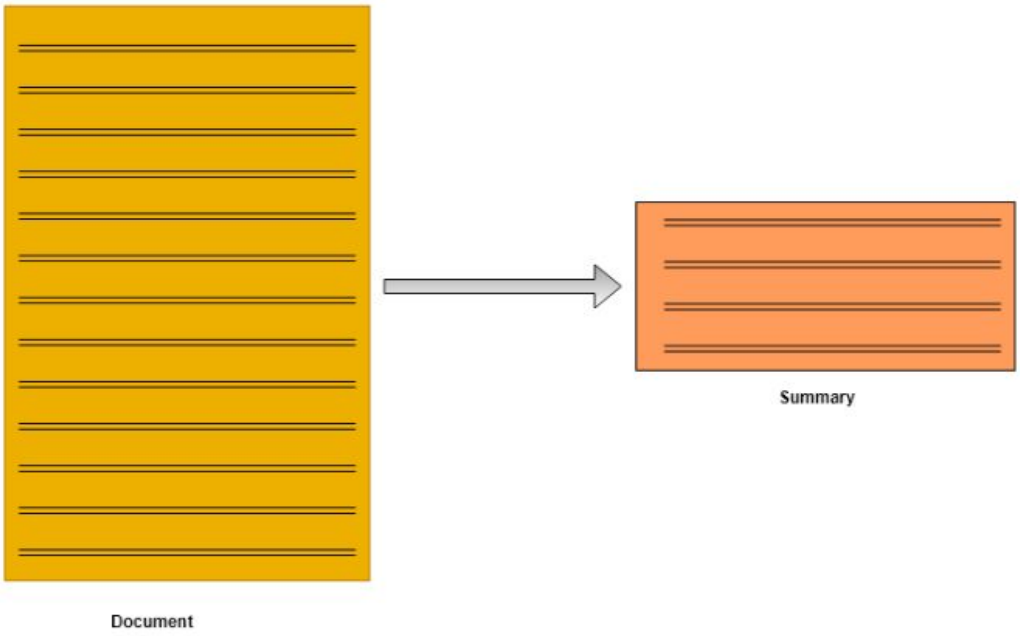


*Summarization can be defined as a task of producing a concise and fluent summary while preserving key information and overall meaning.*

An Extractive Summarization method is used to retain the most important points in a document to form a summary.

Input document → sentences similarity → weight sentences → select sentences with higher rank.

The DeepMind Q&A Dataset is a large collection of news articles from CNN and the Daily Mail with associated questions.This dataset contains more than 93,000 news articles where each article is stored in a single "*.story*" file. The dataset is available at "https://cs.nyu.edu/~kcho/DMQA/".

## 1. Import Necessary Libraries

```
from nltk.corpus import stopwords
from nltk.cluster.util import cosine_distance
import numpy as np
import networkx as nx
from os import listdir
import string
```

## 2. Load Data:

```
# Here we define a function to load the file(single document) to obtain its textual content
def load_doc(filename):
    # open the file in read mode
    file = open(filename, encoding='utf-8')
    text = file.read()
    # close the file and return its textual content
    file.close()
    return text
```

By observing the dataset, we can figure out that the document can be separated into news story text and the summary text. We aim to separate the two. After separation, we can organize all the highlights into a list.

```
# This function splits the doc into story text and summary text
def split_story(doc):
    # We find the index of the occurrence of the phrase '@highlight'
    index = doc.find('@highlight')
    # We split the document by this index
    story, highlights = doc[:index], doc[index:].split('@highlight')
    # Convert the highlights into a list
    highlights = [h.strip() for h in highlights if len(h) > 0]
    return story,highlights
```

```
# Here we define a function to open all files in the directory and load their content
# by calling the above load_doc function
def load_stories(directory):
    # Creating an empty list of all stories
    all_stories = []
    for name in listdir(directory):
        filename = directory + '/' + name
        # Calling the load doc function to load the content of each file
        doc = load_doc(filename)
        # Calling the split document function to separate the story text and summary text
        story,highlight = split_story(doc)
```

```
        all_stories.append({'story':story, 'highlights':highlight})
    return all_stories
```

