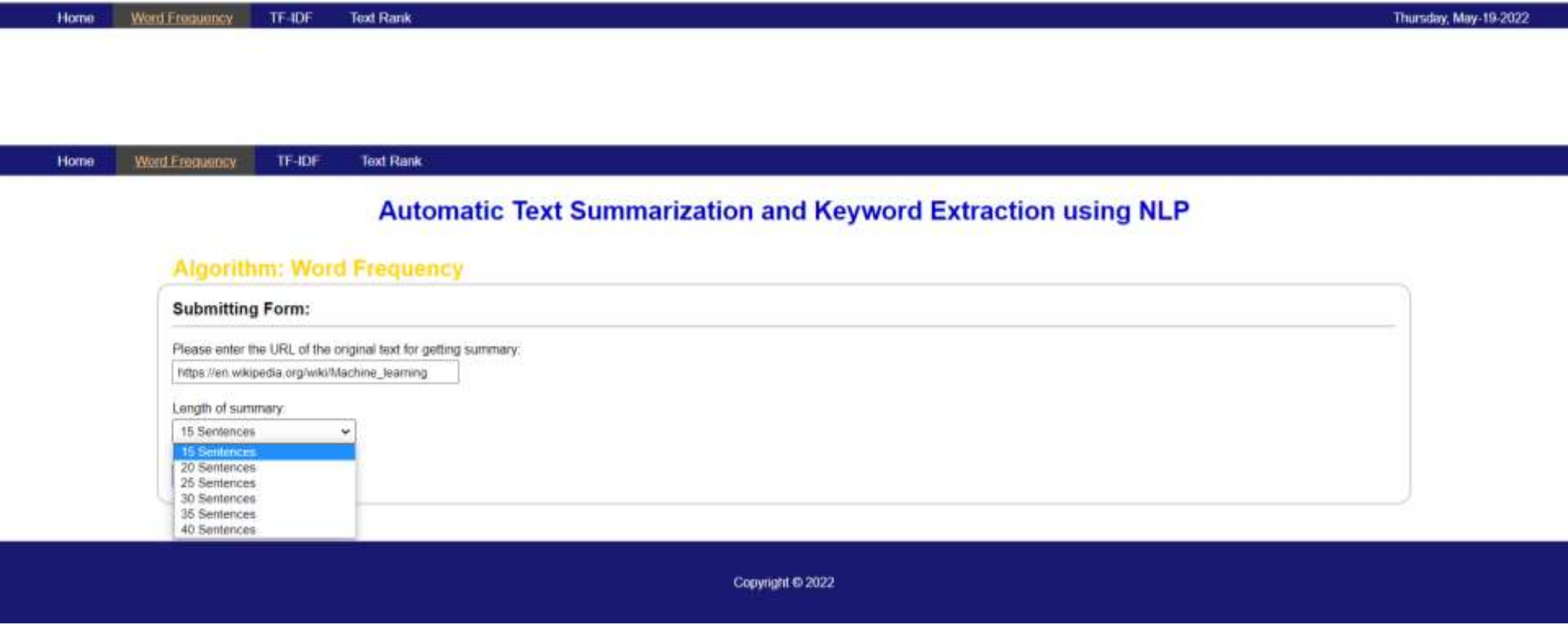
How the word frequency algorithm works:

Step 5: Text summary is obtained by sorting the sentences based on sentence scores. The keywords are the most frequently repeated words in the given input text.



```
it encourages team structures and attitudes that make communication (among team members, between technologists and business people, between software engineers and their managers) more facile. support for changes should be built-in everything we do in software, something we embrace because it is the heart and soul of software. ivar jacobson [jac02a] provides a useful discussion: agility  has become today's buzzword when describing a modern software process. software engineers must be quick on their feet if they are to accommodate the rapid changes that jacobson describes. an agile team is a nimble team able to appropriately respond to changes.
```

The top 10 frequent words are considered as **keywords**.



Automatic Text Summarization and Keyword Extraction using NLP

Algorithm: Word Frequency

Submitting Form:

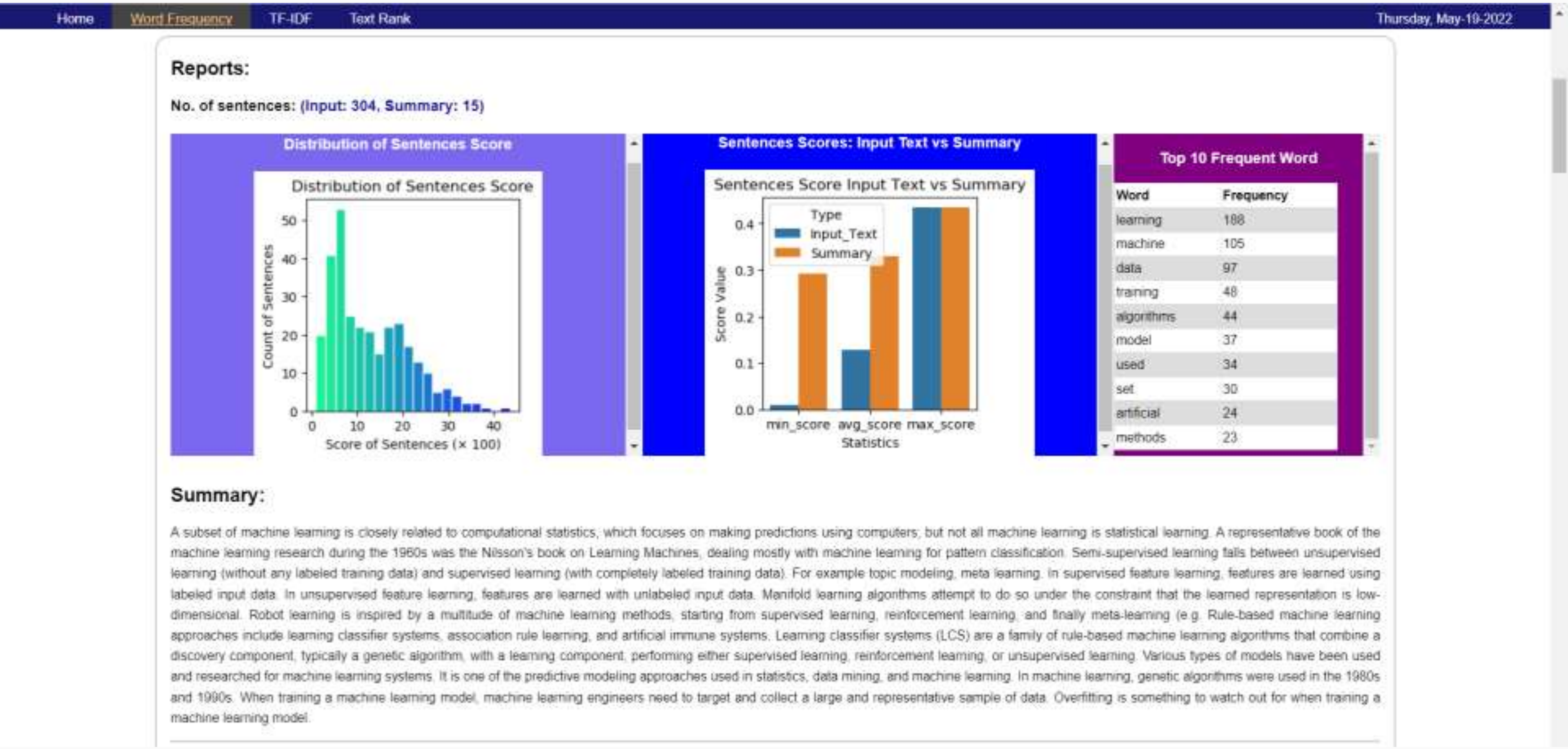
Please enter the URL of the original text for getting summary:

https://en.wikipedia.org/wiki/Machine_learning

Length of summary:

15 Sentences

Output Screen : Giving input link and selecting the summary length.



Reports:

No. of sentences: (Input: 304, Summary: 15)

Top 10 Frequent Word

Word	Frequency
learning	188
machine	105
data	97
training	48
algorithms	44
model	37
used	34
set	30
artificial	24
methods	23

Summary:

