SYNOPSIS for NLP Text Summarization:-

In Natural Language Processing, or NLP, Text Summarization refers to the process of using Deep Learning and Machine Learning models to synthesize large bodies of texts into their most important parts. Text Summarization can be applied to static, pre-existing texts, like research papers or news stories, or to audio or video streams, like a podcast or YouTube video, with the help of Speech-to-Text APIs.

Say, for example, you wanted to summarize the 2021 State of the Union Address–an hour and 43 minute long video.

Using a Text Summarization API with time stamps, you might be able to generate the following summaries for key sections of the video:

1:45: I have the high privilege and distinct honor to present to you the President of the United States.

31:42: 90% of Americans now live within 5 miles of a vaccination site.

44:28: The American job plan is going to create millions of good paying jobs.

47:59: No one working 40 hours a week should live below the poverty line.

48:22: American jobs finally be the biggest increase in non defense research and development.

49:21: The National Institute of Health, the NIH, should create a similar advanced research Projects agency for Health.

50:31: It would have a singular purpose to develop breakthroughs to prevent, detect and treat diseases like Alzheimer's, diabetes and cancer.

51:29: I wanted to lay out before the Congress my plan.

52:19: When this nation made twelve years of public education universal in the last century, it made us the best educated, best prepared nation in the world.

54:25: The American Family's Plan guarantees four additional years of public education for every person in America, starting as early as we can.

57:08: American Family's Plan will provide access to quality, affordable childcare.

61:58: I will not impose any tax increase on people making less than $400,000.

67:34: He said the U.S. will become an Arsenal for vaccines for other countries.

74:12: After 20 years of value, Valor and sacrifice, it's time to bring those troops home.

76:01: We have to come together to heal the soul of this nation.

80:02: Gun violence has become an epidemic in America.

84:23: If you believe we need to secure the border, pass it.

85:00: Congress needs to pass legislation this year to finally secure protection for dreamers.

87:02: If we want to restore the soul of America, we need to protect the right to vote.

This makes the video much more understandable at a glance.

Text Preprocessing steps required:-

In NLP, text preprocessing is the first step in the process of building a model.

The various text preprocessing steps are:

1.Tokenization

2.Lower casing

3.Stop words removal

4.Stemming

5.Lemmatization

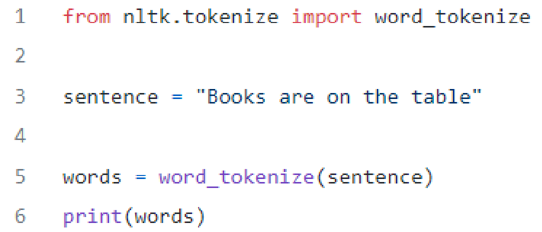
"nlkt" is one of the python library to implement text processing.

It can installed as ,

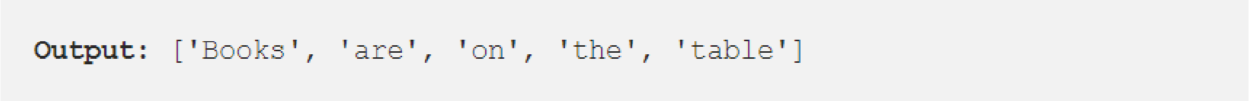
```
pip install nltk==3.4.5
```

TOKENIZATION:-

Splitting the sentence into words.



```python
from nltk.tokenize import word_tokenize

sentence = "Books are on the table"

words = word_tokenize(sentence)
print(words)
```

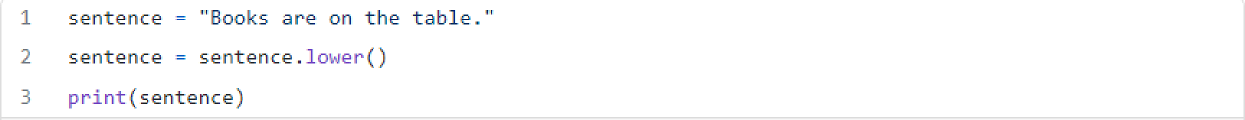


```
Output: ['Books', 'are', 'on', 'the', 'table']
```

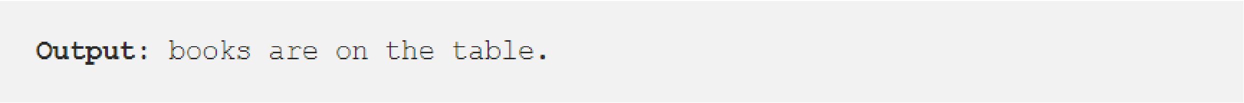
Lower casing:-

Converting a word to lower case (NLP -> nlp).

