In [1]:
```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
from docx import Document
```

In [2]:

```
def read_word_files(file_paths):
    all_content = []
    for file_path in file_paths:
        try:
            doc = Document(file_path)
            text = []
            for paragraph in doc.paragraphs:
                text.append(paragraph.text)
            all_content.append('\n'.join(text))
        except Exception as e:
            print(f"Error reading the Word file '{file_path}': {e}")
    return all_content

word_file_paths = [r"D:\code\NLP\doc\Doc1.docx", r"D:\code\NLP\doc\Doc 2.docx", r"D:\code\N


contents = read_word_files(word_file_paths)

def tokenize_and_remove_stopwords(contents):
    stop_words = set(stopwords.words('english'))
    tokenized_docs = []
    for doc_content in contents:
        words = word_tokenize(doc_content)
        words = [word for word in words if word.lower() not in stop_words]
        sentences = sent_tokenize(doc_content)
        tokenized_docs.append((words, sentences))
    return tokenized_docs

tokenized_docs_no_stopwords = tokenize_and_remove_stopwords(contents)

for i, (words, sentences) in enumerate(tokenized_docs_no_stopwords, start=1):
    print(f"\nTokens for Document {i} after removing stopwords:")
    print("Words:", words)
    print("Sentences:", sentences)
```

Tokens for Document 1 after removing stopwords:
Words: ['Formula', 'One', ',', 'commonly', 'known', 'Formula', '1', 'F1', ',', 'highest',
'class', 'international', 'racing', 'open-wheel', 'single-seater', 'formula', 'racing', 'c
ars', 'sanctioned', 'Fédération', 'Internationale', 'de', "l'Automobile", '(', 'FIA', ')',
'.', 'FIA', 'Formula', 'One', 'World', 'Championship', 'one', 'premier', 'forms', 'racin
g', 'around', 'world', 'since', 'inaugural', 'running', '1950', '.', 'word', 'formula', 'n
ame', 'refers', 'set', 'rules', 'participants', "'", 'cars', 'must', 'conform', '.', 'Form
ula', 'One', 'season', 'consists', 'series', 'races', ',', 'known', 'Grands', 'Prix', '.',
'Grands', 'Prix', 'take', 'place', 'multiple', 'countries', 'continents', 'around', 'worl
d', 'either', 'purpose-built', 'circuits', 'closed', 'public', 'roads', '.', 'point-syste
m', 'used', 'Grands', 'Prix', 'determine', 'two', 'annual', 'World', 'Championships', ':',
'one', 'drivers', ',', 'one', 'constructors', '(', 'teams', ')', '.', 'driver', 'must', 'h
old', 'valid', 'Super', 'Licence', ',', 'highest', 'class', 'racing', 'licence', 'issued',
'FIA', ',', 'races', 'must', 'held', 'grade', 'one', 'tracks', ',', 'highest', 'grade-rati
ng', 'issued', 'FIA', 'tracks', '.']
Sentences: ["Formula One, commonly known as Formula 1 or F1, is the highest class of inter
national racing for open-wheel single-seater formula racing cars sanctioned by the Fédérat
ion Internationale de l'Automobile (FIA).", 'The FIA Formula One World Championship has be
en one of the premier forms of racing around the world since its inaugural running in 195
0.', "The word formula in the name refers to the set of rules to which all participants' c
ars must conform.", 'A Formula One season consists of a series of races, known as Grands P
rix.', 'Grands Prix take place in multiple countries and continents around the world on ei
ther purpose-built circuits or closed public roads.', 'A point-system is used at Grands Pr
ix to determine two annual World Championships: one for the drivers, and one for the const
ructors (the teams).', 'Each driver must hold a valid Super Licence, the highest class of
racing licence issued by the FIA, and the races must be held on grade one tracks, the high
est grade-rating issued by the FIA for tracks.']