## 3. Data Cleaning

Now we will process the story data obtained above in all_stories which is a list/array of dictionaries containing story text and summary text for each file separately.
Observing the contents of the dataset, we see that many articles start with source information, presumably the CNN office that produced the story. We clean that along with the punctuations from every token for each line. We also remove numbers and empty strings. Also, normalize to lower-case and remove low frequency words.

```
def clean_lines(lines):
    cleaned = [ ]
    # Prepare a translation table to remove punctuations
    table = str.maketrans('', '', string.punctuation)
    for line in lines:
        index = line.find('(CNN) -- ')
        # Strip source CNN office if it exists
        if index > -1:
            line = line[index+len('(CNN)'):]
        # Tokenize on white space
        line = line.split()
        # Convert to lower case
        line = [word.lower() for word in line]
        # Remove punctuation from each token
        line = [w.translate(table) for w in line]
        # Remove tokens with numbers in them
        line = [word for word in line if word.isalpha()]
        # Store as a string
        cleaned.append(' '.join(line))
    # Remove empty strings
    cleaned = [c for c in cleaned if len(c) > 0]
    return cleaned
```

## 4. Save Clean Data

```
# Save to file for future use
from pickle import dump
dump(stories, open('cnn_dataset.pkl', 'wb'))

# Load from file if needed
stories = load(open('cnn_dataset.pkl', 'rb'))
print('Loaded Stories %d' % len(stories))
```

## 5. Similarity Matrix

```
# This function is helpful in building the similarity matrix by
# finding sentence similarity using cosine distance
def sentence_similarity(sent1, sent2, stopwords=None):
    if stopwords is None:
        stopwords = []

    sent1 = [w.lower() for w in sent1]
    sent2 = [w.lower() for w in sent2]

    all_words = list(set(sent1 + sent2))

    vector1 = [0] * len(all_words)
    vector2 = [0] * len(all_words)

    # Build the vector for the first sentence
    for w in sent1:
        if w in stopwords:
            continue
        vector1[all_words.index(w)] += 1

    # Build the vector for the second sentence
    for w in sent2:
        if w in stopwords:
            continue
        vector2[all_words.index(w)] += 1

    return 1 - cosine_distance(vector1, vector2)
```

```
# This is where we will be using cosine similarity to find similarity between sentences.
def build_similarity_matrix(sentences,stop_words):
    # Create an empty similarity matrix
    similarity_matrix = np.zeros((len(sentences), len(sentences)))
    for idx1 in range(len(sentences)):
        for idx2 in range(len(sentences)):
            # Ignore if both sentences are same
            if idx1 == idx2:
                continue
            similarity_matrix[idx1][idx2] = sentence_similarity(sentences[idx1],
sentences[idx2], stop_words)
    return similarity_matrix
```

## 6. Generate Summary method

```
# This function generates summary and prints it
def generate_summary(sentences, top_n=5):
    stop_words = stopwords.words('english')
    summarize_text = []

    # Generate Similarity Matrix across sentences
    sentence_similarity_martix = build_similarity_matrix(sentences, stop_words)

    # Rank sentences in similarity matrix
    sentence_similarity_graph = nx.from_numpy_array(sentence_similarity_martix)
    scores = nx.pagerank(sentence_similarity_graph)

    # Sort the rank and pick top n sentences
    ranked_sentence = sorted(((scores[i],s) for i,s in enumerate(sentences)), reverse=True)
    print("Indexes of top ranked_sentence order are ", ranked_sentence)

    for i in range(top_n):
        summarize_text.append("".join(ranked_sentence[i][1]))

    # Of Course, output the summarize text
    print("Summarize Text: \n", ". ".join(summarize_text))
```