Words like Book and book mean the same but when not converted to the lower case those two are represented as two different words in the vector space model (resulting in more dimensions).



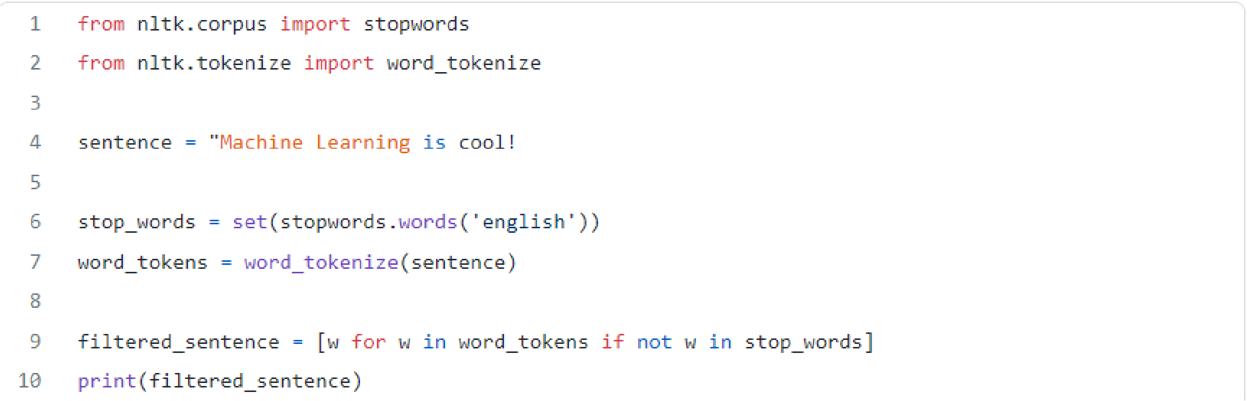
```python
sentence = "Books are on the table."
sentence = sentence.lower()
print(sentence)
```



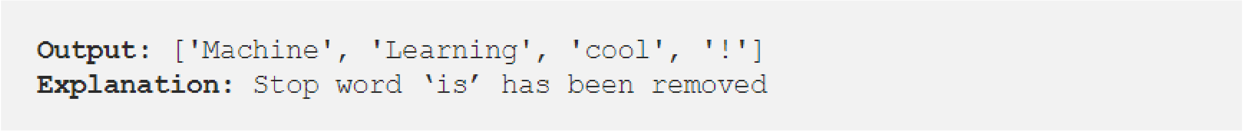
```
Output: books are on the table.
```

Stop words removal:-

Stop words are very commonly used words (a, an, the, etc.) in the documents. These words do not really signify any importance as they do not help in distinguishing two documents.



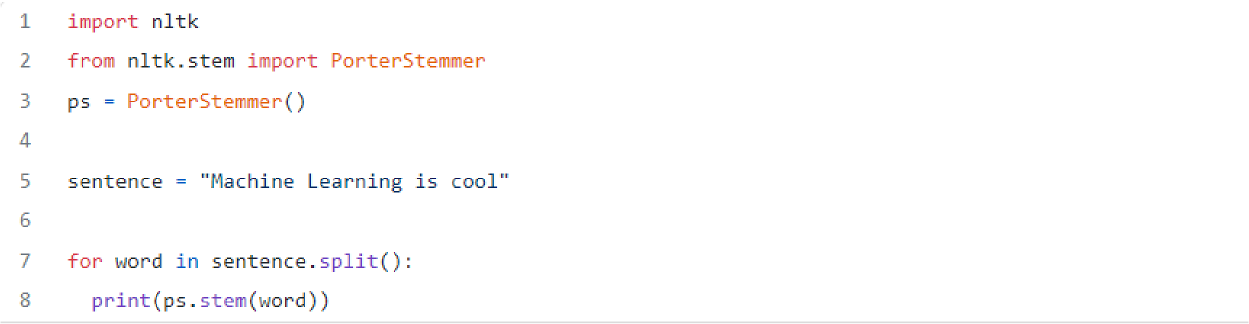
```python
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

sentence = "Machine Learning is cool!

stop_words = set(stopwords.words('english'))
word_tokens = word_tokenize(sentence)

filtered_sentence = [w for w in word_tokens if not w in stop_words]
print(filtered_sentence)
```