Tokens for Document 2 after removing stopwords:
Words: ['track', ',', 'McLaren', 'Williams', 'teams', 'dominated', '1980s', '1990s', '.',
'Brabham', 'also', 'competitive', 'early', 'part', '1980s', ',', 'winning', 'two', 'Driver
s', "'", 'Championships', 'Nelson', 'Piquet', '.', 'Powered', 'Porsche', ',', 'Honda',
',', 'Mercedes-Benz', ',', 'McLaren', 'sixteen', 'championships', '(', 'seven', 'construct
ors', "'", 'nine', 'drivers', "'", ')', 'period', ',', 'Williams', 'used', 'engines', 'For
d', ',', 'Honda', ',', 'Renault', 'also', 'win', 'sixteen', 'titles', '(', 'nine', 'constr
uctors', "'", 'seven', 'drivers', "'", ')', '.', 'rivalry', 'racers', 'Ayrton', 'Senna',
'Alain', 'Prost', 'became', 'F1', "'s", 'central', 'focus', '1988', 'continued', 'Prost',
'retired', 'end', '1993', '.', 'Senna', 'died', '1994', 'San', 'Marino', 'Grand', 'Prix',
'crashing', 'wall', 'exit', 'notorious', 'curve', 'Tamburello', '.', 'FIA', 'worked', 'imp
rove', 'sport', "'s", 'safety', 'standards', 'since', 'weekend', ',', 'Roland', 'Ratzenber
ger', 'also', 'died', 'accident', 'Saturday', 'qualifying', '.', 'driver', 'died', 'injuri
es', 'sustained', 'track', 'wheel', 'Formula', 'One', 'car', '20', 'years', '2014', 'Japan
ese', 'Grand', 'Prix', ',', 'Jules', 'Bianchi', 'collided', 'recovery', 'vehicle', 'aquapl
aning', 'circuit', ',', 'dying', 'nine', 'months', 'later', 'injuries', '.', 'Since', '199
4', ',', 'three', 'track', 'marshals', 'died', ',', 'one', '2000', 'Italian', 'Grand', 'Pr
ix', ',', '[', '31', ']', 'second', '2001', 'Australian', 'Grand', 'Prix', '[', '31', ']',
'third', '2013', 'Canadian', 'Grand', 'Prix', '.']
Sentences: ['On the track, the McLaren and Williams teams dominated the 1980s and 1990s.',
"Brabham were also being competitive during the early part of the 1980s, winning two Drive
rs' Championships with Nelson Piquet.", "Powered by Porsche, Honda, and Mercedes-Benz, McL
aren won sixteen championships (seven constructors' and nine drivers') in that period, whi
le Williams used engines from Ford, Honda, and Renault to also win sixteen titles (nine co
nstructors' and seven drivers').", "The rivalry between racers Ayrton Senna and Alain Pros
t became F1's central focus during 1988 and continued until Prost retired at the end of 19
93.", 'Senna died at the 1994 San Marino Grand Prix after crashing into a wall on the exit
of the notorious curve Tamburello.', "The FIA worked to improve the sport's safety standar
ds since that weekend, during which Roland Ratzenberger also died in an accident during Sa
turday qualifying.", 'No driver died of injuries sustained on the track at the wheel of a
Formula One car for 20 years until the 2014 Japanese Grand Prix, where Jules Bianchi colli
ded with a recovery vehicle after aquaplaning off the circuit, dying nine months later fro
m his injuries.', 'Since 1994, three track marshals have died, one at the 2000 Italian Gra
nd Prix,[31] the second at the 2001 Australian Grand Prix[31] and the third at the 2013 Ca
nadian Grand Prix.']