## 7. Driver code

```
# Load stories
directory = '/home/sunishka/Downloads/NLP/cnn_stories/cnn/stories'
stories = load_stories(directory)
print('Loaded Stories %d' % len(stories))
```

```
# Load stories
directory = '/home/sunishka/Downloads/NLP/cnn_stories/cnn/stories'
stories = load_stories(directory)
print('Loaded Stories %d' % len(stories))
```

```
Loaded Stories 92579
```

```
# Printing an example
print(stories[9]['story'])
print("-----------------------------------------------")
print(stories[9]['highlights'])
```

```
print(stories[9]['story'])
print("-----------------------------------------------")
print(stories[9]['highlights'])
```

```
['we know the devastating force of class hurricanes which have sustained winds exceeding miles per hour or meters
per second like hurricane katrina', 'now imagine winds that are times faster stripping a galaxy of its future ligh
t and heat devastating doesnt begin to describe it', 'data from the new atacama large millimeter array a growing a
rray of radio telescopes in the high desert of chile have mapped a superwind flowing out of a nearby galaxy', 'thi
s galaxy named ngc because it is the object in the new general catalog of galaxies is a bit like our own milky way
galaxy in that it has a large disk of cold gas atoms and molecules of matter out of which stars are constantly for
ming', 'but ngc is a galaxy on steroids it is forming stars at about times the rate of the milky way thats why its
called a starburst galaxy this made it a great target to observe with alma which can see light from the gas from w
hich those stars form', 'this light is not visible with the human eye its electromagnetic radiation like visible l
ight but with a much longer wavelength to observe it we use radio telescopes which look a lot like your basic sate
llite dish only much bigger and more numerous eventually alma will consist of dishes each feet in diameter', 'astr
onomers used alma to measure the amount of carbon monoxide molecules in ngc superwind and to extrapolate the total
amount of cold gas being blasted out of the galaxy', 'whats the big deal well for the first time the new alma anal
ysis shows enough mass in this wind to carry away a large fraction of the gas in the galaxy this suggests an answe
r to a longstanding mystery about why galaxies today do not have more stars', 'lets delve more into this mystery b
y looking at galaxies grow over the billion years since the big bang origin of the universe familiar building bloc
ks of matter electrons protons and neutrons cooled and combined to form atoms then as this gas cooled further some
of the atoms combined to form molecules meanwhile gravity amplified regions of high density so stars and whole gal
axies formed out of the cold dense gas', 'but straightforward calculations describing this process tell us that to
days galaxies should be filled with more stars shining more brightly than we see something must disrupt the ordina
ry gravitypowered process of coalescence and star formation', 'whats missing cold gas', 'the new alma data show cl
early that the gas from which those stars would form is being blown out of the galaxy its a funny kind of death by
excess the very high number of stars forming in ngc generates a lot of heat blowing winds off their surfaces and t
hese streams of heated stellar material unite to push away the surrounding gas that has not formed stars', 'when a
superwind disrupts the gas clouds that might have formed more stars we call it feedback because if the star format
ion rate is very high the superwind is strong so a lot of gas is heated andor ejected from the galaxy so fewer sta
rs form and the superwind dies down conspicuous overproducers like ngc are doomed it seems to cut off their own st
ar production', 'think of it as appetite control gravity has none gravity just keeps pulling on mass particles ato
ms molecules dark matter if gravity were the only important effect the clouds of gas would keep condensing and for
ming many more stars than we see but the superwind keeps this from happening', 'alberto bolatto of the university
of maryland who led the alma study of ngc explained for the first time we can clearly see massive concentrations o
f cold molecular gas being jettisoned by expanding shells of intense pressure created by young stars the amount of
gas we measure gives us very convincing evidence that some growing galaxies blow out more gas than they take in sl
owing star formation down to a crawl', 'a talented science team was behind this result but it wouldnt have been po
ssible without the new alma telescope array which was built by the us national science foundation in partnership w
ith europe japan canada taiwan and the host country chile', 'alma is the largest groundbased astronomical project
ever taking more than years from conception to operation and costing about billion with roughly onethird funded by
the national science foundation it is a model of international cooperation each partner supplied a share of the an
tennas and an international organization oversees telescope operations', 'alma is sited in chile because the ataca
ma desert is ideal for millimeter astronomy its extreme dryness and high altitude about feet above sea level mean
```

```
ma desert is ideal for millimeter astronomy its extreme dryness and high altitude about feet above sea level mean
greater atmospheric transparency to the highfrequency radio waves alma was designed to detect the wavelength of th
is light is about millimeter hence the m in the name', 'people often debate the value of astronomy it wont cure ca
ncer or eradicate poverty but it has its practical value well beyond the primary goal of understanding how the pre
sentday universe and our earth solar system and milky way galaxy came to be', 'astronomy pushes technology advance
ment those digital images you take with your camera or phone are possible because of tools that were developed for
astronomy about years ago', 'think we need a work force better trained in science and technology astronomy gets ki
ds and the curious inner kid in all of us interested in science in chile astronomy is an important part of the hig
htech economy', 'could we live without knowing about the superwind in ngc sure but like art astronomy is part of w
hat enriches our lives for some of us learning where we came from is what its all about', 'the opinions expressed
in this commentary are solely those of meg urry']
--------------------------------------------------
['meg urry new data have mapped a superwind flowing out of a nearby galaxy', 'urry there is enough mass in the win
d to carry away a lot of gas in the galaxy', 'she says this suggests an answer to why galaxies today do not have m
ore stars', 'urry its because cold gas which forms stars is being blown out of galaxies']
```

```
# Clean stories
for example in stories:
    example['story'] = clean_lines(example['story'].split('\n'))
    example['highlights'] = clean_lines(example['highlights'])

# Now, we have a choice to use the code for 'Save Clean Data' in Section 4
# Code to print text summaries for over 92,000 files (not used here due to
# space constraints)
for example in stories:
    sentences = example['story']
    generate_summary(sentences)

# Printing an example story
sentences = stories[9]['story']
generate_summary(sentences)
```

```
# Printing an example story
sentences = stories[9]['story']
generate_summary(sentences)
```

```
Indexes of top ranked_sentence order are  [(0.044376276617326446, 'alberto bolatto of the university of maryland w
ho led the alma study of ngc explained for the first time we can clearly see massive concentrations of cold molecu
lar gas being jettisoned by expanding shells of intense pressure created by young stars the amount of gas we measu
re gives us very convincing evidence that some growing galaxies blow out more gas than they take in slowing star f
ormation down to a crawl'), (0.044359396836348486, 'lets delve more into this mystery by looking at galaxies grow
over the billion years since the big bang origin of the universe familiar building blocks of matter electrons prot
ons and neutrons cooled and combined to form atoms then as this gas cooled further some of the atoms combined to f
orm molecules meanwhile gravity amplified regions of high density so stars and whole galaxies formed out of the co
ld dense gas'), (0.04418493637455351, 'people often debate the value of astronomy it wont cure cancer or eradicate
poverty but it has its practical value well beyond the primary goal of understanding how the presentday universe a
nd our earth solar system and milky way galaxy came to be'), (0.04411348281772223, 'think of it as appetite contro
l gravity has none gravity just keeps pulling on mass particles atoms molecules dark matter if gravity were the on
ly important effect the clouds of gas would keep condensing and forming many more stars than we see but the superw
ind keeps this from happening'), (0.04410374963590429, 'the new alma data show clearly that the gas from which tho
se stars would form is being blown out of the galaxy its a funny kind of death by excess the very high number of s
tars forming in ngc generates a lot of heat blowing winds off their surfaces and these streams of heated stellar m
aterial unite to push away the surrounding gas that has not formed stars'), (0.04406252456286719, 'when a superwin
d disrupts the gas clouds that might have formed more stars we call it feedback because if the star formation rate
is very high the superwind is strong so a lot of gas is heated andor ejected from the galaxy so fewer stars form a
nd the superwind dies down conspicuous overproducers like ngc are doomed it seems to cut off their own star produc
tion'), (0.044060853208742874, 'data from the new atacama large millimeter array a growing array of radio telescop
es in the high desert of chile have mapped a superwind flowing out of a nearby galaxy'), (0.04400865153623515, 'as
tronomy pushes technology advancement those digital images you take with your camera or phone are possible because
of tools that were developed for astronomy about years ago'), (0.0440035078995747, 'but straightforward calculatio
ns describing this process tell us that todays galaxies should be filled with more stars shining more brightly tha
n we see something must disrupt the ordinary gravitypowered process of coalescence and star formation'), (0.043926
82582996099, 'astronomers used alma to measure the amount of carbon monoxide molecules in ngc superwind and to ext
rapolate the total amount of cold gas being blasted out of the galaxy'), (0.0438933212642684, 'could we live witho
ut knowing about the superwind in ngc sure but like art astronomy is part of what enriches our lives for some of u
s learning where we came from is what its all about'), (0.043887754301988026, 'now imagine winds that are times fa
ster stripping a galaxy of its future light and heat devastating doesnt begin to describe it'), (0.043774585800944
56, 'whats the big deal well for the first time the new alma analysis shows enough mass in this wind to carry away
a large fraction of the gas in the galaxy this suggests an answer to a longstanding mystery about why galaxies tod
ay do not have more stars'), (0.04377353254744397, 'this light is not visible with the human eye its electromagnet
ic radiation like visible light but with a much longer wavelength to observe it we use radio telescopes which look
a lot like your basic satellite dish only much bigger and more numerous eventually alma will consist of dishes eac
h feet in diameter'), (0.04367545757752415, 'this galaxy named ngc because it is the object in the new general cat
alog of galaxies is a bit like our own milky way galaxy in that it has a large disk of cold gas atoms and molecule
s of matter out of which stars are constantly forming'), (0.04356316678612012, 'alma is sited in chile because the
atacama desert is ideal for millimeter astronomy its extreme dryness and high altitude about feet above sea level

atacama desert is ideal for millimeter astronomy its extreme dryness and high altitude about feet above sea level
mean greater atmospheric transparency to the highfrequency radio waves alma was designed to detect the wavelength
of this light is about millimeter hence the m in the name'), (0.04353087876034059, 'think we need a work force bet
ter trained in science and technology astronomy gets kids and the curious inner kid in all of us interested in sci
ence in chile astronomy is an important part of the hightech economy'), (0.04343824155472863, 'a talented science
team was behind this result but it wouldnt have been possible without the new alma telescope array which was built
by the us national science foundation in partnership with europe japan canada taiwan and the host country chile'),
(0.043315654932074274, 'the opinions expressed in this commentary are solely those of meg urry'), (0.0432734435137
1074, 'we know the devastating force of class hurricanes which have sustained winds exceeding miles per hour or me
ters per second like hurricane katrina'), (0.043251995636492194, 'alma is the largest groundbased astronomical pro
ject ever taking more than years from conception to operation and costing about billion with roughly onethird fund
ed by the national science foundation it is a model of international cooperation each partner supplied a share of
the antennas and an international organization oversees telescope operations'), (0.04314432455861896, 'but ngc is
a galaxy on steroids it is forming stars at about times the rate of the milky way thats why its called a starburst
galaxy this made it a great target to observe with alma which can see light from the gas from which those stars fo
rm'), (0.03627743744650947, 'whats missing cold gas')]
Summarize Text:
 alberto bolatto of the university of maryland who led the alma study of ngc explained for the first time we can c
learly see massive concentrations of cold molecular gas being jettisoned by expanding shells of intense pressure c
reated by young stars the amount of gas we measure gives us very convincing evidence that some growing galaxies bl
ow out more gas than they take in slowing star formation down to a crawl. lets delve more into this mystery by loo
king at galaxies grow over the billion years since the big bang origin of the universe familiar building blocks of
matter electrons protons and neutrons cooled and combined to form atoms then as this gas cooled further some of th
e atoms combined to form molecules meanwhile gravity amplified regions of high density so stars and whole galaxies
formed out of the cold dense gas. people often debate the value of astronomy it wont cure cancer or eradicate pove
rty but it has its practical value well beyond the primary goal of understanding how the presentday universe and o
ur earth solar system and milky way galaxy came to be. think of it as appetite control gravity has none gravity ju
st keeps pulling on mass particles atoms molecules dark matter if gravity were the only important effect the cloud
s of gas would keep condensing and forming many more stars than we see but the superwind keeps this from happenin
g. the new alma data show clearly that the gas from which those stars would form is being blown out of the galaxy
its a funny kind of death by excess the very high number of stars forming in ngc generates a lot of heat blowing w
inds off their surfaces and these streams of heated stellar material unite to push away the surrounding gas that h
as not formed stars
```