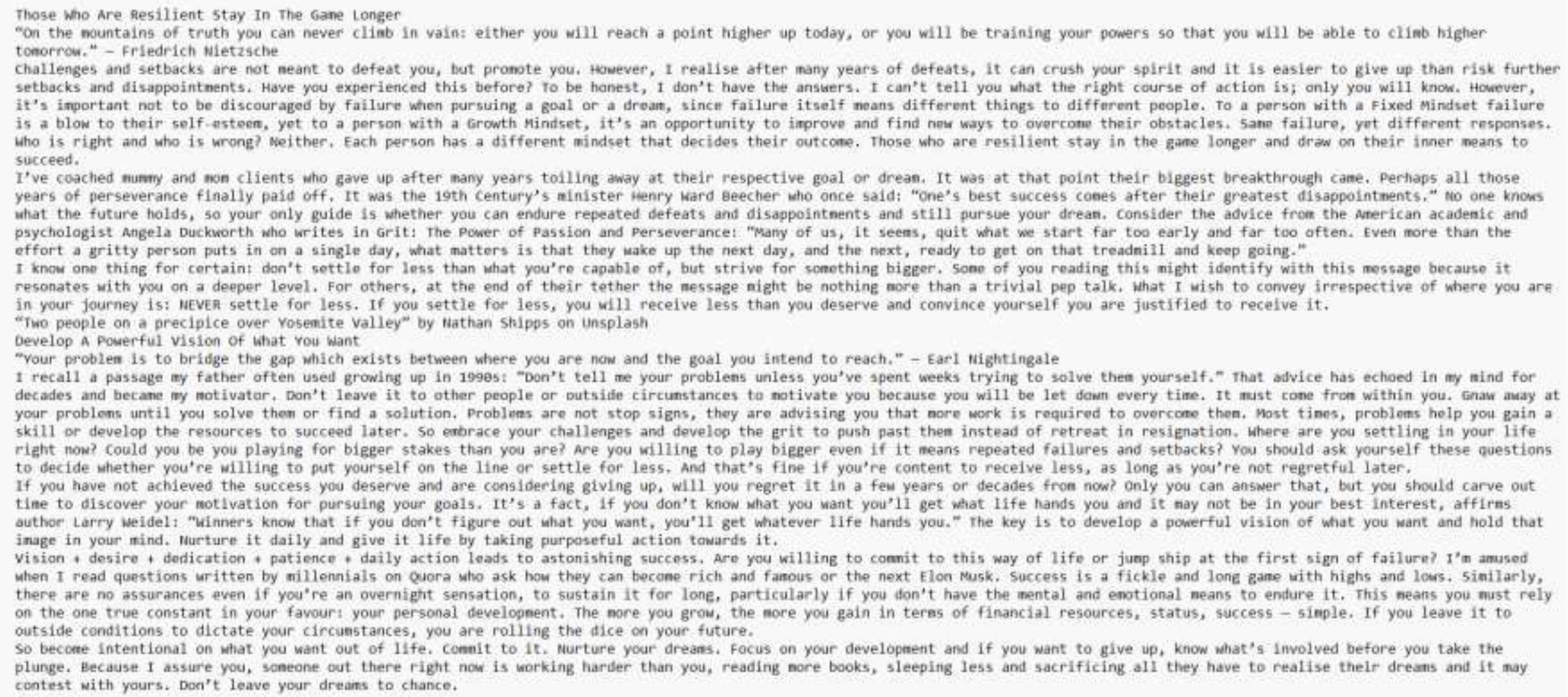
A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. A representative book of the machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). For example topic modeling, meta learning. In supervised feature learning, features are learned using labeled input data. In unsupervised feature learning, features are learned with unlabeled input data. Manifold learning algorithms attempt to do so under the constraint that the learned representation is low-dimensional. Robot learning is inspired by a multitude of machine learning methods, starting from supervised learning, reinforcement learning, and finally meta-learning (e.g. Rule-based machine learning approaches include learning classifier systems, association rule learning, and artificial immune systems. Learning classifier systems (LCS) are a family of rule-based machine learning algorithms that combine a discovery component, typically a genetic algorithm, with a learning component, performing either supervised learning, reinforcement learning, or unsupervised learning. Various types of models have been used and researched for machine learning systems. It is one of the predictive modeling approaches used in statistics, data mining, and machine learning. In machine learning, genetic algorithms were used in the 1980s and 1990s. When training a machine learning model, machine learning engineers need to target and collect a large and representative sample of data. Overfitting is something to watch out for when training a machine learning model.

Final output screen using word frequency algorithm.

TF-IDF Algorithm

Text Summarization and Keyword Extraction using TF-IDF algorithm.

First a url of the website which contains the text is given as input.For example the text is like this from which the summary and keywords must be drawn :



Those Who Are Resilient Stay In The Game Longer

“On the mountains of truth you can never climb in vain: either you will reach a point higher up today, or you will be training your powers so that you will be able to climb higher tomorrow.” – Friedrich Nietzsche

Challenges and setbacks are not meant to defeat you, but promote you. However, I realise after many years of defeats, it can crush your spirit and it is easier to give up than risk further setbacks and disappointments. Have you experienced this before? To be honest, I don’t have the answers. I can’t tell you what the right course of action is; only you will know. However, it’s important not to be discouraged by failure when pursuing a goal or a dream, since failure itself means different things to different people. To a person with a Fixed Mindset failure is a blow to their self-esteem, yet to a person with a Growth Mindset, it’s an opportunity to improve and find new ways to overcome their obstacles. Same failure, yet different responses. Who is right and who is wrong? Neither. Each person has a different mindset that decides their outcome. Those who are resilient stay in the game longer and draw on their inner means to succeed.

I’ve coached mummy and mom clients who gave up after many years toiling away at their respective goal or dream. It was at that point their biggest breakthrough came. Perhaps all those years of perseverance finally paid off. It was the 19th Century’s minister Henry Ward Beecher who once said: “One’s best success comes after their greatest disappointments.” No one knows what the future holds, so your only guide is whether you can endure repeated defeats and disappointments and still pursue your dream. Consider the advice from the American academic and psychologist Angela Duckworth who writes in Grit: The Power of Passion and Perseverance: "Many of us, it seems, quit what we start far too early and far too often. Even more than the effort a gritty person puts in on a single day, what matters is that they wake up the next day, and the next, ready to get on that treadmill and keep going."

I know one thing for certain: don’t settle for less than what you’re capable of, but strive for something bigger. Some of you reading this might identify with this message because it resonates with you on a deeper level. For others, at the end of their tether the message might be nothing more than a trivial pep talk. What I wish to convey irrespective of where you are in your journey is: NEVER settle for less. If you settle for less, you will receive less than you deserve and convince yourself you are justified to receive it.

“Two people on a precipice over Yosemite Valley” by Nathan Shipps on Unsplash

Develop A Powerful Vision Of What You Want

“Your problem is to bridge the gap which exists between where you are now and the goal you intend to reach.” – Earl Nightingale

I recall a passage my father often used growing up in 1990s: "Don’t tell me your problems unless you’ve spent weeks trying to solve them yourself." That advice has echoed in my mind for decades and became my motivator. Don’t leave it to other people or outside circumstances to motivate you because you will be let down every time. It must come from within you. Gnaw away at your problems until you solve them or find a solution. Problems are not stop signs, they are advising you that more work is required to overcome them. Most times, problems help you gain a skill or develop the resources to succeed later. So embrace your challenges and develop the grit to push past them instead of retreat in resignation. Where are you settling in your life right now? Could you be you playing for bigger stakes than you are? Are you willing to play bigger even if it means repeated failures and setbacks? You should ask yourself these questions to decide whether you’re willing to put yourself on the line or settle for less. And that’s fine if you’re content to receive less, as long as you’re not regretful later.

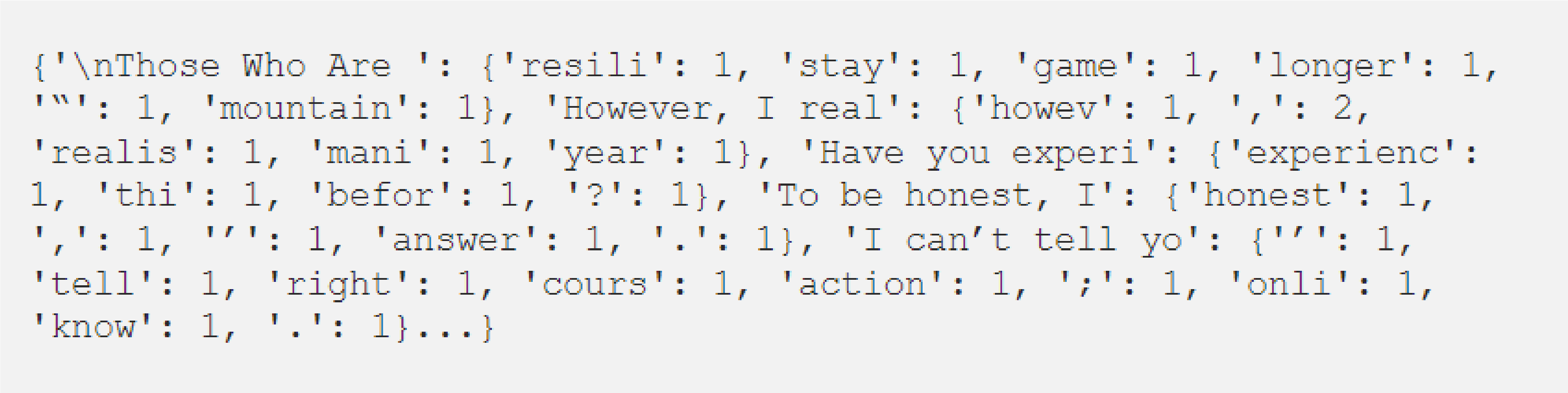
If you have not achieved the success you deserve and are considering giving up, will you regret it in a few years or decades from now? Only you can answer that, but you should carve out time to discover your motivation for pursuing your goals. It’s a fact, if you don’t know what you want you’ll get what life hands you and it may not be in your best interest, affirms author Larry Weidel: "Winners know that if you don’t figure out what you want, you'll get whatever life hands you." The key is to develop a powerful vision of what you want and hold that image in your mind. Nurture it daily and give it life by taking purposeful action towards it.

Vision + desire + dedication + patience + daily action leads to astonishing success. Are you willing to commit to this way of life or jump ship at the first sign of failure? I’m amused when I read questions written by millennials on Quora who ask how they can become rich and famous or the next Elon Musk. Success is a fickle and long game with highs and lows. Similarly, there are no assurances even if you're an overnight sensation, to sustain it for long, particularly if you don’t have the mental and emotional means to endure it. This means you must rely on the one true constant in your favour: your personal development. The more you grow, the more you gain in terms of financial resources, status, success — simple. If you leave it to outside conditions to dictate your circumstances, you are rolling the dice on your future.

So become intentional on what you want out of life. Commit to it. Nurture your dreams. Focus on your development and if you want to give up, know what’s involved before you take the plunge. Because I assure you, someone out there right now is working harder than you, reading more books, sleeping less and sacrificing all they have to realise their dreams and it may contest with yours. Don’t leave your dreams to chance.

How the TF-IDF algorithm works:

Step 1 : Tokenize the sentences.
Step 2: Create the frequency matrix of words in each sentence.



```
{'\nThose Who Are ': {'resili': 1, 'stay': 1, 'game': 1, 'longer': 1,
'“': 1, 'mountain': 1}, 'However, I real': {'howev': 1, ',': 2,
'realis': 1, 'mani': 1, 'year': 1}, 'Have you experi': {'experienc':
1, 'thi': 1, 'befor': 1, '?': 1}, 'To be honest, I': {'honest': 1,
',': 1, '’': 1, 'answer': 1, '.': 1}, 'I can’t tell yo': {'’': 1,
'tell': 1, 'right': 1, 'cours': 1, 'action': 1, ';': 1, 'onli': 1,
'know': 1, '.': 1}...}
```

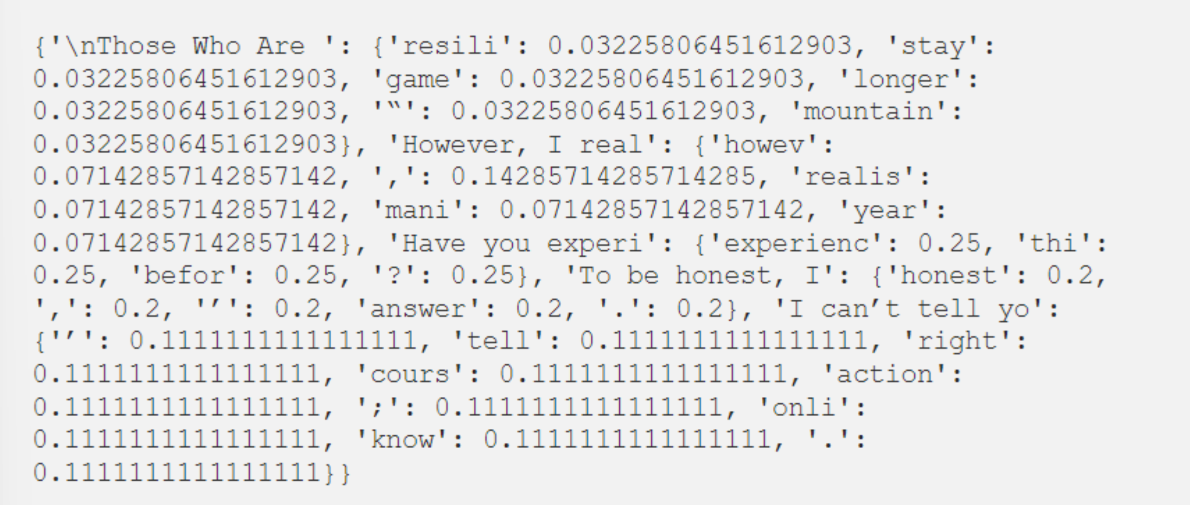
Here, each **sentence is the key** and the **value is a dictionary of word frequency.**

How the TF-IDF algorithm works:

Step 3: Calculate TermFrequency (TF) and generate a matrix.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

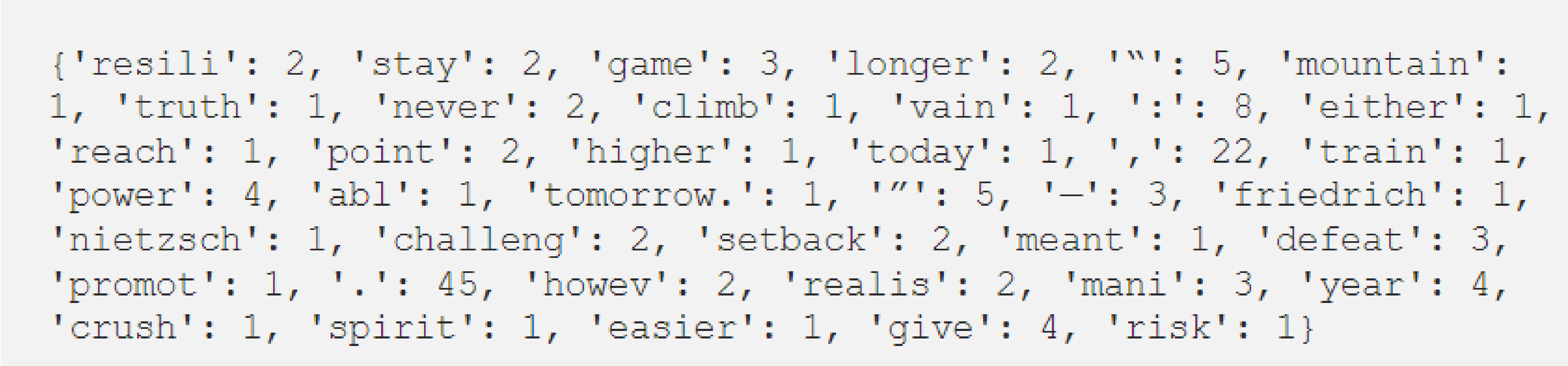
Here, the document is a paragraph, the term is a word in a paragraph.



```
{'\nThose Who Are ': {'resili': 0.03225806451612903, 'stay':
0.03225806451612903, 'game': 0.03225806451612903, 'longer':
0.03225806451612903, '“': 0.03225806451612903, 'mountain':
0.03225806451612903}, 'However, I real': {'howev':
0.07142857142857142, ',': 0.14285714285714285, 'realis':
0.07142857142857142, 'mani': 0.07142857142857142, 'year':
0.07142857142857142}, 'Have you experi': {'experienc': 0.25, 'thi':
0.25, 'befor': 0.25, '?': 0.25}, 'To be honest, I': {'honest': 0.2,
',': 0.2, '’': 0.2, 'answer': 0.2, '.': 0.2}, 'I can’t tell yo':
{'’': 0.1111111111111111, 'tell': 0.1111111111111111, 'right':
0.1111111111111111, 'cours': 0.1111111111111111, 'action':
0.1111111111111111, ';': 0.1111111111111111, 'onli':
0.1111111111111111, 'know': 0.1111111111111111, '.':
0.1111111111111111}}
```

How the TF-IDF algorithm works:

Step 4: Creating a table for documents per words.
we calculate, **how many sentences contain a word**.

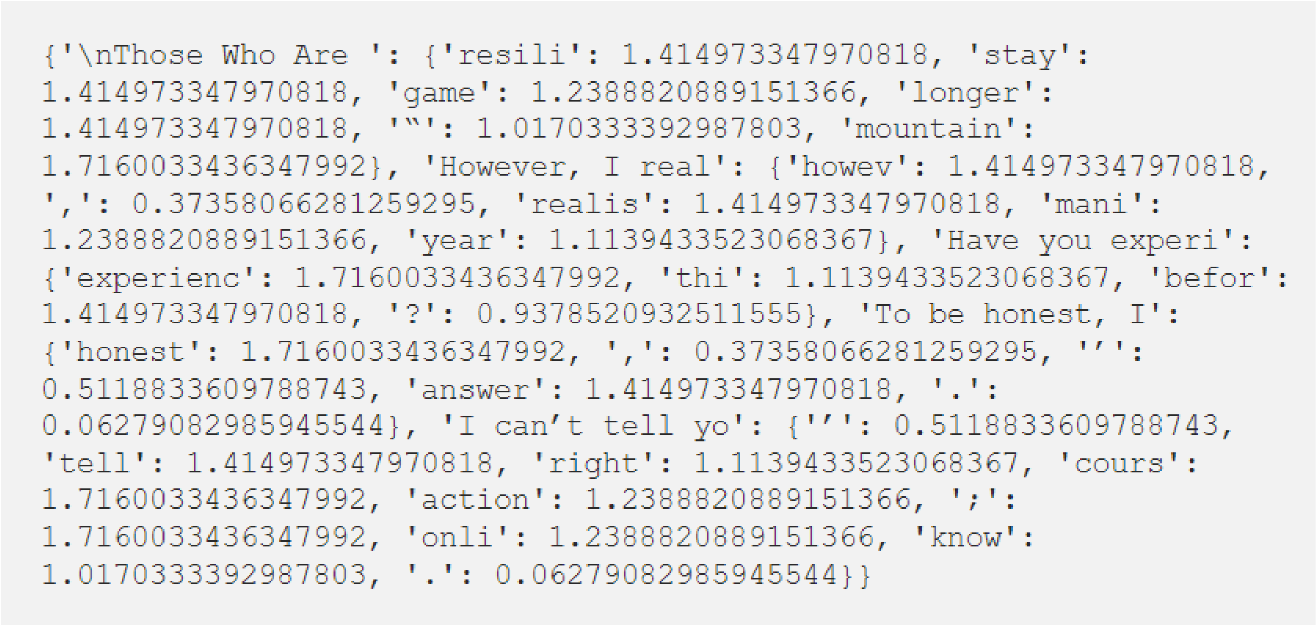


```
{'resili': 2, 'stay': 2, 'game': 3, 'longer': 2, '“': 5, 'mountain':
1, 'truth': 1, 'never': 2, 'climb': 1, 'vain': 1, ':': 8, 'either': 1,
'reach': 1, 'point': 2, 'higher': 1, 'today': 1, ',': 22, 'train': 1,
'power': 4, 'abl': 1, 'tomorrow.': 1, '”': 5, '—': 3, 'friedrich': 1,
'nietzsch': 1, 'challeng': 2, 'setback': 2, 'meant': 1, 'defeat': 3,
'promot': 1, '.': 45, 'howev': 2, 'realis': 2, 'mani': 3, 'year': 4,
'crush': 1, 'spirit': 1, 'easier': 1, 'give': 4, 'risk': 1}
```

i.e, the word resili appears in 2 sentences, game appears in 3 sentences.

How the TF-IDF algorithm works:

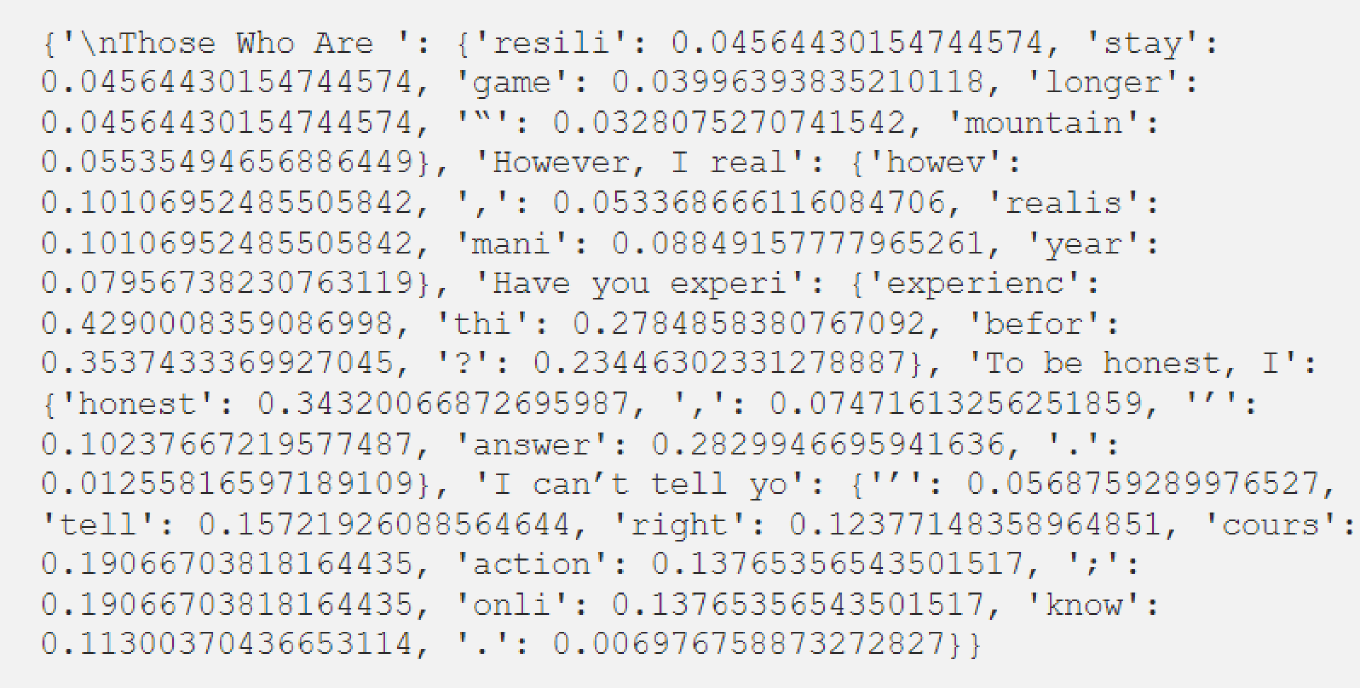
Step 5: Calculate **IDF** and generate matrix.
IDF(t) = log_e(Total number of documents / Number of documents with term t in it)
Here, the document is a paragraph, the term is a word in a paragraph.



```
{'\nThose Who Are ': {'resili': 1.414973347970818, 'stay':
1.414973347970818, 'game': 1.2388820889151366, 'longer':
1.414973347970818, '“': 1.0170333392987803, 'mountain':
1.7160033436347992}, 'However, I real': {'howev': 1.414973347970818,
',': 0.37358066281259295, 'realis': 1.414973347970818, 'mani':
1.2388820889151366, 'year': 1.1139433523068367}, 'Have you experi':
{'experienc': 1.7160033436347992, 'thi': 1.1139433523068367, 'befor':
1.414973347970818, '?': 0.9378520932511555}, 'To be honest, I':
{'honest': 1.7160033436347992, ',': 0.37358066281259295, '’':
0.5118833609788743, 'answer': 1.414973347970818, '.':
0.06279082985945544}, 'I can’t tell yo': {'’': 0.5118833609788743,
'tell': 1.414973347970818, 'right': 1.1139433523068367, 'cours':
1.7160033436347992, 'action': 1.2388820889151366, ';':
1.7160033436347992, 'onli': 1.2388820889151366, 'know':
1.0170333392987803, '.': 0.06279082985945544}}
```

How the TF-IDF algorithm works:

Step 6: Calculate **TF-IDF** and generate a matrix.
TF-IDF algorithm is made of 2 algorithms multiplied together.

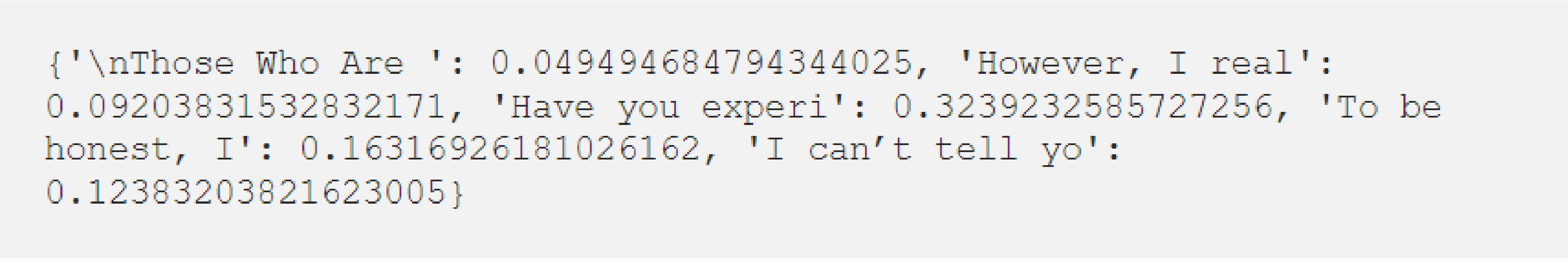


```
{'\nThose Who Are ': {'resili': 0.04564430154744574, 'stay':
0.04564430154744574, 'game': 0.03996393835210118, 'longer':
0.04564430154744574, '“': 0.032807527070741542, 'mountain':
0.05535494656886449}, 'However, I real': {'howev':
0.10106952485505842, ',': 0.05336866116084706, 'realis':
0.10106952485505842, 'mani': 0.08849157777965261, 'year':
0.07956738230763119}, 'Have you experi': {'experienc':
0.4290008359086998, 'thi': 0.2784858380767092, 'befor':
0.3537433369927045, '?': 0.2344630233127887}, 'To be honest, I':
{'honest': 0.34320066872695987, ',': 0.07471613256251859, '’':
0.10237667219577487, 'answer': 0.2829946695941636, '.':
0.01255816597189109}, 'I can’t tell yo': {'’': 0.0568759289976527,
'tell': 0.15721926088564644, 'right': 0.12377148358964851, 'cours':
0.19066703818164435, 'action': 0.13765356543501517, ';':
0.19066703818164435, 'onli': 0.13765356543501517, 'know':
0.11300370436653114, '.': 0.006976758873272827}}
```

How the TF-IDF algorithm works:

Step 7: Score the sentences.

Here, we are using TF-IDF score of words in a sentence to give weight to the paragraph.

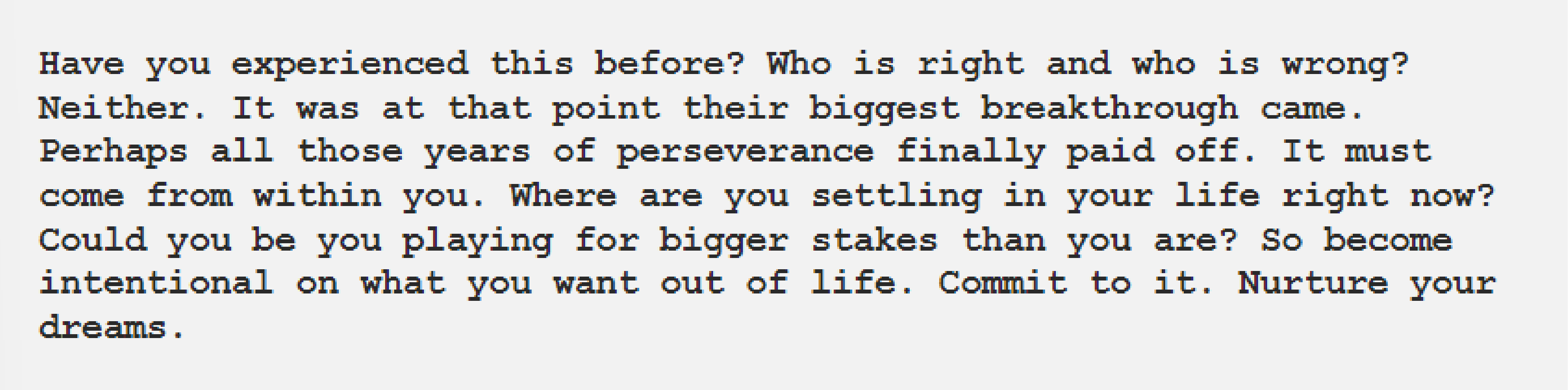


```
{'\nThose Who Are ': 0.049494684794344025, 'However, I real':
0.09203831532832171, 'Have you experi': 0.3239232585727256, 'To be
honest, I': 0.16316926181026162, 'I can’t tell yo':
0.12383203821623005}
```

How the TF-IDF algorithm works:

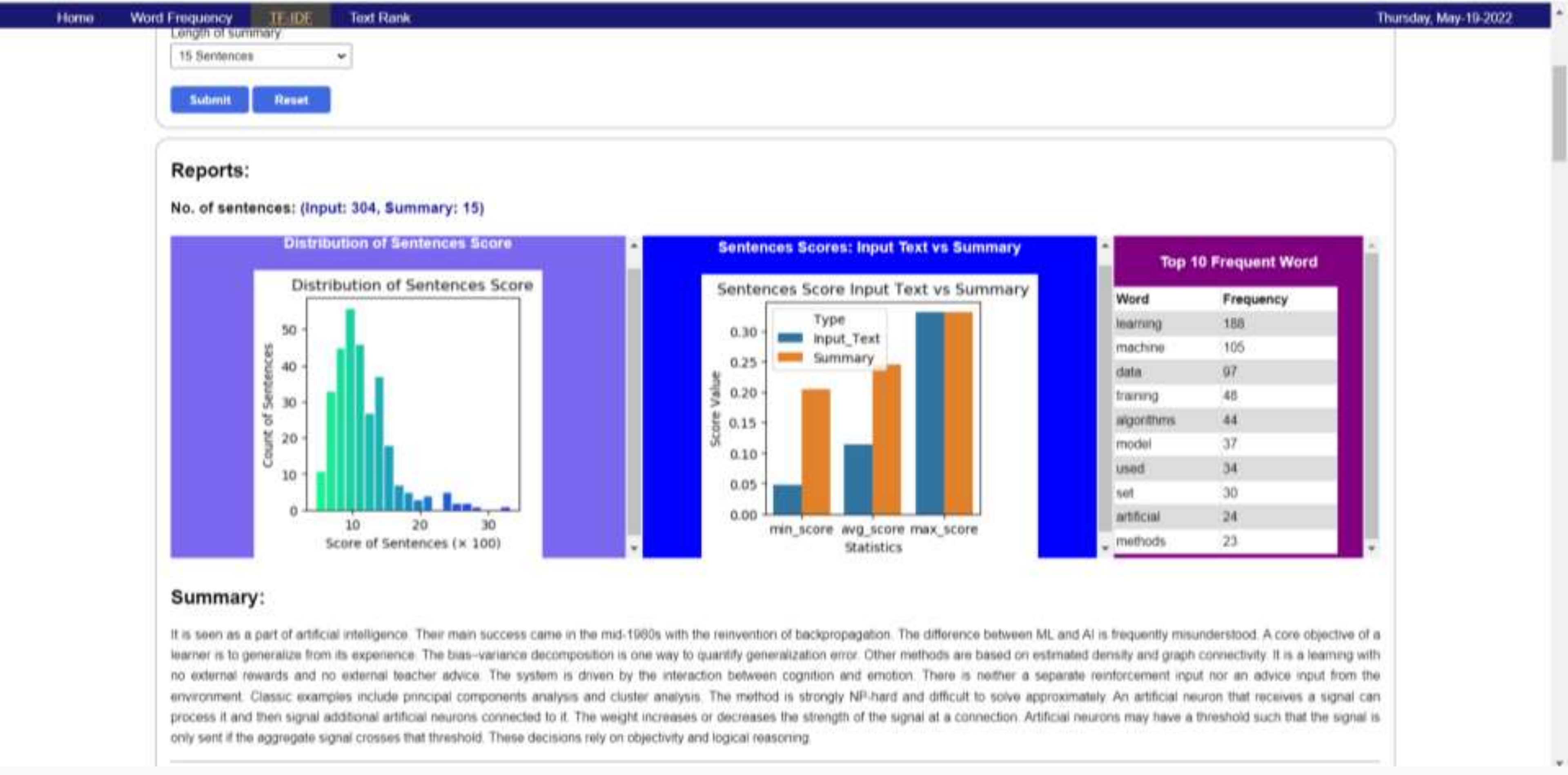
Step 8: Text summary is obtained by sorting the sentences based on sentence scores.

The keywords are the most frequently repeated words in the given input text.



```
Have you experienced this before? Who is right and who is wrong?
Neither. It was at that point their biggest breakthrough came.
Perhaps all those years of perseverance finally paid off. It must
come from within you. Where are you settling in your life right now?
Could you be you playing for bigger stakes than you are? So become
intentional on what you want out of life. Commit to it. Nurture your
dreams.
```

The top 10 frequent words are considered as **keywords**.



Length of summary: 15 Sentences

Reports:

No. of sentences: (Input: 304, Summary: 15)

Top 10 Frequent Word

Word	Frequency
learning	188
machine	105
data	97
training	48
algorithms	44
model	37
used	34
set	30
artificial	24
methods	23

Summary:

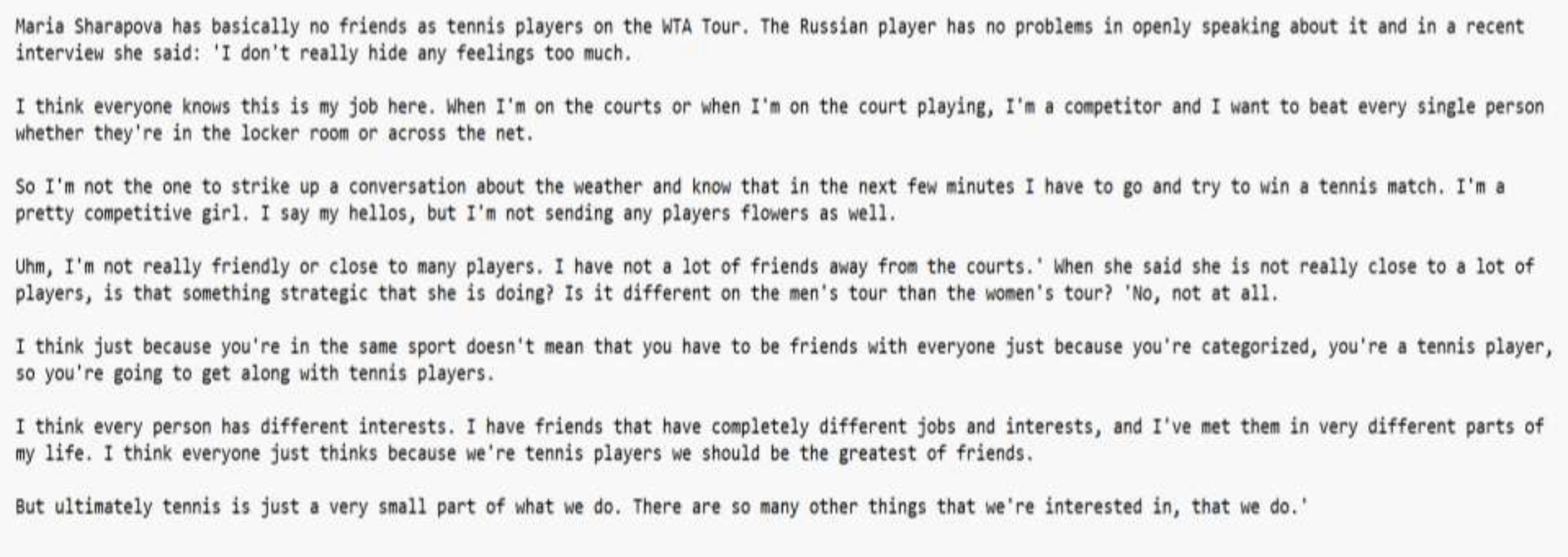
It is seen as a part of artificial intelligence. Their main success came in the mid-1980s with the reinvention of backpropagation. The difference between ML and AI is frequently misunderstood. A core objective of a learner is to generalize from its experience. The bias–variance decomposition is one way to quantify generalization error. Other methods are based on estimated density and graph connectivity. It is a learning with no external rewards and no external teacher advice. The system is driven by the interaction between cognition and emotion. There is neither a separate reinforcement input nor an advice input from the environment. Classic examples include principal components analysis and cluster analysis. The method is strongly NP-hard and difficult to solve approximately. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. These decisions rely on objectivity and logical reasoning.

Final output screen using TF-IDF algorithm.

Text Rank algorithm

Text Summarization and Keyword Extraction using TextRank algorithm.

First a url of the website which contains the text is given as input. For example the text is like this from which the summary and keywords must be drawn :



Maria Sharapova has basically no friends as tennis players on the WTA Tour. The Russian player has no problems in openly speaking about it and in a recent interview she said: 'I don't really hide any feelings too much.

I think everyone knows this is my job here. When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether they're in the locker room or across the net.

So I'm not the one to strike up a conversation about the weather and know that in the next few minutes I have to go and try to win a tennis match. I'm a pretty competitive girl. I say my hellos, but I'm not sending any players flowers as well.

Uhm, I'm not really friendly or close to many players. I have not a lot of friends away from the courts.' When she said she is not really close to a lot of players, is that something strategic that she is doing? Is it different on the men's tour than the women's tour? 'No, not at all.

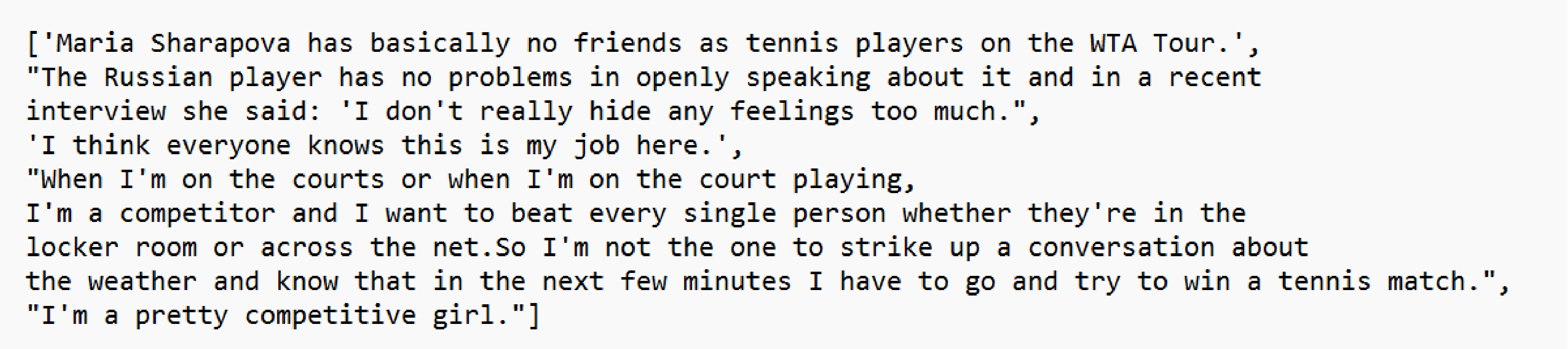
I think just because you're in the same sport doesn't mean that you have to be friends with everyone just because you're categorized, you're a tennis player, so you're going to get along with tennis players.

I think every person has different interests. I have friends that have completely different jobs and interests, and I've met them in very different parts of my life. I think everyone just thinks because we're tennis players we should be the greatest of friends.

But ultimately tennis is just a very small part of what we do. There are so many other things that we're interested in, that we do.'

How the TextRank algorithm works:

Step 1 : Split the text into individual sentences.



```
['Maria Sharapova has basically no friends as tennis players on the WTA Tour.',
"The Russian player has no problems in openly speaking about it and in a recent
interview she said: 'I don't really hide any feelings too much.",
'I think everyone knows this is my job here.',
"When I'm on the courts or when I'm on the court playing,
I'm a competitor and I want to beat every single person whether they're in the
locker room or across the net.So I'm not the one to strike up a conversation about
the weather and know that in the next few minutes I have to go and try to win a tennis match.",
"I'm a pretty competitive girl."]
```

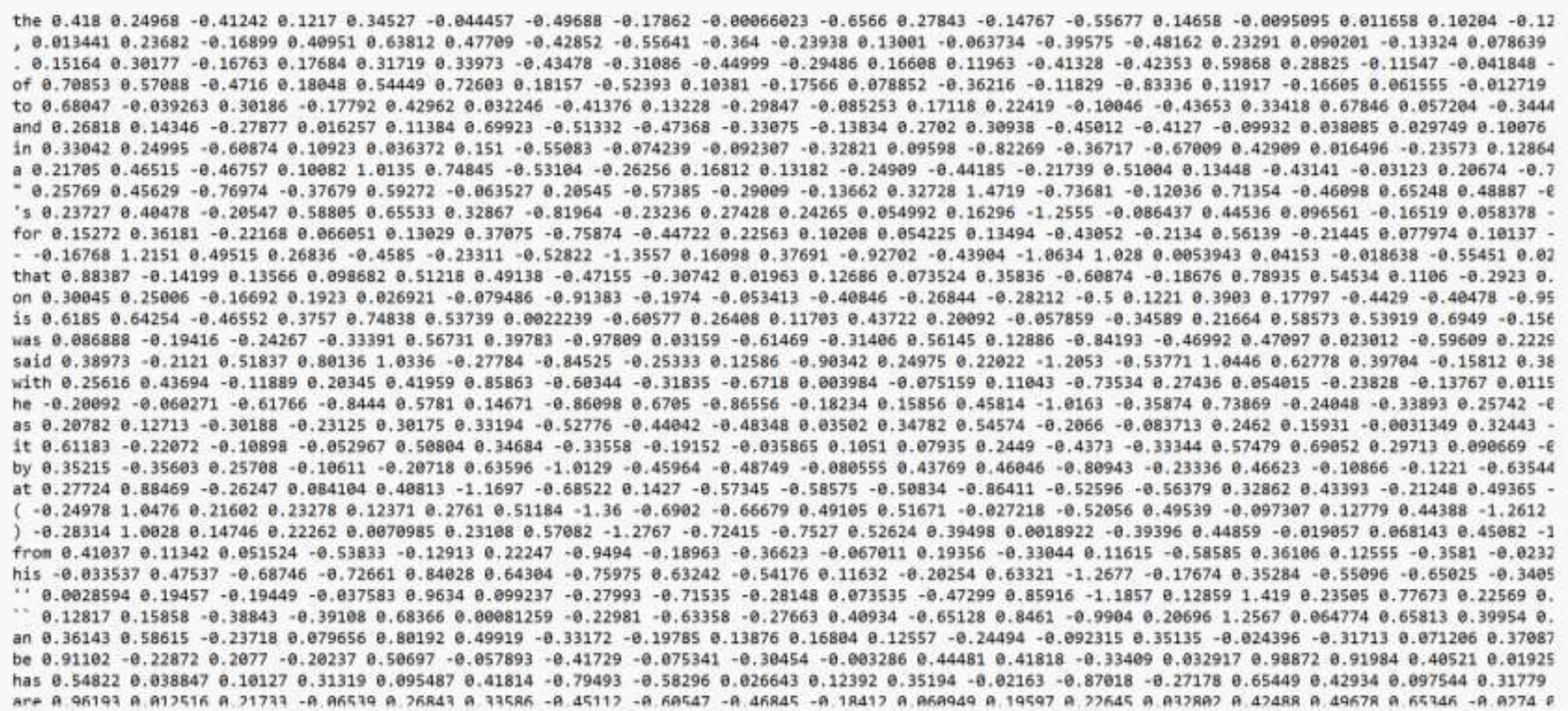
Here, the first 5 sentences are displayed.

Next, text preprocessing is performed on these sentences.

How the TextRank algorithm works:

Step 2: Find the vector representation for every sentence using GloVe embedding.

GloVe embedding is imported which contains the following.