Tokens for Document 3 after removing stopwords:
Words: ['major', 'rule', 'shake-up', '2014', 'saw', '2.4-litre', 'naturally', 'aspirated',
'V8', 'engines', 'replaced', '1.6-litre', 'turbocharged', 'hybrid', 'power', 'units', '.',
'prompted', 'Honda', 'return', 'sport', '2015', 'championship', "'s", 'fourth', 'power',
'unit', 'manufacturer', '.', 'Mercedes', 'emerged', 'dominant', 'force', 'rule', 'shake-u
p', ',', 'Lewis', 'Hamilton', 'winning', 'championship', 'closely', 'followed', 'main', 'r
ival', 'teammate', ',', 'Nico', 'Rosberg', ',', 'team', 'winning', '16', '19', 'races', 's

```
eason', '.', 'team', 'continued', 'form', 'following', 'two', 'seasons', ',', 'winning',
'16', 'races', '2015', 'taking', 'record', '19', 'wins', '2016', ',', 'Hamilton', 'claimin
g', 'title', 'former', 'year', 'Rosberg', 'winning', 'latter', 'five', 'points', '.', '201
6', 'season', 'also', 'saw', 'new', 'team', ',', 'Haas', ',', 'join', 'grid', ',', 'Max',
'Verstappen', 'became', 'youngest-ever', 'race', 'winner', 'age', '18', 'Spain', '.']
Sentences: ['A major rule shake-up in 2014 saw the 2.4-litre naturally aspirated V8 engine
s replaced by 1.6-litre turbocharged hybrid power units.', "This prompted Honda to return
to the sport in 2015 as the championship's fourth power unit manufacturer.", 'Mercedes eme
rged as the dominant force after the rule shake-up, with Lewis Hamilton winning the champi
onship closely followed by his main rival and teammate, Nico Rosberg, with the team winnin
g 16 out of the 19 races that season.', 'The team continued this form in the following two
seasons, again winning 16 races in 2015 before taking a record 19 wins in 2016, with Hamil
ton claiming the title in the former year and Rosberg winning it in the latter by five poi
nts.', 'The 2016 season also saw a new team, Haas, join the grid, while Max Verstappen bec
ame the youngest-ever race winner at the age of 18 in Spain.']

Tokens for Document 4 after removing stopwords:
Words: ['race', 'begins', 'warm-up', 'lap', ',', 'cars', 'assemble', 'starting', 'grid',
'order', 'qualified', '.', 'lap', 'often', 'referred', 'formation', 'lap', ',', 'cars', 'l
ap', 'formation', 'overtaking', '(', 'although', 'driver', 'makes', 'mistake', 'may', 'reg
ain', 'lost', 'ground', ')', '.', 'warm-up', 'lap', 'allows', 'drivers', 'check', 'conditi
on', 'track', 'car', ',', 'gives', 'tyres', 'chance', 'warm', 'increase', 'traction', 'gri
p', ',', 'also', 'gives', 'pit', 'crews', 'time', 'clear', 'equipment', 'grid', 'race', 's
tart', '.']
Sentences: ['The race begins with a warm-up lap, after which the cars assemble on the star
ting grid in the order they qualified.', 'This lap is often referred to as the formation l
ap, as the cars lap in formation with no overtaking (although a driver who makes a mistake
may regain lost ground).', 'The warm-up lap allows drivers to check the condition of the t
rack and their car, gives the tyres a chance to warm up to increase traction and grip, and
also gives the pit crews time to clear themselves and their equipment from the grid for th
e race start.']

Tokens for Document 5 after removing stopwords:
Words: ['Formula', 'One', 'constructor', 'entity', 'credited', 'designing', 'chassis', 'en
gine', '.', '[', '97', ']', 'designed', 'company', ',', 'company', 'receives', 'sole', 'cr
edit', 'constructor', '(', 'e.g.', ',', 'Ferrari', ')', '.', 'designed', 'different', 'com
panies', ',', 'credited', ',', 'name', 'chassis', 'designer', 'placed', 'engine', 'designe
r', '(', 'e.g.', ',', 'McLaren-Mercedes', ')', '.', 'constructors', 'scored', 'individuall
y', ',', 'even', 'share', 'either', 'chassis', 'engine', 'another', 'constructor', '(',
'e.g.', ',', 'Williams-Ford', ',', 'Williams-Honda', '1983', ')', '.']
Sentences: ['A Formula One constructor is the entity credited for designing the chassis an
d the engine.', '[97] If both are designed by the same company, that company receives sole
credit as the constructor (e.g., Ferrari).', 'If they are designed by different companies,
both are credited, and the name of the chassis designer is placed before that of the engin
e designer (e.g., McLaren-Mercedes).', 'All constructors are scored individually, even if
they share either chassis or engine with another constructor (e.g., Williams-Ford, William
s-Honda in 1983).']

Tokens for Document 6 after removing stopwords:
Words: ['use', 'volunteers', 'integral', 'making', 'maintaining', 'Wikipedia', '.', 'Howev
er', ',', 'even', 'without', 'internet', ',', 'huge', 'complex', 'projects', 'similar', 'n
ature', 'made', 'use', 'volunteers', '.', 'Specifically', ',', 'creation', 'Oxford', 'Engl
ish', 'Dictionary', 'conceived', 'speech', 'London', 'Library', ',', 'Guy', 'Fawkes', 'Da
y', ',', '5', 'November', '1857', ',', 'Richard', 'Chenevix', 'Trench', '.', 'took', '70',
'years', 'complete', '.', 'Dr.', 'Trench', 'envisioned', 'grand', 'new', 'dictionary', 'ev
ery', 'word', 'English', 'language', ',', 'used', 'democratically', 'freely', '.', 'Accord
ing', 'author', 'Simon', 'Winchester', ',', '``', 'undertaking', 'scheme', ',', 'said',
',', 'beyond', 'ability', 'one', 'man', '.', 'peruse', 'English', 'literature', '–', 'com
b', 'London', 'New', 'York', 'newspapers', 'literate', 'magazines', 'journals', '–', 'mus
t', 'instead', "'the", 'combined', 'action', 'many', '.', "'", 'would', 'necessary', 'recr
uit', 'team', '–', 'moreover', ',', 'huge', 'one', '–', 'probably', 'comprising', 'hundred
s', 'hundreds', 'unpaid', 'amateurs', ',', 'working', 'volunteers', '.']
Sentences: ['The use of volunteers was integral in making and maintaining Wikipedia.', 'Ho
wever, even without the internet, huge complex projects of similar nature had made use of
volunteers.', 'Specifically, the creation of the Oxford English Dictionary was conceived w
ith the speech at the London Library, on Guy Fawkes Day, 5 November 1857, by Richard Chene
vix Trench.', 'It took about 70 years to complete.', 'Dr. Trench envisioned a grand new di
ctionary of every word in the English language, and to be used democratically and freel
y.', 'According to author Simon Winchester, "The undertaking of the scheme, he said, was b
eyond the ability of any one man.', "To peruse all of English literature – and to comb the
London and New York newspapers and the most literate of the magazines and journals – must
be instead 'the combined action of many.'", 'It would be necessary to recruit a team – mor
```

```
eover, a huge one – probably comprising hundreds and hundreds of unpaid amateurs, all of t
hem working as volunteers.']
```