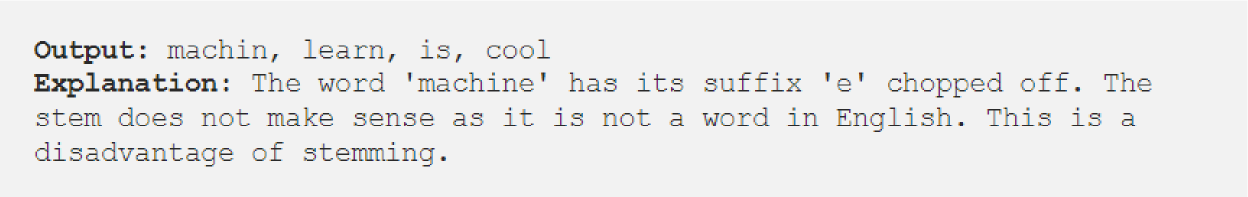
```
Output: ['Machine', 'Learning', 'cool', '!']
Explanation: Stop word 'is' has been removed
```

Stemming:-

It is a process of transforming a word to its root form.



```python
import nltk
from nltk.stem import PorterStemmer
ps = PorterStemmer()

sentence = "Machine Learning is cool"

for word in sentence.split():
  print(ps.stem(word))
```

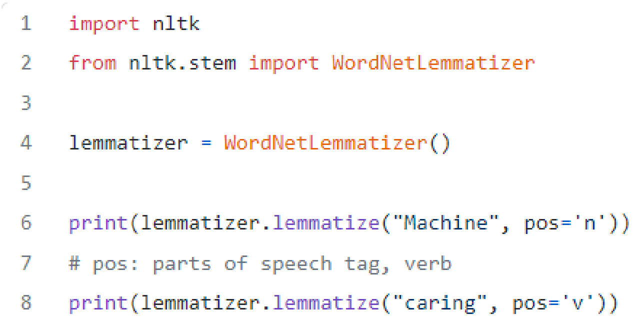


```
Output: machin, learn, is, cool
Explanation: The word 'machine' has its suffix 'e' chopped off. The
stem does not make sense as it is not a word in English. This is a
disadvantage of stemming.
```

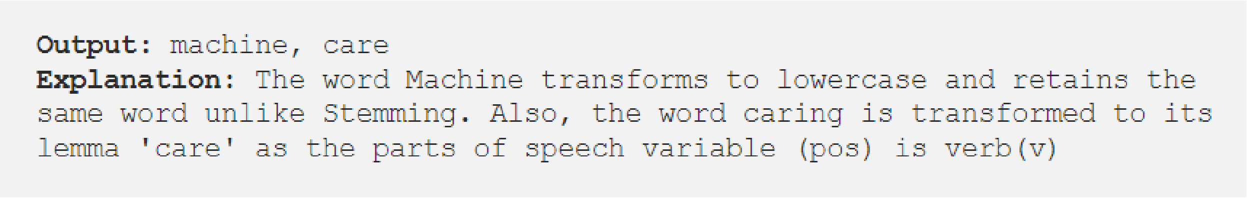
Lemmatization:-

Unlike stemming, lemmatization reduces the words to a word existing in the language.

Lemmatization is preferred over Stemming because lemmatization does a morphological analysis of the words.



```python
import nltk
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print(lemmatizer.lemmatize("Machine", pos='n'))
# pos: parts of speech tag, verb
print(lemmatizer.lemmatize("caring", pos='v'))
```



```
Output: machine, care
Explanation: The word Machine transforms to lowercase and retains the
same word unlike Stemming. Also, the word caring is transformed to its
lemma 'care' as the parts of speech variable (pos) is verb(v)
```

Techniques for Text Summarization:-

Text summarization methods can be grouped into two main categories: **Extractive** and **Abstractive methods**

- **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.

- **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.

Text Summarization using Gensim with TextRank:-

“genism” is a very handy python library for performing NLP tasks. The text summarization process using gensim library is based on **TextRank Algorithm.**

What is TextRank algorithm?

TextRank is an extractive summarization technique. It is based on the concept that words which occur more frequently are significant. Hence , the sentences containing highly frequent words are important .

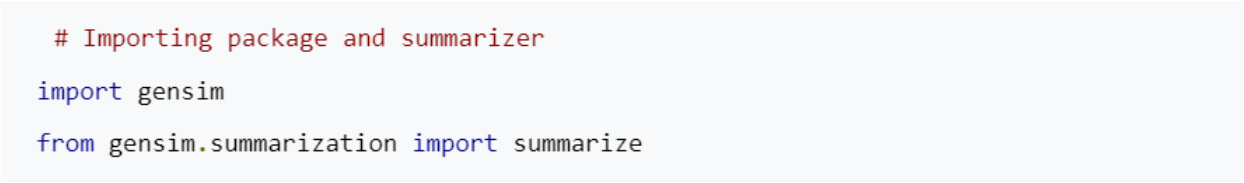
Based on this , the algorithm assigns scores to each sentence in the text . The top-ranked sentences make it to the summary.

Consider the below article on junk foods which has to be summarized.

```
original_text = 'Junk foods taste good that’s why it is mostly liked by everyone of any age group especially kids and
school going children. They generally ask for the junk food daily because they have been trend so by their parents from
the childhood. They never have been discussed by their parents about the harmful effects of junk foods over health.
According to the research by scientists, it has been found that junk foods have negative effects on the health in many
ways. They are generally fried food found in the market in the packets. They become high in calories, high in
cholesterol, low in healthy nutrients, high in sodium mineral, high in sugar, starch, unhealthy fat, lack of protein and
lack of dietary fibers. Processed and junk foods are the means of rapid and unhealthy weight gain and negatively
impact the whole body throughout the life. It makes able a person to gain excessive weight which is called as obesity.
Junk foods tastes good and looks good however do not fulfil the healthy calorie requirement of the body. Some of the
foods like french fries, fried foods, pizza, burgers, candy, soft drinks, baked goods, ice cream, cookies, etc are the
example of high-sugar and high-fat containing foods. It is found according to the Centres for Disease Control and
Prevention that Kids and children eating junk food are more prone to the type-2 diabetes. In type-2 diabetes our body
```

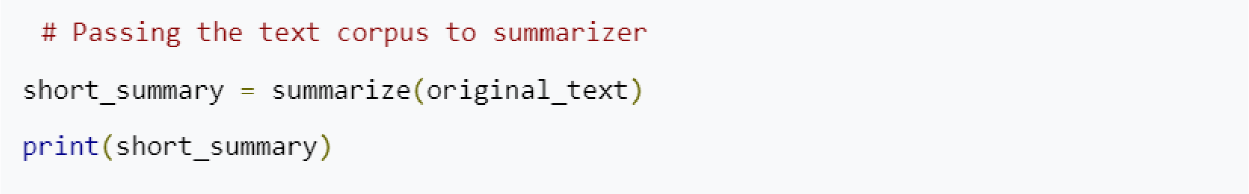
become unable to regulate blood sugar level. Risk of getting this disease is increasing as one become more obese or overweight. It increases the risk of kidney failure. Eating junk food daily lead us to the nutritional deficiencies in the body because it is lack of essential nutrients, vitamins, iron, minerals and dietary fibers. It increases risk of cardiovascular diseases because it is rich in saturated fat, sodium and bad cholesterol. High sodium and bad cholesterol diet increases blood pressure and overloads the heart functioning. One who like junk food develop more risk to put on extra weight and become fatter and unhealthier. Junk foods contain high level carbohydrate which spike blood sugar level and make person more lethargic, sleepy and less active and alert. Reflexes and senses of the people eating this food become dull day by day thus they live more sedentary life. Junk foods are the source of constipation and other disease like diabetes, heart ailments, clogged arteries, heart attack, strokes, etc because of being poor in nutrition. Junk food is the easiest way to gain unhealthy weight. The amount of fats and sugar in the food makes you gain weight rapidly. However, this is not a healthy weight. It is more of fats and cholesterol which will have a harmful impact on your health. Junk food is also one of the main reasons for the increase in obesity nowadays.This food only looks and tastes good, other than that, it has no positive points. The amount of calorie your body requires to stay fit is not fulfilled by this food. For instance, foods like French fries, burgers, candy, and cookies, all have high amounts of sugar and fats. Therefore, this can result in long-term illnesses like diabetes and high blood pressure. This may also result in kidney failure. Above all, you can get various nutritional deficiencies when you don’t consume the essential nutrients, vitamins, minerals and more. You become prone to cardiovascular diseases due to the consumption of bad cholesterol and fat plus sodium. In other words, all this interferes with the functioning of your heart. Furthermore, junk food contains a higher level of carbohydrates. It will instantly spike your blood sugar levels. This will result in lethargy, inactiveness, and sleepiness. A person reflex becomes dull overtime and they lead an inactive life. To make things worse, junk food also clogs your arteries and increases the risk of a heart attack. Therefore, it must be avoided at the first instance to save your life from becoming ruined.The main problem with junk food is that people don’t realize its ill effects now. When the time comes, it is too late. Most importantly, the issue is that it does not impact you instantly. It works on your overtime; you will face the consequences sooner or later. Thus, it is better to stop now.You can avoid junk food by encouraging your children from an early age to eat green vegetables. Their taste buds must be developed as such that they find healthy food tasty. Moreover, try to mix things up. Do not serve the same green vegetable daily in the same style. Incorporate different types of healthy food in their diet following different recipes. This will help them to try foods at home rather than being attracted to junk food.In short, do not deprive them completely of it as that will not help. Children will find one way or the other to have it. Make sure you give them junk food in limited quantities and at healthy periods of time. '

After importing the gensim package, the first step is to import summarize from gensim.summarization. It is an in-built function that implements TextRank.



```
 # Importing package and summarizer
import gensim
from gensim.summarization import summarize
```

Next, pass the text corpus as input to summarize function.



```python
 # Passing the text corpus to summarizer

short_summary = summarize(original_text)

print(short_summary)
```

```
They become high in calories, high in cholesterol, low in healthy
nutrients, high in sodium mineral, high in sugar, starch, unhealthy fat,
lack of protein and lack of dietary fibers.

Processed and junk foods are the means of rapid and unhealthy weight gain
and negatively impact the whole body throughout the life.

Junk foods tastes good and looks good however do not fulfil the healthy
calorie requirement of the body.

It is found according to the Centres for Disease Control and Prevention
that Kids and children eating junk food are more prone to the type-2
diabetes.

Eating junk food daily lead us to the nutritional deficiencies in the body
because it is lack of essential nutrients, vitamins, iron, minerals and
dietary fibers.

It increases risk of cardiovascular diseases because it is rich in
saturated fat, sodium and bad cholesterol.

High sodium and bad cholesterol diet increases blood pressure and
overloads the heart functioning.

One who like junk food develop more risk to put on extra weight and become
fatter and unhealthier.

Junk foods contain high level carbohydrate which spike blood sugar level
and make person more lethargic, sleepy and less active and alert.

For instance, foods like French fries, burgers, candy, and cookies, all
have high amounts of sugar and fats.
```

Seems too long right!

Yes, but you can control how long your summarized text should be.

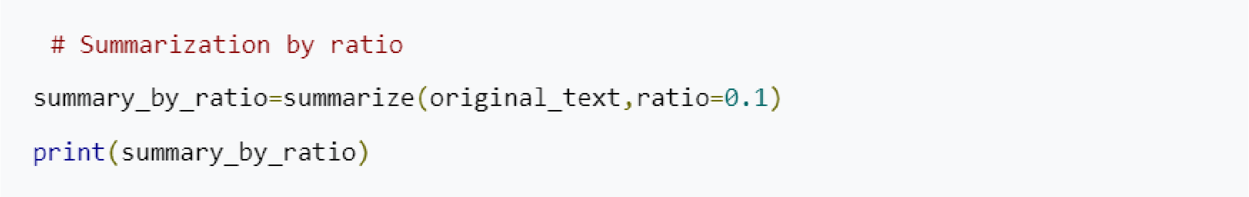
You can change the default parameters of the summarize function according to your requirements.

The parameters are:

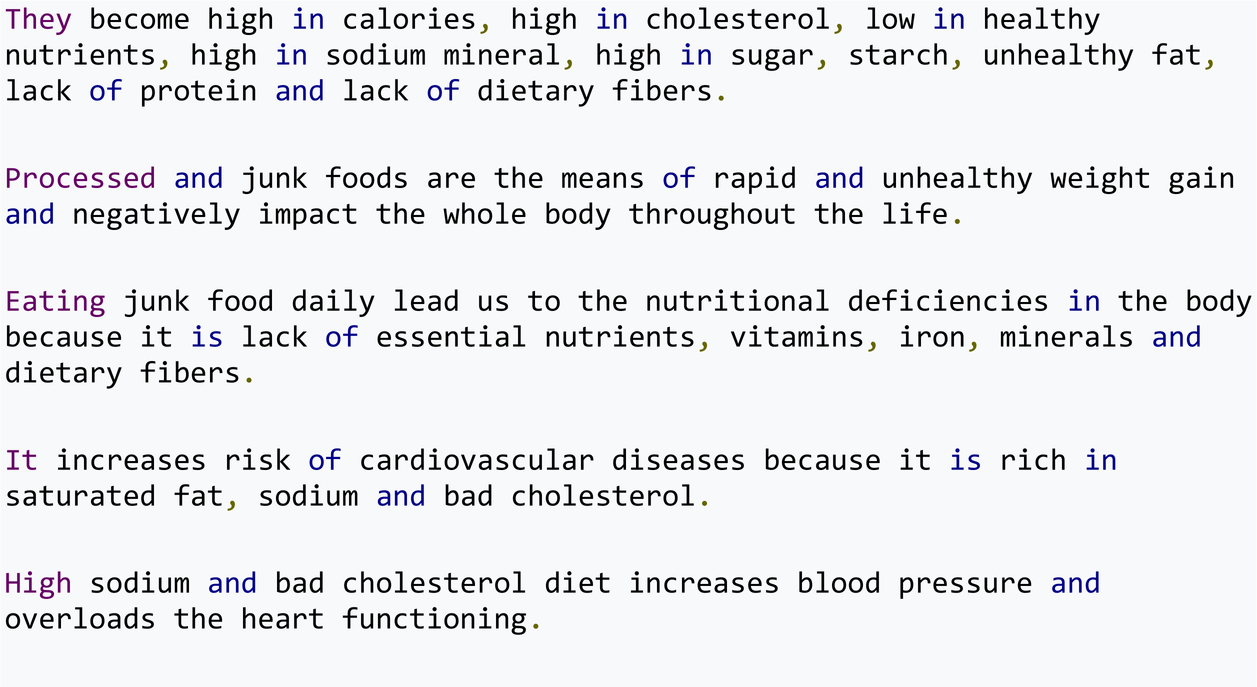
ratio: It can take values between 0 to 1. It represents the proportion of the summary compared to the original text.

word_count: It decides the no of words in the summary.

Let me show you how to use the parameters in above example.



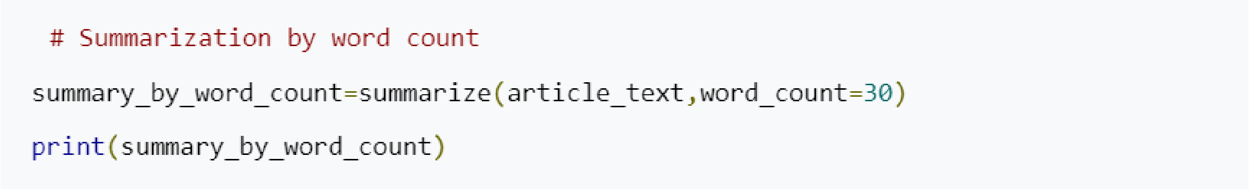
```
 # Summarization by ratio
summary_by_ratio=summarize(original_text,ratio=0.1)
print(summary_by_ratio)
```



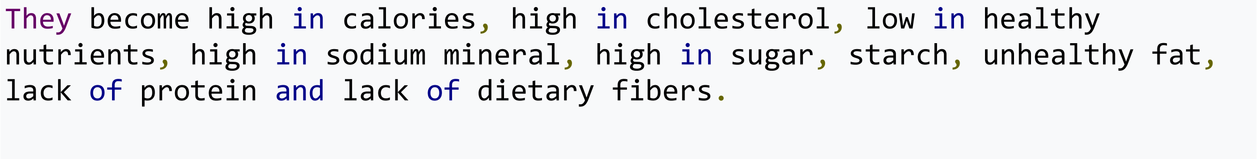
```
They become high in calories, high in cholesterol, low in healthy
nutrients, high in sodium mineral, high in sugar, starch, unhealthy fat,
lack of protein and lack of dietary fibers.

Processed and junk foods are the means of rapid and unhealthy weight gain
and negatively impact the whole body throughout the life.

Eating junk food daily lead us to the nutritional deficiencies in the body
because it is lack of essential nutrients, vitamins, iron, minerals and
dietary fibers.

It increases risk of cardiovascular diseases because it is rich in
saturated fat, sodium and bad cholesterol.

High sodium and bad cholesterol diet increases blood pressure and
overloads the heart functioning.
```

In the above output, you can notice that only 10% of original text is taken as summary.

Likewise, you can summarize using word_count.



```
 # Summarization by word count
summary_by_word_count=summarize(article_text,word_count=30)
print(summary_by_word_count)
```



```
They become high in calories, high in cholesterol, low in healthy
nutrients, high in sodium mineral, high in sugar, starch, unhealthy fat,
lack of protein and lack of dietary fibers.
```

Similar to TextRank , there are various other algorithms which perform summarization. Let's look at it one by one.

Text Summarization with Sumy:-

Along with TextRank , there are various other algorithms to summarize text.

Don't you think it would be very smooth and beneficial to have a library, which will let you perform summarization through multiple algorithms?

Fortunately, we already have the sumy library for it !

sumy libraray provides you several algorithms to implement Text Summarzation. Just import your desired algorithm rather having to code it on your own.

In this section, I shall discuss on implementation of the below algorithms for summarization using sumy :

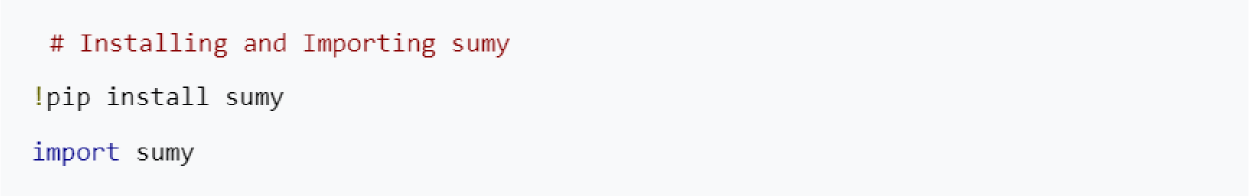
LexRank

Luhn

Latent Semantic Analysis, LSA

KL-Sum

First , import the library through below command



```
 # Installing and Importing sumy
!pip install sumy
import sumy
```

You can acesss different summarizers available through `sumy.summarizers` module.

```
sumy.summarizers

<module 'sumy.summarizers' from '/usr/local/lib/python3.6/dist-packages/sumy/summarizers/__init__.py'>
```

LexRank:-

A sentence which is similar to many other sentences of the text has a high probability of being important. The approach of LexRank is that a particular sentence is recommended by other similar sentences and hence is ranked higher.

Higher the rank, higher is the priority of being included in the summarized text.

I will demonstrate step-by-step on how to summarize the below text

```
original_text='Junk foods taste good that’s why it is mostly liked by
everyone of any age group especially kids and school going children. They
generally ask for the junk food daily because they have been trend so by
their parents from the childhood. They never have been discussed by their
parents about the harmful effects of junk foods over health. According to
the research by scientists, it has been found that junk foods have
negative effects on the health in many ways. They are generally fried food
found in the market in the packets. They become high in calories, high in
cholesterol, low in healthy nutrients, high in sodium mineral, high in
sugar, starch, unhealthy fat, lack of protein and lack of dietary fibers.
Processed and junk foods are the means of rapid and unhealthy weight gain
and negatively impact the whole body throughout the life. It makes able a
person to gain excessive weight which is called as obesity. Junk foods
tastes good and looks good however do not fulfil the healthy calorie
requirement of the body. Some of the foods like french fries, fried foods,
pizza, burgers, candy, soft drinks, baked goods, ice cream, cookies, etc
are the example of high-sugar and high-fat containing foods. It is found
according to the Centres for Disease Control and Prevention that Kids and
children eating junk food are more prone to the type-2 diabetes. In type-2
diabetes our body become unable to regulate blood sugar level. Risk of
getting this disease is increasing as one become more obese or overweight.
It increases the risk of kidney failure. Eating junk food daily lead us to
the nutritional deficiencies in the body because it is lack of essential
nutrients, vitamins, iron, minerals and dietary fibers. It increases risk
of cardiovascular diseases because it is rich in saturated fat, sodium and
bad cholesterol. High sodium and bad cholesterol diet increases blood
pressure and overloads the heart functioning. One who like junk food
develop more risk to put on extra weight and become fatter and
unhealthier. Junk foods contain high level carbohydrate which spike blood
sugar level and make person more lethargic, sleepy and less active and
alert. Reflexes and senses of the people eating this food become dull day
by day thus they live more sedentary life. Junk foods are the source of
constipation and other disease like diabetes, heart ailments, clogged
arteries, heart attack, strokes, etc because of being poor in nutrition.
Junk food is the easiest way to gain unhealthy weight. The amount of fats
and sugar in the food makes you gain weight rapidly. However, this is not
a healthy weight. It is more of fats and cholesterol which will have a
harmful impact on your health. Junk food is also one of the main reasons
for the increase in obesity nowadays.This food only looks and tastes good,
other than that, it has no positive points. The amount of calorie your
body requires to stay fit is not fulfilled by this food. For instance,
foods like French fries, burgers, candy, and cookies, all have high
amounts of sugar and fats. Therefore, this can result in long-term
illnesses like diabetes and high blood pressure. This may also result in
kidney failure. Above all, you can get various nutritional deficiencies
when you don’t consume the essential nutrients, vitamins, minerals and
more. You become prone to cardiovascular diseases due to the consumption
of bad cholesterol and fat plus sodium. In other words, all this
```

```
interferes with the functioning of your heart. Furthermore, junk food
contains a higher level of carbohydrates. It will instantly spike your
blood sugar levels. This will result in lethargy, inactiveness, and
sleepiness. A person reflex becomes dull overtime and they lead an
inactive life. To make things worse, junk food also clogs your arteries
and increases the risk of a heart attack. Therefore, it must be avoided at
the first instance to save your life from becoming ruined.The main problem
with junk food is that people don’t realize its ill effects now. When the
time comes, it is too late. Most importantly, the issue is that it does
not impact you instantly. It works on your overtime; you will face the
consequences sooner or later. Thus, it is better to stop now.You can avoid
junk food by encouraging your children from an early age to eat green
vegetables. Their taste buds must be developed as such that they find
healthy food tasty. Moreover, try to mix things up. Do not serve the same
green vegetable daily in the same style. Incorporate different types of
healthy food in their diet following different recipes. This will help
them to try foods at home rather than being attracted to junk food.In
short, do not deprive them completely of it as that will not help.
Children will find one way or the other to have it. Make sure you give
them junk food in limited quantities and at healthy periods of time. '
```

Next, import `PlaintextParser`. Here, we have a article stored as a string hence we use it. In case of using website sources etc, there are other parsers available. Along with parser, you have to import `Tokenizer` for segmenting the raw text into tokens.

```
# Importing the parser and tokenizer

from sumy.parsers.plaintext import PlaintextParser

from sumy.nlp.tokenizers import Tokenizer
```

You can access the summarizers available through `sumy.summarizers`. Here, I have imported the `LexRankSummarizer`

```
# Import the LexRank summarizer

from sumy.summarizers.lex_rank import LexRankSummarizer
```

As the text source here is a string, you need to use `PlainTextParser.from_string()` function to initialize the parser. You can specify the language used as input to the `Tokenizer`.

syntax : `PlaintextParser.from_string(cls, string, tokenizer)`

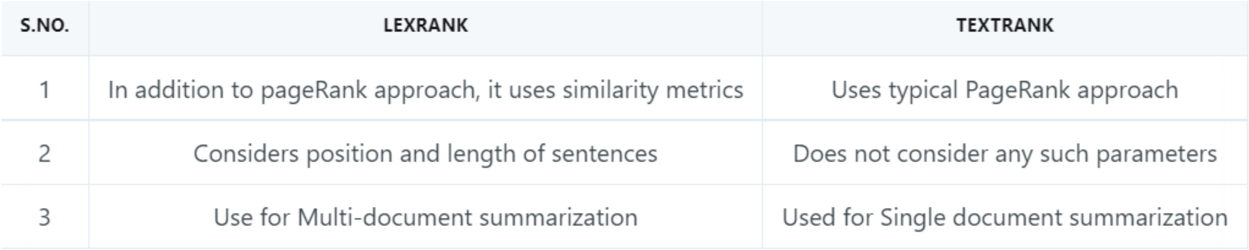
```
# Initializing the parser

my_parser =
PlaintextParser.from_string(original_text,Tokenizer('english'))
```

Next create a summarizer model `lex_rank_summarizer` to fit your text. The syntax is: `lex_rank_summarizer(document, sentences_count)`.

You can decide the number of sentences you want in the summary through parameter `sentences_count`.

```
# Creating a summary of 3 sentences.

lex_rank_summarizer = LexRankSummarizer()

lexrank_summary =
lex_rank_summarizer(my_parser.document,sentences_count=3)



# Printing the summary

for sentence in lexrank_summary:

  print(sentence)

It is found according to the Centres for Disease Control and Prevention
that Kids and children eating junk food are more prone to the type-2
diabetes.

It is more of fats and cholesterol which will have a harmful impact on
your health.

Children will find one way or the other to have it.
```

Advanatge over TextRank is :-



S.NO.	LEXRANK	TEXTRANK
1	In addition to pageRank approach, it uses similarity metrics	Uses typical PageRank approach
2	Considers position and length of sentences	Does not consider any such parameters
3	Use for Multi-document summarization	Used for Single document summarization