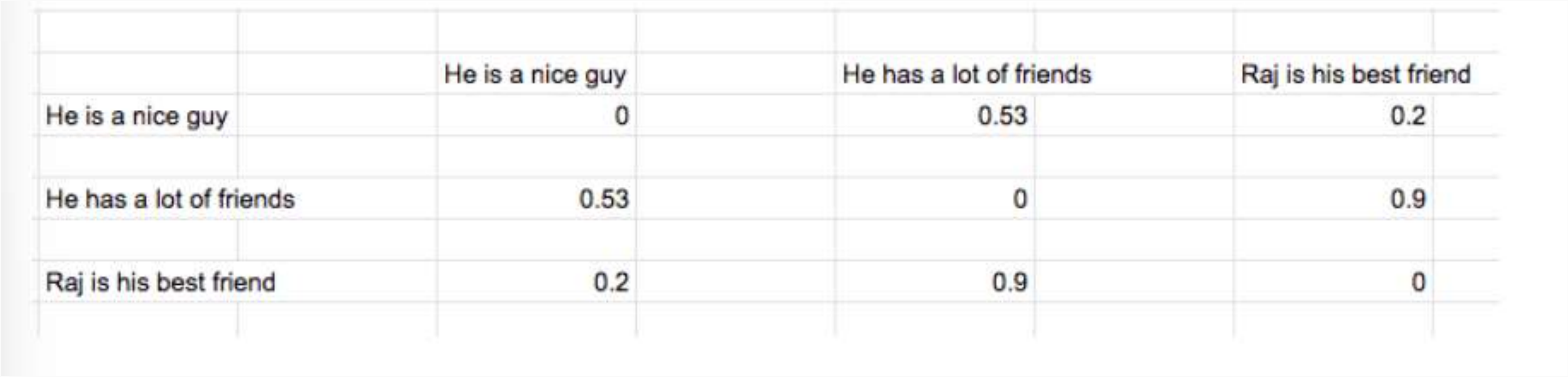
```
the 0.418 0.24968 -0.41242 0.1217 0.34527 -0.044457 -0.49688 -0.17862 -0.00066023 -0.6566 0.27843 -0.14767 -0.55677 0.14658 -0.0095095 0.011658 0.10204 -0.12
, 0.013441 0.23682 -0.16899 0.40951 0.63812 0.47709 -0.42852 -0.55641 -0.364 -0.23938 0.13001 -0.063734 -0.39575 -0.48162 0.23291 0.090201 -0.13324 0.078639
. 0.15164 0.30177 -0.16763 0.17684 0.31719 0.33973 -0.43478 -0.31086 -0.44999 -0.29486 0.16608 0.11963 -0.41328 -0.42353 0.59868 0.28825 -0.11547 -0.041848 -
of 0.70853 0.57088 -0.4716 0.18048 0.54449 0.72603 0.18157 -0.52393 0.10381 -0.17566 0.078852 -0.36216 -0.11829 -0.83336 0.11917 -0.16605 0.061555 -0.012719
to 0.68047 -0.039263 0.30186 -0.17792 0.42962 0.032246 -0.41376 0.13228 -0.29847 -0.085253 0.17118 0.22419 -0.10046 -0.43653 0.33418 0.67846 0.057204 -0.3444
and 0.26818 0.14346 -0.27877 0.016257 0.11384 0.69923 -0.51332 -0.47368 -0.33075 -0.13834 0.2702 0.30938 -0.45012 -0.4127 -0.09932 0.038085 0.029749 0.10076
in 0.33042 0.24995 -0.60874 0.10923 0.036372 0.151 -0.55083 -0.074239 -0.092307 -0.32821 0.09598 -0.82269 -0.36717 -0.67009 0.42909 0.016496 -0.23573 0.12864
a 0.21705 0.46515 -0.46757 0.10082 1.0135 0.74845 -0.53104 -0.26256 0.16812 0.13182 -0.24909 -0.44185 -0.21739 0.51004 0.13448 -0.43141 -0.03123 0.20674 -0.7
" 0.25769 0.45629 -0.76974 -0.37679 0.59272 -0.063527 0.20545 -0.57385 -0.29009 -0.13662 0.32728 1.4719 -0.73681 -0.12036 0.71354 -0.46098 0.65248 0.48887 -0
's 0.23727 0.40478 -0.20547 0.58805 0.65533 0.32867 -0.81964 -0.23236 0.27428 0.24265 0.054992 0.16296 -1.2555 -0.086437 0.44536 0.096561 -0.16519 0.058378 -
for 0.15272 0.36181 -0.22168 0.066051 0.13029 0.37075 -0.75874 -0.44722 0.22563 0.10208 0.054225 0.13494 -0.43052 -0.2134 0.56139 -0.21445 0.077974 0.10137 -
- -0.16768 1.2151 0.49515 0.26836 -0.4585 -0.23311 -0.52822 -1.3557 0.16098 0.37691 -0.92702 -0.43904 -1.0634 1.028 0.0053943 0.04153 -0.018638 -0.55451 0.02
that 0.88387 -0.14199 0.13566 0.098682 0.51218 0.49138 -0.47155 -0.30742 0.01963 0.12686 0.073524 0.35836 -0.60874 -0.18676 0.78935 0.54534 0.1106 -0.2923 0.
on 0.30045 0.25006 -0.16692 0.1923 0.026921 -0.079486 -0.91383 -0.1974 -0.053413 -0.40846 -0.26844 -0.28212 -0.5 0.1221 0.3903 0.17797 -0.4429 -0.40478 -0.95
is 0.6185 0.64254 -0.46552 0.3757 0.74838 0.53739 0.0022239 -0.60577 0.26408 0.11703 0.43722 0.20092 -0.057859 -0.34589 0.21664 0.58573 0.53919 0.6949 -0.156
was 0.086888 -0.19416 -0.24267 -0.33391 0.56731 0.39783 -0.97809 0.03159 -0.61469 -0.31406 0.56145 0.12886 -0.84193 -0.46992 0.47097 0.023012 -0.59609 0.2229
said 0.38973 -0.2121 0.51837 0.80136 1.0336 -0.27784 -0.84525 -0.25333 0.12586 -0.90342 0.24975 0.22022 -1.2053 -0.53771 1.0446 0.62778 0.39704 -0.15812 0.38
with 0.25616 0.43694 -0.11889 0.20345 0.41959 0.85863 -0.60344 -0.31835 -0.6718 0.003984 -0.075159 0.11043 -0.73534 0.27436 0.054015 -0.23828 -0.13767 0.0115
he -0.20092 -0.060271 -0.61766 -0.8444 0.5781 0.14671 -0.86098 0.6705 -0.86556 -0.18234 0.15856 0.45814 -1.0163 -0.35874 0.73869 -0.24048 -0.33893 0.25742 -0
as 0.20782 0.12713 -0.30188 -0.23125 0.30175 0.33194 -0.52776 -0.44042 -0.48348 0.03502 0.34782 0.54574 -0.2066 -0.083713 0.2462 0.15931 -0.0031349 0.32443 -
it 0.61183 -0.22072 -0.10898 -0.052967 0.50804 0.34684 -0.33558 -0.19152 -0.035865 0.1051 0.07935 0.2449 -0.4373 -0.33344 0.57479 0.69052 0.29713 0.090669 -0
by 0.35215 -0.35603 0.25708 -0.10611 -0.20718 0.63596 -1.0129 -0.45964 -0.48749 -0.080555 0.43769 0.46046 -0.80943 -0.23336 0.46623 -0.10866 -0.1221 -0.63544
at 0.27724 0.88469 -0.26247 0.084104 0.40813 -1.1697 -0.68522 0.1427 -0.57345 -0.58575 -0.50834 -0.86411 -0.52596 -0.56379 0.32862 0.43393 -0.21248 0.49365 -
( -0.24978 1.0476 0.21602 0.23278 0.12371 0.2761 0.51184 -1.36 -0.6902 -0.66679 0.49105 0.51671 -0.027218 -0.52056 0.49539 -0.097307 0.12779 0.44388 -1.2612
) -0.28314 1.0028 0.14746 0.22262 0.0070985 0.23108 0.57082 -1.2767 -0.72415 -0.7527 0.52624 0.39498 0.0018922 -0.39396 0.44859 -0.019057 0.068143 0.45082 -1
from 0.41037 0.11342 0.051524 -0.53833 -0.12913 0.22247 -0.9494 -0.18963 -0.36623 -0.067011 0.19356 -0.33044 0.11615 -0.58585 0.36106 0.12555 -0.3581 -0.0232
his -0.033537 0.47537 -0.68746 -0.72661 0.84028 0.64304 -0.75975 0.63242 -0.54176 0.11632 -0.20254 0.63321 -1.2677 -0.17674 0.35284 -0.55096 -0.65025 -0.3405
'' 0.0028594 0.19457 -0.19449 -0.037583 0.9634 0.099237 -0.27993 -0.71535 -0.28148 0.073535 -0.47299 0.85916 -1.1857 0.12859 1.419 0.23505 0.77673 0.22569 0.
`` 0.12817 0.15858 -0.38843 -0.39108 0.68366 0.00081259 -0.22981 -0.63358 -0.27663 0.40934 -0.65128 0.8461 -0.9904 0.20696 1.2567 0.064774 0.65813 0.39954 0.
an 0.36143 0.58615 -0.23718 0.079656 0.80192 0.49919 -0.33172 -0.19785 0.13876 0.16804 0.12557 -0.24494 -0.092315 0.35135 -0.024396 -0.31713 0.071206 0.37087
be 0.91102 -0.22872 0.2077 -0.20237 0.50697 -0.057893 -0.41729 -0.075341 -0.30454 -0.003286 0.44481 0.41818 -0.33409 0.032917 0.98872 0.91984 0.40521 0.01925
has 0.54822 0.038847 0.10127 0.31319 0.095487 0.41814 -0.79493 -0.58296 0.026643 0.12392 0.35194 -0.02163 -0.87018 -0.27178 0.65449 0.42934 0.097544 0.31779
are 0.96193 0.012516 0.21733 -0.06539 0.26843 0.33586 -0.45112 -0.60547 -0.46845 -0.18412 0.060949 0.19597 0.22645 0.032802 0.42488 0.49678 0.65346 -0.0274 0
```

How the TextRank algorithm works:

Step 3: The similarities between sentence vectors are then calculated and stored in a matrix, which is called a similarity matrix.
We will use the cosine similarity approach for this.

For example consider the following text:

‘He is a nice guy. He has a lot of friends. Raj is his best friend’

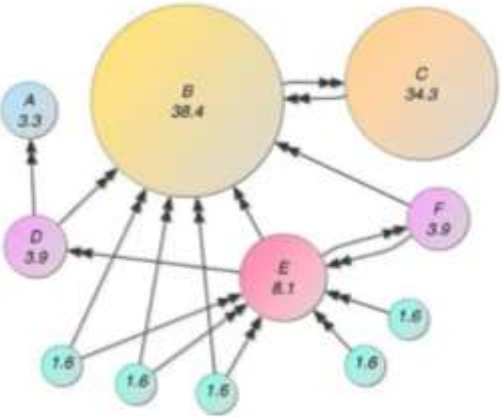


	He is a nice guy	He has a lot of friends	Raj is his best friend
He is a nice guy	0	0.53	0.2
He has a lot of friends	0.53	0	0.9
Raj is his best friend	0.2	0.9	0

How the TextRank algorithm works:

Step 4: The similarity matrix is converted into graph which is necessary for sentence rank calculation.

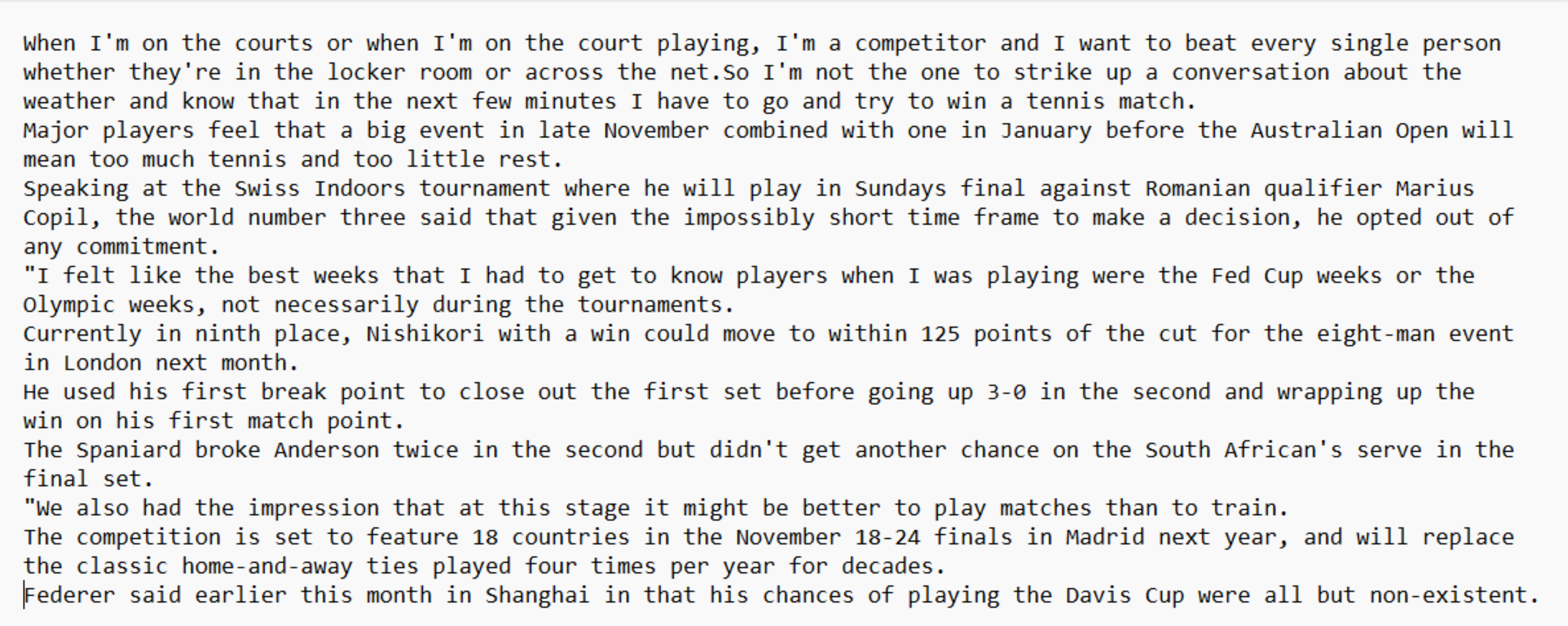
On this graph, we will apply the PageRank algorithm to arrive at the sentence rankings.



PageRank illustration

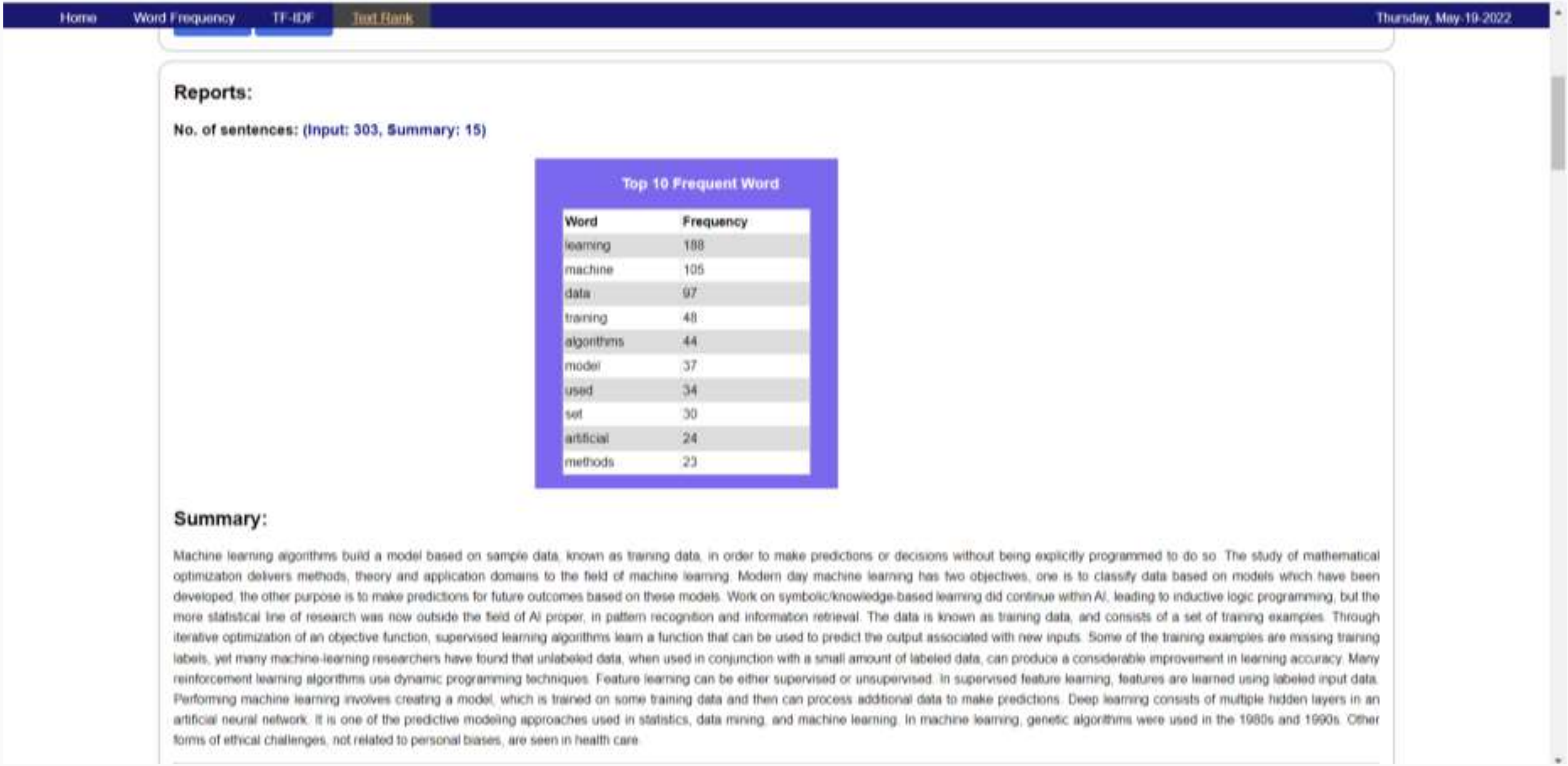
How the TextRank algorithm works:

Step 5: The top-ranked sentences form the final summary as the output.



```
When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person
whether they're in the locker room or across the net.So I'm not the one to strike up a conversation about the
weather and know that in the next few minutes I have to go and try to win a tennis match.
Major players feel that a big event in late November combined with one in January before the Australian Open will
mean too much tennis and too little rest.
Speaking at the Swiss Indoors tournament where he will play in Sundays final against Romanian qualifier Marius
Copil, the world number three said that given the impossibly short time frame to make a decision, he opted out of
any commitment.
"I felt like the best weeks that I had to get to know players when I was playing were the Fed Cup weeks or the
Olympic weeks, not necessarily during the tournaments.
Currently in ninth place, Nishikori with a win could move to within 125 points of the cut for the eight-man event
in London next month.
He used his first break point to close out the first set before going up 3-0 in the second and wrapping up the
win on his first match point.
The Spaniard broke Anderson twice in the second but didn't get another chance on the South African's serve in the
final set.
"We also had the impression that at this stage it might be better to play matches than to train.
The competition is set to feature 18 countries in the November 18-24 finals in Madrid next year, and will replace
the classic home-and-away ties played four times per year for decades.
Federer said earlier this month in Shanghai in that his chances of playing the Davis Cup were all but non-existent.
```

The top 10 frequent words are considered as **keywords**.



Reports:

No. of sentences: (Input: 303, Summary: 15)

Top 10 Frequent Word

Word	Frequency
learning	188
machine	105
data	97
training	48
algorithms	44
model	37
used	34
set	30
artificial	24
methods	23

Summary:

Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Modern day machine learning has two objectives, one is to classify data based on models which have been developed, the other purpose is to make predictions for future outcomes based on these models. Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming, but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval. The data is known as training data, and consists of a set of training examples. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. Some of the training examples are missing training labels, yet many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy. Many reinforcement learning algorithms use dynamic programming techniques. Feature learning can be either supervised or unsupervised. In supervised feature learning, features are learned using labeled input data. Performing machine learning involves creating a model, which is trained on some training data and then can process additional data to make predictions. Deep learning consists of multiple hidden layers in an artificial neural network. It is one of the predictive modeling approaches used in statistics, data mining, and machine learning. In machine learning, genetic algorithms were used in the 1980s and 1990s. Other forms of ethical challenges, not related to personal biases, are seen in health care.

Final output screen using TextRank algorithm.

Conclusion

In this project 3 different algorithms are implemented, with choice to select the desired algorithm, which will reduce the text size and create a summary of our input text data and also provide keywords.

Automatic text summarization is a tool that enables a quantum leap in human productivity by simplifying the sheer volume of information that humans interact with daily. This not only allows people to cut down on the reading necessary but also frees up time to read and understand otherwise overlooked written works.

The keyword extraction process helps us in identifying the important words. These keywords will provide the gist of complete text.

Thank you.