In [9]:

```
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.tokenize import sent_tokenize, word_tokenize
```

In [10]:
```
def select_content_tfidf(tokenized_docs, num_sentences):
    documents = [' '.join(words) for words, _ in tokenized_docs]

    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform(documents)
    feature_names = vectorizer.get_feature_names_out()
    word_tfidf_means = np.mean(tfidf_matrix, axis=0)
    word_tfidf_means = np.array(word_tfidf_means).reshape(-1)
    top_word_indices = np.argsort(word_tfidf_means)[::-1][:5]

    print("Top 5 words based on TF-IDF scores:")
    for idx in top_word_indices[:5]:
        word = feature_names[idx]
        tfidf_score = word_tfidf_means[idx]
        print(f"{word}: {tfidf_score}")

    selected_sentences = []
    for doc_idx, (words, _) in enumerate(tokenized_docs):
        doc_content = ' '.join(words)
        sentences = sent_tokenize(doc_content)
        sentence_tfidf_scores = []
        for sentence in sentences:
            sentence_words = word_tokenize(sentence.lower())
            sentence_tfidf = 0
            for word_idx in top_word_indices:
                word = feature_names[word_idx]
                if word in sentence_words:
                    word_tfidf = word_tfidf_means[word_idx]
                    sentence_tfidf += word_tfidf
            sentence_tfidf_scores.append(sentence_tfidf)
        top_sentence_indices = np.argsort(sentence_tfidf_scores)[::-1][:num_sentences]
        selected_sentences.extend([sentences[idx] for idx in top_sentence_indices])
    return selected_sentences

selected_sentences_tfidf = select_content_tfidf(tokenized_docs_no_stopwords, num_sentences=
print("\nSelected Sentences based on TF-IDF:")
for i, sentence in enumerate(selected_sentences_tfidf, start=1):
    print(f"{i}. {sentence}")
```

```
Top 5 words based on TF-IDF scores:
one: 0.09446063755636129
lap: 0.09063245606237069
prix: 0.07810972780528723
formula: 0.07234073035134353
chassis: 0.061396642684729705

Selected Sentences based on TF-IDF:
1. Formula One season consists series races , known Grands Prix .
2. point-system used Grands Prix determine two annual World Championships : one drivers ,
one constructors ( teams ) .
3. driver died injuries sustained track wheel Formula One car 20 years 2014 Japanese Grand
Prix , Jules Bianchi collided recovery vehicle aquaplaning circuit , dying nine months lat
er injuries .
4. Since 1994 , three track marshals died , one 2000 Italian Grand Prix , [ 31 ] second 20
01 Australian Grand Prix [ 31 ] third 2013 Canadian Grand Prix .
5. 2016 season also saw new team , Haas , join grid , Max Verstappen became youngest-ever
race winner age 18 Spain .
6. team continued form following two seasons , winning 16 races 2015 taking record 19 wins
2016 , Hamilton claiming title former year Rosberg winning latter five points .
7. warm-up lap allows drivers check condition track car , gives tyres chance warm increase
traction grip , also gives pit crews time clear equipment grid race start .
8. lap often referred formation lap , cars lap formation overtaking ( although driver make
s mistake may regain lost ground ) .
9. Formula One constructor entity credited designing chassis engine .
10. constructors scored individually , even share either chassis engine another constructo
r ( e.g.
11. would necessary recruit team – moreover , huge one – probably comprising hundreds hund
reds unpaid amateurs , working volunteers .
12. According author Simon Winchester , `` undertaking scheme , said , beyond ability one
man .
```

In [ ]: